Modeling Junior Hockey

J.W.W Pohlkamp-Hartt

June 3, 2015

1 Introduction

As ice hockey continues to grow in popularity, we are seeing a shift in the way we evaluate the quality of the contestants. Teams, media, and the public are moving away from qualitative valuation based on personal observation and towards quantitative valuation of a player's success. The current surge in statistical analysis of hockey games has lead to improvements in data collection and analysis methods but as a sport there is still a lack of quantitative methods for evaluating many areas of the game.

Four problems within hockey that lack significant statistical analysis are: How do we evaluate neutral zone play? What are the optimal line combinations for a teams players? Can we identify when a player is under or over performing during a game? And lastly, can we predict how with any precision how a game will progress and manage your roster to optimize your potential of winning? We set out to provide solutions to these problems by developing a new model of neutral zone play, modeling player effects on game events, using machine learning techniques to choose optimal lines, and time series methods to predict future events.

2 Neutral Zone Play

In a recent article published in the Toronto Sun [**?**] they spoke of the lack of valuation of neutral zone play that is provided by today's myriad of player metrics calculated. For this reason we find it essential to introduce a way of understanding neutral zone play from a statistical vantage and several metrics that result from this methodology.

The neutral zone in hockey is thought of by many as where games are won and lost. The New Jersey Devils' famous trap system is an example of how strong neutral zone play generate success on the score sheet. Outside of qualitative observations on the strength of a team, tactic, or player in the neutral zone there is little in the way of information about the play in this are of the ice. With few recorded game events coming from the neutral zone, hits and penalties being the most common, there is little data to employ and no identifiable event that is considered a success in the same light as a shot on net or goal.

This leads us to evaluate what we should consider a success within neutral zone play. At its most basic, success can be defined as the ability to control which zone the puck moves to from the neutral zone. We can break this idea of success down into independent events; when a team or player are entering the neutral zone with the puck from the defensive zone and successfully enter the offensive zone or when the opposing team is bringing the puck from their defensive zone into the neutral zone and a team or player takes the puck under their control and navigates out of the neutral zone.

These situations can be treated as 3 separate processes and each process can be modeled as a Bernoulli trial with an estimated probability of success. Another way to think of this is, if player X bring the puck into the neutral zone what is the probability of his team successfully bringing the puck into the offensive zone.

To model these processes we will use a maximum likelihood estimator of the probability of success in a Bernoulli trial. We consider the sum of all neutral zone successes for one situation as $Y_s \sim Binomia(n, p)$ and one games realization as a fixed value y from Y_s. With this we are able to estimate the probability of success, p, to be

$$\hat{p} = \frac{y}{n}, \tag{1}$$

where n is the number of neutral zone events that occurred during the game.

The obvious question here now is "How do we apply this model to ice hockey over several games or even a season?", for this we need only to extend this estimator to include the larger portion of data. That is to say or an entire season with the games numbered $1, 2, ..., N-1, N$, the estimator becomes,

$$\hat{p} = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} n_i}. \tag{2}$$

Now that we have a model to describe the events in the neutral zone, we can create a set of metrics for evaluating the play of an individual or team. The first set of metrics would be direct computation of equation **??** for a given type of success. This would allow us to have statistics for offensive and defensive neutral zone success. These would give us an easily comparable value of each player in a league and 2 simple metrics for a players success in the neutral zone.

We can take these statistics one step further to attain an overall neutral zone effectiveness score. For this we want a statistic that combines both situations and compares a player's skills to that of a league or team average. To do this we determine the cumulative difference of the estimated probability of success over the average for the group. Thus, we define the estimated neutral zone differential (END) for player X relative to group G as

$$END(X, G) = (\hat{p}_O(X) - \hat{p}_O(G)) + (\hat{p}_D(X) - \hat{p}_D(G)), \tag{3}$$

where $\hat{p}_s$ is their estimated probability of success, in for either situation for player X and group G.

This statistic gives an estimated of the effect success a player can be expected to have in the neutral zone, a negative value would indicate their play does not positively affect neutral zone control. The opposite is true for positive scores where they can be expected to control the neutral zone when on the ice.

Having these statistics is a good start to evaluating neutral zone play but without the ability to compute them they become useless. Hockey presents several unique challenges in designing useful statistics. The free flowing manner of play and limited data on puck possession that is recorded for current statistics makes direct calculation of these statistics from the existing data nearly impossible.

The primary method we recommend is direct game time recording of neutral zone attempts leading to estimates of the values for y_O, n_O, y_D, n_D. This method does have some difficulties as there are some neutral zone plays within hockey that may considered to be non events or difficult to describe as failures or successes. For this reason we define an offensive zone success to be any play that begins with the puck in the defensive zone and control in the offensive zone obtained without the other team gaining control of the puck and leaving the neutral zone. We include in this situation the case of dump in attempts that are recovered by the attacking team. In addition non events occur when a stoppage of play is had in the neutral zone for almost any reason, all other events are considered failures. The exception here is offside offensive player stoppages, which we consider failures.

As for defensive neutral zone play, the opposite hold true, any event where the offensive team does not have control of the puck in offensive zone after playing through the neutral zone is a success. It does not matter which end of the ice the defending team skates into when they gain control of the puck. There may be some ambiguity when they puck becomes contested in the neutral zone but we will maintain this is one event until the puck leaves the neutral zone under one teams control.

Differentiating between a team and players neutral zone statistic will be done by using the entire games worth of data for the team and only the subset of data when a player is on the ice for the individual's statistic. This may be problematic in fast paced leagues where the teams change players quickly on the fly but a keen eye and use of existing methods for tracking players can alleviate this issue.

In the event that recording neutral zone data is not feasible, we can identify events within a game through the use of already reported events as proxies. Depending on the league and data available it may possible to identify reasonable proxies of each type of event. Using the NHL as an example of the possible data available, we can use their online interface to gain data in real time of: where the puck is on the ice, who is on the ice, passing attempts, shot attempts, face off outcomes, and giveaways. Using some of this information we can produce a decision process for determining if an event has occurred and whether it is successful. This decision process would vary based on the league and as such should only be thought of a crude approximation for the neutral zone processes.

We have developed a new and statistically well founded metric for valuing the neutral zone play of players and teams in hockey. The difficulties in defining good proxies for events across a variety of leagues will make it difficult to implement this method but as more statistical analysis is performed on the average game, we will be able to use reliable proxies or record the neutral zone events with consistency. The next step with this research is to conduct a data collection study and demonstrate the quality of this statistic.

3 Optimizing Line Selection

Making the choices of who to play and what lines to use during a hockey game can a large impact of the outcome. Historically, many of the top teams in hockey have had strong lines with a significant effect on the game. Usually the problem of choosing lines is resolved using anecdotal evidence to justify a certain set of lines. This practice is not particularly scientific, barely resembling a qualitative analysis. This problem can be quantified by using statistics, allowing us to optimize for a choice of line combinations.

No matter the metric we wish to optimize over for each line combination, it is unfeasible to check every potential set of lines for the teams roster. For example, if we had 16 forwards and 8 defensemen, there would

be 1.7 trillion possible choices, far more that we could measure in a reasonable amount of time with standard computing power. To make this problem more feasible we work with a teams coaching staff to produce a set of potential line combinations.

To be able to optimize the line combinations for a team we need quantify success on the ice and define a metric for a particular combinations effect on this success. Success on the ice can be considered as scoring a goal on the other teams net, while we can also claim that stopping the other team from scoring is a success of another type. Assuming that the actions in the current play can lead to changes in momentum that can lead to scoring within the next minute, we propose using two boolean response variables for measuring on ice success, goal scored by your team in the next 60 seconds and goal scored by the other team in the next 60 seconds. If we take a game a divide it into 30 second blocks, identifying which players were on the ice in that period, we can now use this data as indicator variables for a regression problem.

Using logistic regression we can identify a players effect of both the probability of goals for and against during a game. If we also include the line combinations as 2 and 3 player interaction terms we now have to full structure we expect during the game. This could be extended to include all 5 player combinations also, but we assume that these high level interaction terms are negligible. It is worth noting we assume that the teams are at even strength during the 30 second intervals used.

Now for each set of line combinations we can perform a logistic regression for both dependent variables and obtain estimates of the effects of each line through their coefficients. Then for the probability of goals for we have,

$$P_{gf} = \sigma(\beta_0 + \sum_{i=1}^{18} \beta_i p(i) + \sum_{j=1}^{3} \gamma_j d_j + \sum_{k=1}^{4} \zeta_k f_k), \tag{4}$$

where $p(i)$ is an indicator variable for each player being on the ice with $d(j)$ and $f(k)$ being the indicators for the whole line being on and $\sigma()$ is the logit function.

We would like a pessimistic estimate of the overall effect each possible line combination can have on the game. Knowing that the coefficients from our regression are normally distributed we can estimate the lower 10% confidence limit for the coefficients on the goals for model and the upper 90% confidence limit on the goals against model. If we were to look at the pessimistic net difference in effect each line has on the goals for and against probabilities, we will have a measure of the effect on success the potential line combination has.

$$\begin{aligned}\Delta = &(\sum_{i=1}^{18}(\beta_{i,gf} + \Phi(.1)S(\beta_{i,gf})) + \sum_{j=1}^{3}(\gamma_{j,gf} + \Phi(.1)S(\gamma_{j,gf})) + \sum_{k=1}^{4}(\zeta_{k,gf} + \Phi(.1)S(\zeta_{k,gf})))\\ &- (\sum_{i=1}^{18}(\beta_{i,ga} + \Phi(.9)S(\beta_{i,ga}) + \sum_{j=1}^{3}(\gamma_{j,ga} + \Phi(.9)S(\gamma_{j,ga})) + \sum_{k=1}^{4}(\zeta_{k,ga} + \Phi(.9)S(\zeta_{k,ga})))\end{aligned} \tag{5}$$

Determining Δ for each combination of lines in our potential set, we can identify the optimal combination by the one with the largest Δ. Teams can adjust the significance level used for the pessimistic difference from 10% depending on the situation they are in. Line combinations with little data will have larger variation on their coefficients causing new lines to be undervalued. With large changes in personnel of a team they may want to relax the significance to greater than10% or in situations where a team has greater prior success and little change you may want to use a smaller value. Another variation that can be made is to weight the effect of your data to taper the older the events are. With this you will tune more to how your team is performing now. [expand on this]

4 In-Game Player Monitoring

Many of the methods designed to aid in the evaluation of a players impact on the the outcome of a game is done historically, in an offline context. It is conceivable that simple count statistics like shots, saves, etc. could be reported as the game progresses but these lack context to a players typical play or give unsupervised continuous monitoring of key attributes of play. In an aim to develop a straight forward and visually compelling method to provide real-time in game player monitoring, we look to methods from traditional quality control theory.

An appealing method for providing this real-time reporting is the Exponentially Weighted Moving Average control chart ($EWMA$). An $EWMA$ chart reports the exponentially weighted cumulative standardized deviation of a player's stat values during a game to their expected value. As an example a player's $EWMA$ value for corsi is,

$$EWMA_{corsi}(t) = \lambda \frac{T - corsi(t)}{S} + (1-\lambda) EWMA_{corsi}(t-1), \tag{6}$$

or equivalently,

$$EWMA_{corsi}(t) = \sum_{i=1}^{t} \lambda(1-\lambda)^{i-1} \frac{T - corsi(i)}{T}, \tag{7}$$

where T is the expected or target value and S is the standard deviation of a players corsi.

The value of our $EWMA$ statistics will rely on a choice of T and estimation of S. Depending on the outcome desired and the stationarity of the players statistics throughout the year, the target value can be the mean of last stationary section of games or an entire season if the individuals play has not changed significantly throughout. The standard deviation estimate should be made from the same set of data as the mean. Another option for the target value is to choose a desired value which you would like the player to meet.

Then under the null hypothesis that $H_0 : \mu corsi = T$, we should have

$$EWMA_{corsi}(t) \sim N(0, \sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2i}]}). \tag{8}$$

It is important to note that we are standardizing the EWMA metric to allow for reporting of values that are comparable across players. We can now test at any point if their corsi follows the null hypothesis or is significantly different. For this we recommend using $\alpha = .05$ significance limits, ± 1.96 (or ± 2 for simplicity to lay-people).

Now using the data collection methodology described earlier, we can contiguously update the $EWMA$ statistics for each new time period and provide real-time performance of players during the games. We can see here in the example below that this player significantly out performed his season average on mulitple occasions during this game and around half way through the 3^{rd} period he is playing exceptionally. We can lastly see that he has a significant decrease in corsi rating in the final minutes of the game. In game we could use this information to play him more during his hot streak and to play him less near the end where his play has diminished. Obviously, the coaching decisions made based on the available data would alter the future play of an individual and we could not predict those events but in this retroactive example we can identify two occasions where coaching decisions could be made to improve the teams results.

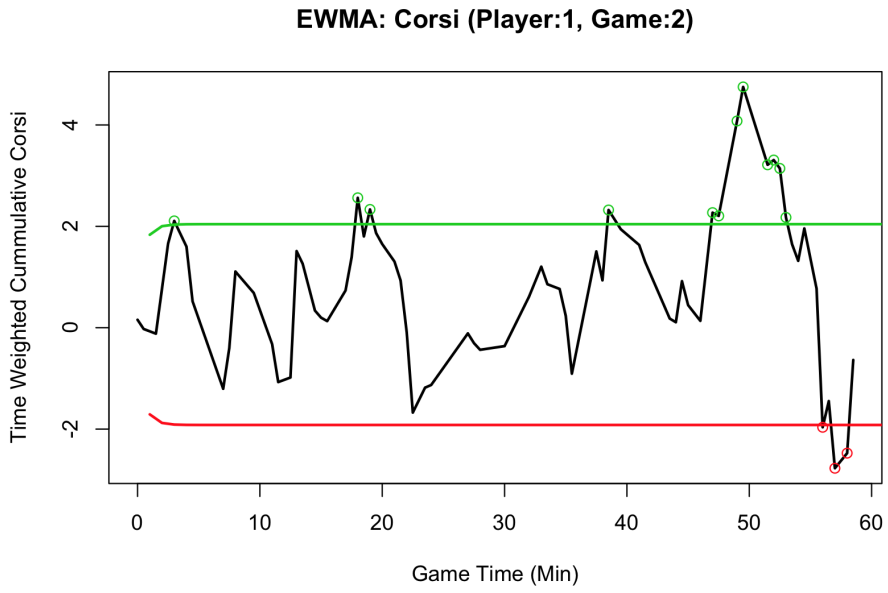

Figure 1: Example of a players $EWMA_{corsi}$ across a game with 2σ limits.

5 Prediction Future Trends in Game Play

Another tool we can use to aid in-game decision making along with player monitoring, is to provide predictions of future trends within the key metrics of game-play. The ebb and flow of every hockey game is different with many unique situations presenting in each one. The objective would not be to predict these situations but to model the controllable effects and attempt to determine if temporal correlation exists in the resulting residuals. If we can model the residuals with a predictive model we can determine potential future values or ranges for key statistics and use the known controllable effects in the model to maintain a statistically advantageous position in the game.

After identifying the optimal line combination for the game we can use this same pool of variables to model the variable of interest. In addition we may want to include uncontrollable variables like opponents, penalties, game location to further explain the outcome of the game. Once we have selected the ideal model for the metric,

which could be achieved in any reasonable manner (stepwise, stagewise, etc.), we will investigate the residuals from this model. Based on the design of our data by organizing the variables into contiguous blocks of 30 seconds, we have a time series from our residuals.

To model the temporal effects in our residuals we will use Thomson's Fourier inverse method for prediction as described in [site section]. Using the residuals from the data earlier in the game we can predict future events by zero-padding (adding zeros to the end of the time series), windowed Fourier transforming the data, regressing Fourier transformed sinusoids onto the transformed data, identifying the sinusoids of interest and inverse Fourier transforming their coefficients. This gives us an estimate of the periodic trends in the residuals and a prediction for the zero-padded part of the series. By sampling the residuals from this model, replacing the zero-padded end of the time series with these samples and computing the Thomson predictions again, we will get an alternate a varying choice for the predicted section. If we repeat the re-sampling a significant number of times we will have a well defined distribution on predicted values.

Using this distribution for the predicted values we can give with relative certainty how the flow of the game will go outside of player choices and uncontrollable variables. We can use this to ensure we play the right people for the situation. For example, when the other team is expected to play poorly, you may put on strong offensive players to try and exploit their diminished play. As the choices a team makes will effect the future outcomes, it becomes important to update the predictions often.

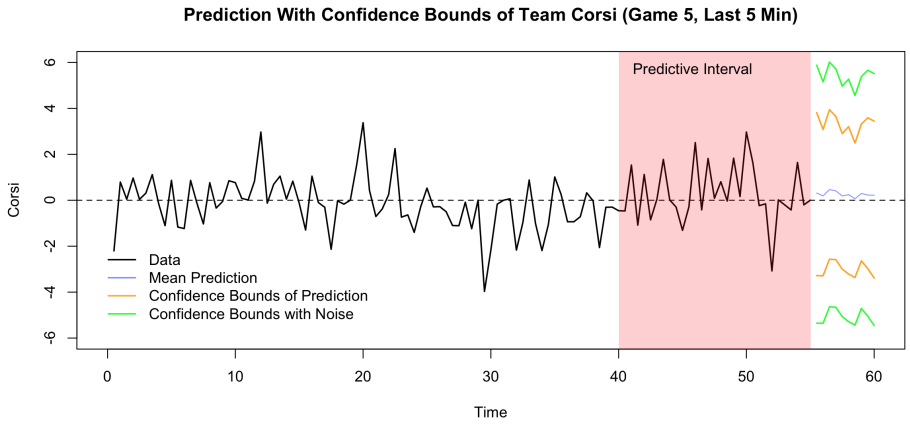

Figure 2: Example of the Prediction of the Final 5 Minutes of Game-play.

6 Data Analysis: Kingston Frontenacs

FILL IN

7 Conclusions and Discussion

As we have shown, there are many ways to improve the decision making in hockey through the use of Statistics. ADD MORE