

Developing an Analytics Program for Major Junior Hockey

Joshua Pohlkamp-Hartt & David Riegert

Queen's University

August 11, 2015

Motivation

*Maple Leafs shake up front office, hire **stats guru** Kyle Dubas, 28, as assistant GM* (Kevin McGran, Toronto Star)

There is no statistic to accurately quantify neutral-zone play. (Steve Simmons, Toronto Star)

Data Collection - Overview

- Tried recording on paper - Impossible for velocity of data!
- Developed JAVA program to meet speed of the game.
- Hired/trained stats undergrads to collect the data.
- Home games are easy but away games are difficult.
- Including scoring and penalty data through web scraping.

Data Collection - Data Recorded

We had three data collection roles: shots, neutral zone play, shifts. The data recorded was:

- Shots: team, type (blocked, missed net, goal, save), and time.
- Neutral Zone: outcome, team, and time.
- Shifts: player, on/off, time.

Data Collection - Java

Example of Java program shots page.

The screenshot shows a web interface for recording hockey shots. At the top, there are three tabs: "Shots" (selected), "Neutral Zone", and "Shift Change". Below the tabs, there is a dropdown menu currently showing "Frontenacs". To the right of the dropdown is the label "Other". Below the dropdown and "Other" are two columns of buttons. The left column contains buttons for "MISS", "GOAL", and "BLOCK". The right column contains buttons for "MISS", "GOAL", and "BLOCK". In the center, between the two columns, there are two "SAVE" buttons.

Data Collection - Home vs Away Games

For home games we were able to watch from the press box. This gave us a full field of view which allowed us to easily monitor all of the play.

For away games we had to watch the scouting film provided by the other team. These films were generally of low quality. An additional issue is the inability to monitor shift changes easily. To avoid collection issues, repeated viewings and strict monitoring of line combinations was needed.

Data Collection - Other Data Sources

In addition to the data we recorded at games we collected information on penalties and scoring.

To do this we scraped the data from the OHL website using the `rvest` package in R.

Data Analysis - Simple Data

The first things we were able to report were simple statistics that were directly calculated from the available data. Some of these statistics were:

- Shooting: Shot Attempts, Corsi, On Net %, Shooting %, and Even Strength Goals per 60 minutes
- Special Teams: Power Play %, Penalty Kill %, and Power Play Goals per 2 minutes.
- Neutral Zone: Offensive Neutral Zone Success, Defensive Neutral Zone Success

All of these stats were broken down into periods and situation. This data was exported to .xls files for the team. We also gave breakdowns for the last 5 and 10 games to show more current performance.

Data Analysis - Estimated Neutral Zone Differential Formula

To give an overall assessment of neutral zone play we developed a statistic to take into account both a player's offensive and defensive neutral zone skills. Known as the Estimated Neutral Zone Differential (END), we referenced a player's success rate to the average rate for similar players. For a player X and reference group G , we have

$$END(X, G) = (\hat{p}_O(X) - \hat{p}_O(G)) + (\hat{p}_D(X) - \hat{p}_D(G)). \quad (1)$$

The group, G , of players referenced against could be other teammates that play the same position (forward or defense) or league wide. This choice is dependent on what you would like to investigate.

Data Analysis - Estimated Neutral Zone Differential Analysis

To examine if the *END* statistic was a good metric of player quality, we examined the relationship between where in the lineup a player would play and their *END* statistic relative to the team.

We modeled the probability of a player being on the top two lines given the p-value for that player's *END* being positive. The fitted model for this past season's data was $\hat{P}(\text{Top 2 lines}) = L(.59 - 1.59 \times \text{p-value})$ with p-value on the significance for the coefficient being .0947.

This indicates that as a player is more likely to have an *END* value that is not positive, he is more likely to be a bottom end player.

Data Analysis - Line Optimization Formulation

The next thing that we thought would be useful for the team was a way to identify strong line combinations. To do this we examined modeled players and their lines (2 or 3 term interactions) with the probability of a goal being scored in the next minute. We performed this for goals for and against separately then evaluated the difference in cumulative pessimistic confidence bounds on the coefficients of our model. That is for a line combination we get the metric,

$$\begin{aligned}
 \Delta_{\alpha} = & \left(\sum_{i=1}^{18} (\beta_{i,gf} + \Phi(\alpha)S(\beta_{i,gf})) \right) + \sum_{j=1}^3 (\gamma_{j,gf} + \Phi(\alpha)S(\gamma_{j,gf})) \\
 & + \sum_{k=1}^4 (\zeta_{k,gf} + \Phi(\alpha)S(\zeta_{k,gf})) - \left(\sum_{i=1}^{18} (\beta_{i,ga} + \Phi(1-\alpha)S(\beta_{i,ga})) \right) \\
 & + \sum_{j=1}^3 (\gamma_{j,ga} + \Phi(1-\alpha)S(\gamma_{j,ga})) + \sum_{k=1}^4 (\zeta_{k,ga} + \Phi(1-\alpha)S(\zeta_{k,ga})).
 \end{aligned} \tag{2}$$

Where α is how pessimistic we are.

Data Analysis - Line Optimization Application

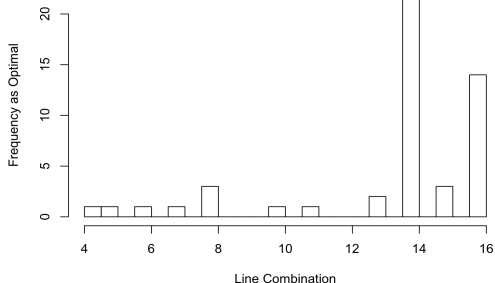
To apply this metric and we start by identifying the line combinations we are interested in using.

Then to avoid issues with unequal use of line combinations in the data, we employed a bagging algorithm. For each sampling from our data we selected the line combination with the largest Δ_α . Then we reported the optimal line combination to the combination selected most often across the samples.

Data Analysis - Line Optimization Example

We wanted to determine the optimal line combination for use in the second round of the playoffs from the data collected in the last 10 games of the season and first round of the playoffs.

Histogram of Line Combination Choices Alpha: 0.1



The choice of α will alter the selection. We often used $\alpha = .1$.

Data Analysis - Player Monitoring Formulation

To provide the coaches with a way to monitor and evaluate player performance within games we used exponentially weighted moving average quality control charts. For a player's Corsi score we have,

$$EWMA_{Corsi}(t) = \sum_{i=1}^t \lambda(1 - \lambda)^{i-1} \frac{T - Corsi(i)}{T}. \quad (3)$$

T being the target value for a player, this could be their season average or a value you expect the player to reach.

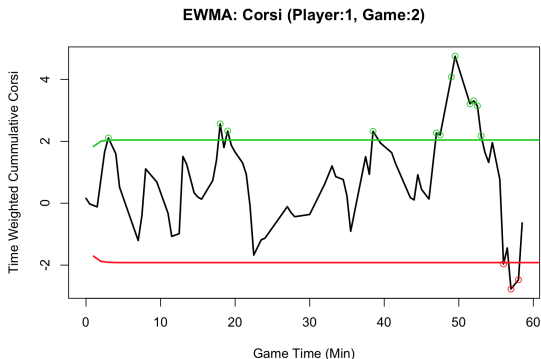
Then under the null hypothesis that $H_0 : \mu_{Corsi} = T$, we have

$$EWMA_{Corsi}(t) \sim N(0, \sqrt{\frac{\lambda}{2 - \lambda} [1 - (1 - \lambda)^{2t}]}). \quad (4)$$

We can then set significance bounds on this statistic and identify when a player is performing significantly different from their target value.

Data Analysis - Player Monitoring Offline Application

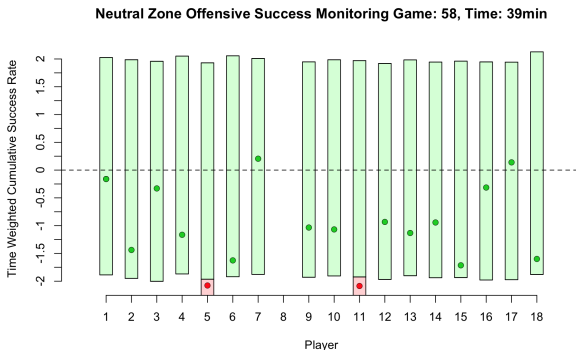
We were able to apply this method in two ways, the first method is to give summaries of the players' performance retrospectively.



We can see this player outperformed his average play several times throughout the game and near the end under performed.

Data Analysis - Player Monitoring Online Application

We can also use this method to give the status of the current entire team simultaneously. This is useful as an in-game reporting tool.



We can see here that players 5 and 11 are under performing at this time with respect to their average.

Data Analysis - Modeling and Prediction Formulation

The last method we developed was to model player effects on different aspects of game play and attempt to predict the residual unexplained temporal correlations.

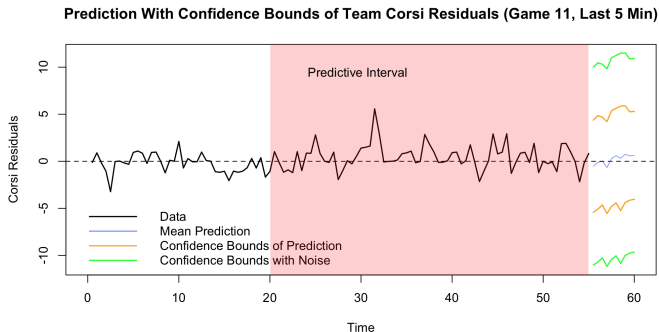
For this we used LASSO regression to identify the optimal model and then used David Thomson's periodic reconstruction method to predict the upcoming residuals.

Thomson's method models the residuals in the frequency domain with sinusoids and then can produce future values through the use of zero-padding.

This method can be used to identify upcoming trends in game play, which can be used with the player effects given from the LASSO model to make personnel decisions for upcoming shifts.

Data Analysis - Modeling and Prediction Example

Here we wanted to model the final 5 minutes of puck possession using the data from the 2nd and 3rd periods.



Care must be taken in selecting the prediction interval. Selecting an interval where stationarity does not hold will cause significant performance issues.









Concluding Remarks

- This is a very simple and easy to construct framework for data collection that can be applied to most levels of hockey.
- The methods shown here are good examples of how statistics can be used in hockey.
- We will continue to work on making hockey more quantitative by continuing our research with the Frontenacs and we hope you will to.
- Go Fronts Go!

Acknowledgments

This research is supported by the Kingston Frontenacs, Queen's University and my supervisors Dr. Takahara and Dr. Thomson.

References

-  S. Lahiri, *Resampling methods for dependent data*, vol. 14, Springer, 2003.
-  D. Montgomery, *Introduction to statistical quality control*, John Wiley & Sons, 2007.
-  T Purdy, *Shots, fenwick and corsi*, February 2011.
-  M. Schuckers and J. Curro, *Total hockey rating (thor): A comprehensive statistical rating of national hockey league forwards and defensemen based upon all on-ice events*, Proceedings of the 2013 MIT Sloan Sports Analytics Conference, 2013.
-  S. Simmons, *Why hockey's trendy advanced stats are a numbers game*, May 2014.
-  R. Tibshirani T. Hastie and J. Friedman, *The elements of statistical learning*, vol. 1, Springer, 2001.
-  D.J. Thomson, *Quadratic-inverse spectrum estimates: applications to paleoclimatology*, Philosophical Transactions: Physical Sciences and Engineering **332** (1990), no. 1627, 536–597.
-  R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological) **58** (1996), no. 1, 267–288.