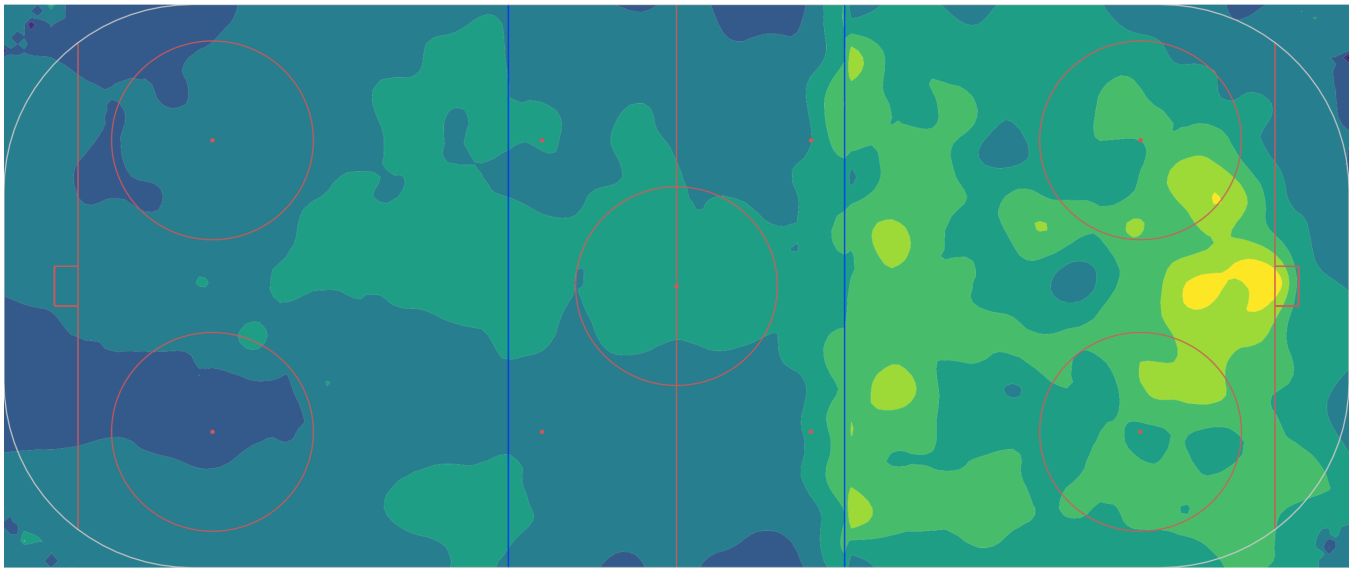
Improving the Condition of the Puck 2

Revisiting the 2022 Big Data Cup

Anders Lempia

May 23, 2023



Introduction

My submission to the 2022 Big Data Cup was named Runner-Up in the Student category. In it, I sought to measure true contribution to play driving numbers. Briefly, I used simple multiple linear regression, including interaction terms, to determine the expected value of a puck touch, where the target variable was the change in the rate of expected goal differential from before and after the touch. I called the resulting metric TVAA, for Touch Value Above Average. For this project, I attempt to apply better techniques and concepts to achieve the same goal.

Certain players in hockey are described as "play-drivers". These are players who post quality offensive and/or defensive on-ice metrics, such as simple Expected Goals For Percentage (xGF%). Something public hockey analysts are very aware of is that these numbers can be skewed by certain play styles. Brady Tkachuk is an easy example in the NHL. He posts excellent offensive play-driving numbers, but these results are to be taken with a grain of salt: Tkachuk has a penchant for jamming pucks into goalies' pads at close range — low-percentage shots that turn up high xG values in public models — which cranks up his on ice xGF numbers. This project seeks to provide a way to identify these players more easily, along with providing credit where it is due to players who directly "improve the condition of the puck", rather than those who might passively benefit from linemates with such a skillset.

This iteration focuses much more on evaluation of players, rather than identifying key features. It is constructed to be completely position-independent, where plays are measured according to their location on the ice surface, where

TVAA was significantly influenced by a players' degree of involvement in offense. And finally, I included measures of uncertainty stemming from each step of the process — another element which was excluded last year.

Methods

I used data from a wider variety of sources, drawing on each of the available files on the Big Data Cup github page. It includes data from the 2018 and 2022 Women's Olympic competitions, NCAA Women's games, and pre-PHF NWHL games. All told, there are over 30 games in the combined data.

The steps of the process, explained in greater detail below, are generally as follows:

- Part 1: Modeling ΔxG
  - Finding the expected value of an individual puck touch, in terms of xG per minute.
- Part 2: Measuring Touch Value Above Expected & Touch Value Created
  - Finding the value of an individual puck touch, relative to some baseline.

Part 1: Modeling ΔxG

This analysis is restricted to 5v5 play. For simplicity's sake, no attempt was made to construct an equivalency model. I built my expected goals model with the `randomForest` package.

The target variable, to which all evaluation metrics make reference is defined as:

$$\Delta xG = (xGF_{after} - xGA_{after})/minute - (xGF_{prior} - xGA_{prior})/minute$$

- This is calculated for all puck touch events, including shots. (excluding FOWs & Penalties)
- The subscript "after" indicates the expected goals occurring after the event and before the next faceoff event.
- The subscript "prior" indicates the expected goals occurring before the event and after the prior faceoff event.
- The xG of a shot is included in xGF_{after}.

A unique model was built for each type of touch, using `gam()` from the `mgcv` package for each, fitting smoothing splines for continuous variables. Last year's attempt used linear modeling for the sake of interpretability, but I made the switch to smoothing splines in order to better capture the relationship between ΔxG and independent variables, as this may not be a linear relationship. This is most obviously demonstrated by the idea that shots are more valuable closer to the goal, but only to the goal line, beyond which shots are coming from behind the goal. A simple linear interaction between the x-coordinate and the offensive zone is thus inappropriate.

Unsuccessful Zone Entries were removed, as not all games in the data included them. Dumped & Played Zone Entries were removed, as they were recorded twice. Each observation was weighted according to the following expression, before being normalized by dividing by the mean:

$$Weight = \frac{t_{after} + 1}{|t_{after} - t_{prior}| + 1}.$$

A brief overview of the terms involved in these models:

- Dump In/Out
  - Location details, location of player's acquisition of the puck, details regarding the prior event, and the target of the dump, proxied by the location of the following puck recovery.
- Pass
  - Location details of both release and target, location of player's acquisition of the puck, direction and distance, details regarding the prior event, and the success of the pass.
- Puck Recovery
  - Location of the recovery, and details regarding the prior event, including the team relinquishing the puck.
- Takeaway
  - Location details and prior event details.
- Zone Entry
  - Location details and prior event details.
- Shot
  - Expected Goal value, pre-shot movement, location details, and information regarding the shot type.

I included xG in the estimation of ΔxG for shots because the value of the shot stemming from goal probability should certainly be considered when evaluating the decision to shoot, with the rest of the output measuring how the play is expected to continue after the shot, considering things like the likelihood of rebounds. The exclusion of the outcome of dump plays versus the inclusion of pass success is also a change from last year. Without tracking data, it is next to impossible to measure the probability of pass success, so I instead make the not-unreasonable assumption that the conditions for which a pass will fail are identifiable prior to the attempt being made by a player, or failing that, the failure is the result of an error in the release of the pass. Thus, a player should in most cases be reasonably punished for failing to complete a pass.

Alternatively, this is not the case for dumps. When dumping the puck, a player is generally willfully relinquishing possession, and the likelihood of a recovery by a teammate is most often independent from the player who dumps the puck. The inclusion of dump target location is meant to bake as much of her influence as possible into the model.

I "trimmed the ends" off some of the predictions generated by these models due to some edge case overfitting. For example, a very small portion of Puck Recoveries were estimated to boost ΔxG by over 18. My solution was, when predictions exhibited such an extreme tail, to replace the estimates of the most extreme 2.5% with a random number generated from the distribution of the next 2.5%, i.e. 95% to 97.5%.

Part 2: Measuring Touch Value Above Expected

In order to measure the value of a puck touch, it is imperative to be conscious of the different types of touches. I broke them up into four categories: decisions (Passes, Zone Entries, Dumps, and Shots), free pucks (Puck Recoveries), takeaways (Takeaways), and shot deflections (Shots labelled "Deflection"). I define *Total Touch Value* (TTV) to be the sum of *Touch Value Above Expected* (TVax) generated from decision events and *Touch Value Created* (TVc) generated from the other — opportunistic — touch types.

Touch Value Above Expected: Decision Events

These are events that occur when a player has clear possession of the puck. She can choose to execute any of them at any time or location. I define *TVax* as

$$TVax = \hat{\Delta}xG - \overline{\mu_{n_{(x,y)}}},$$

Where $\hat{\Delta}xG$ is the model estimate of an event and $\mu_{n_{(x,y)}}$ is the mean of a Bayesian posterior distribution at each pair of X- and Y-coordinates:

$$p(\theta|\sigma^2, \hat{\Delta}xG_1, \dots, \hat{\Delta}xG_n) \propto p(\theta|\sigma^2)p(\hat{\Delta}xG_1, \dots, \hat{\Delta}xG_n|\theta, \sigma^2),$$
$$p(\theta|\sigma^2) \sim N(\mu_0 = 0, \tau_0^2 = 0.0400)\ , \sigma^2 = 0.0225,$$
$$\mu_{n_{(x,y)}} = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n}{\sigma^2}\overline{\hat{\Delta}xG}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}},$$

Where n is the number of observed events at (x, y) and $\overline{\mu_{n_{(x,y)}}}$ is calculated to smooth the distribution of expected touch values at given locations of the ice:

$$\overline{\mu_{n_{(x,y)}}} = \sum_{i=x-5}^{x+5}\sum_{j=y-5}^{y+5} p(i,j)\mu_{n(i,j)},$$

Where, to ensure total probability was equal to one:

$$p(i,j) \sim \frac{\text{Bivariate Normal}(\vec{\mu} = \langle x, y\rangle, \vec{\sigma} = \langle 2.5, 2.5\rangle, \rho = 0)}{\sum_{i=x-5}^{x+5}\sum_{j=y-5}^{y+5}\text{Bivariate Normal}(\vec{\mu} = \langle x, y\rangle, \vec{\sigma} = \langle 2.5, 2.5\rangle, \rho = 0)}.$$

It is important to note that I treated the offensive blue line as an impenetrable wall, so events on the offensive side of the line would not be evaluated according to events in the neutral zone, and vice versa. The defensive line is permeable to the team with possession, so it did not receive the same treatment.

Touch Value Created

Free Pucks To evaluate Puck Recoveries appropriately, the above method would obviously be erroneous. The conditions required for the occurrence of a puck recovery are simply that the puck be "free" — that no player be in possession of the puck. Thus, I made the assumption that, should a player have failed to make a puck recovery, the outcome would instead be a puck recovery for the opponent at the same location. Thus the calculation of Touch Value Created by Puck Recoveries, or TVc_{PR} is:

$$TVc_{PR} = \hat{\Delta}xG + \hat{\Delta}xG',$$

Where $\hat{\Delta}xG'$ is the model estimate of value the opponent would have created for themselves, had they recovered the puck instead.

Takeaways Takeaways are a different case, because there is a less clear baseline case. Consider: a defender makes an attempt to take the puck from its carrier. The defender is unsuccessful. Why? In most cases it is because the puck carrier senses the pressure and moves the puck. Of course, in some cases, it is due to clever maneuvering to maintain possession without moving the puck. However, there is no method in my analysis to handle this case. Thus, I assumed that the value the opponent might generate if the takeaway failed is equal to $\overline{\mu_n}'_{(x,y)}$. Making the necessary coordinate transformation we get:

$$TVc_{TA} = \hat{\Delta}xG + \overline{\mu_n}'_{(x,y)}.$$

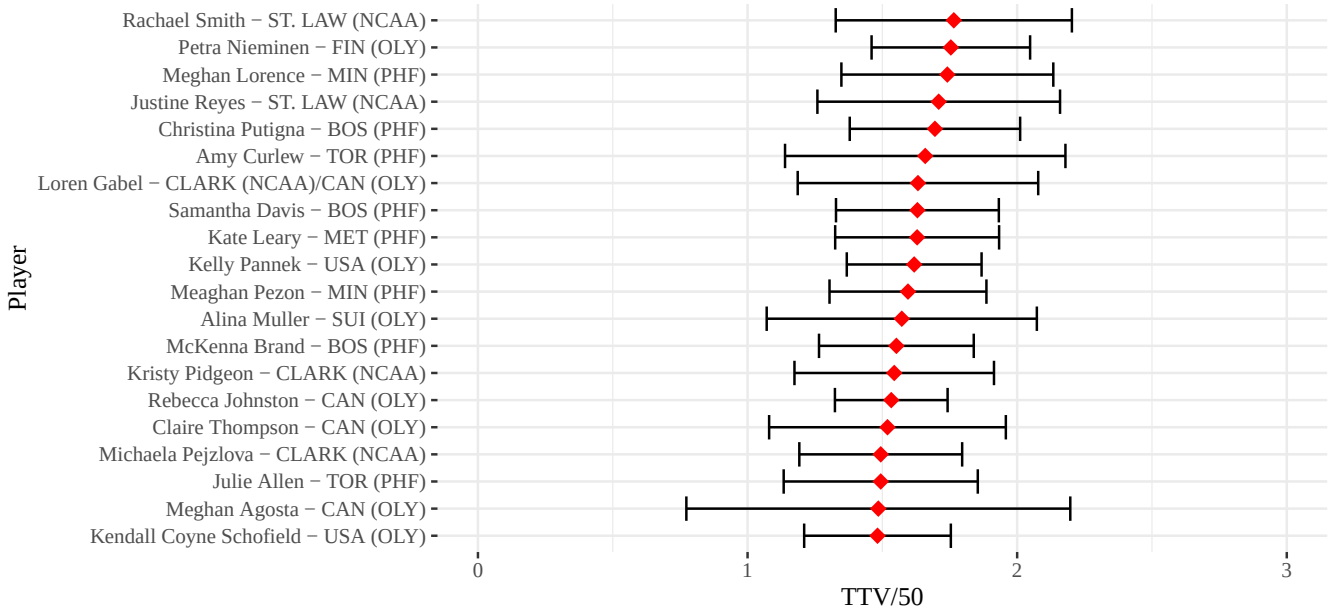
Deflected Shots Touch Value Created by Deflected Shots: $TVc_{DS} = \hat{\Delta}xG$, with no adjustment.

$$TVc = TVc_{DS} + TVc_{TA} + TVc_{PR}$$

Results and Conclusions

A look at the top 20 most efficient players in total value with at least 75 touches, including the bounds of a 95% confidence interval:

Figure 1: Top 20 Player Total Touch Efficiency (min. 75 Touches)

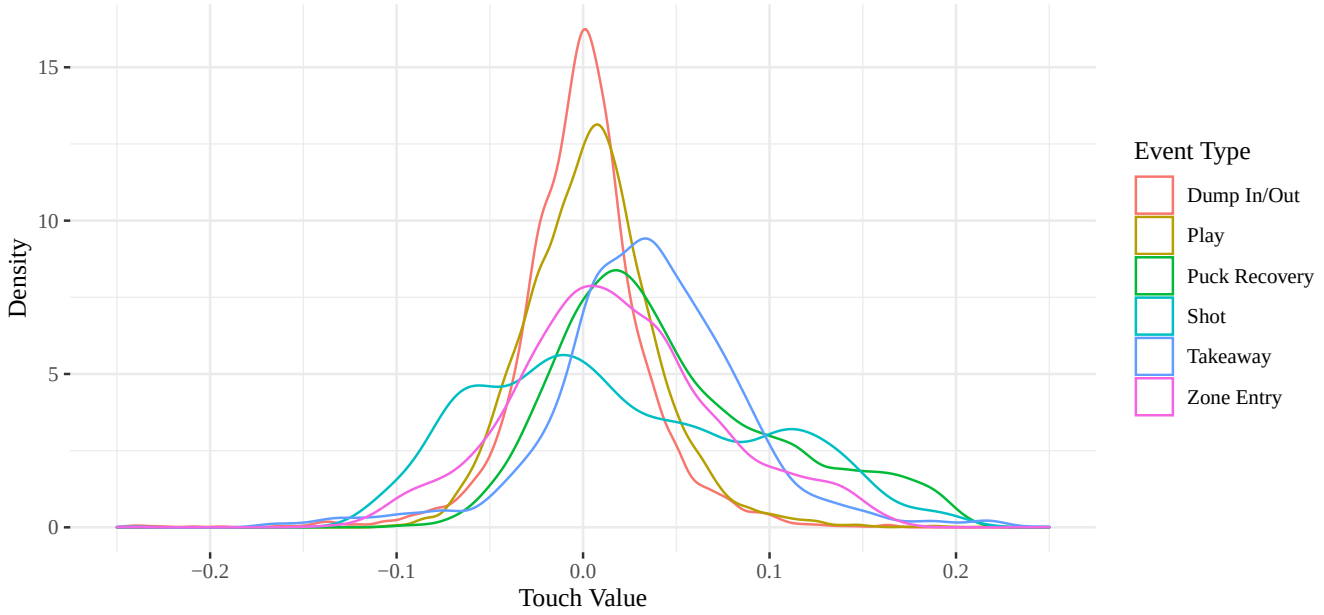


Claire Thompson, the highest rated defender by total efficiency is ranked 16th. This is much better than the previous project, where the rankings were almost perfectly striated by position.

The most obvious issue with this analysis is that TVax and TVc are measured on different scales. TVc, being value created out of thin air by collecting loose pucks or stripping an opposing puck carrier, has a wider range, stemming from a much lower baseline than TVax, which is measured versus the expected value of a play at that point on the ice. As a result, TVc efficiency carries a 0.84 correlation with TTV efficiency (this is why Marie-Philip Poulin is notably absent). They should generally be treated separately.

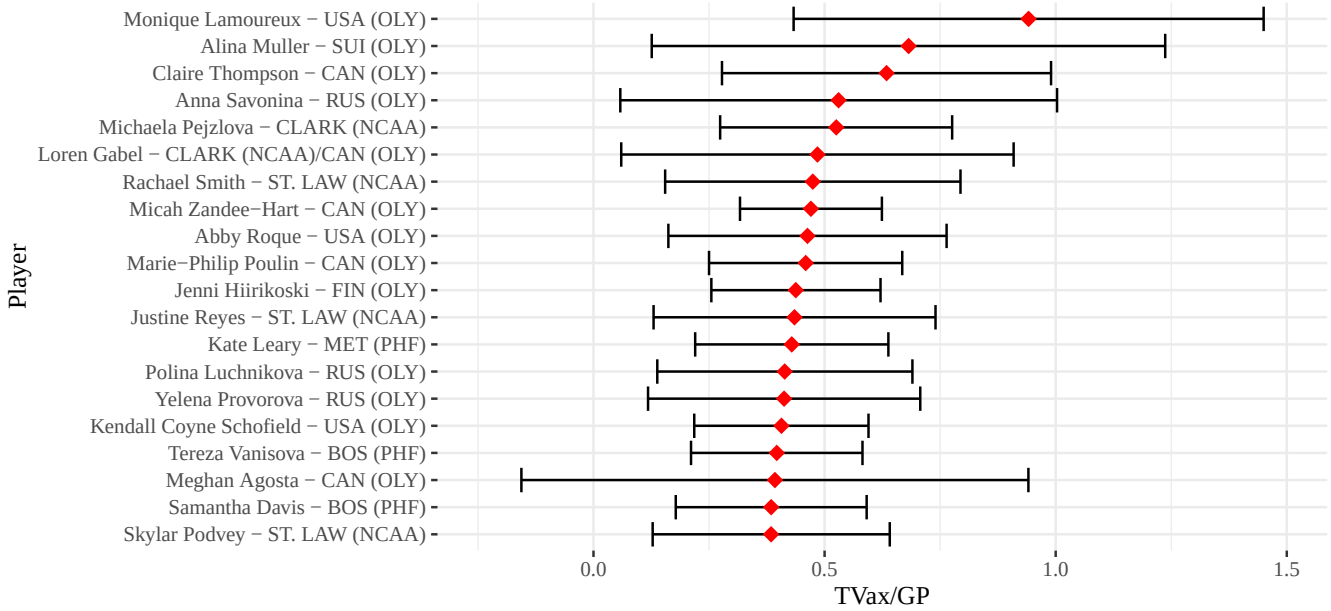
A rough look at the distribution of added Touch Value from each event:

Figure 2: Distributions of Touch Value by Event Type



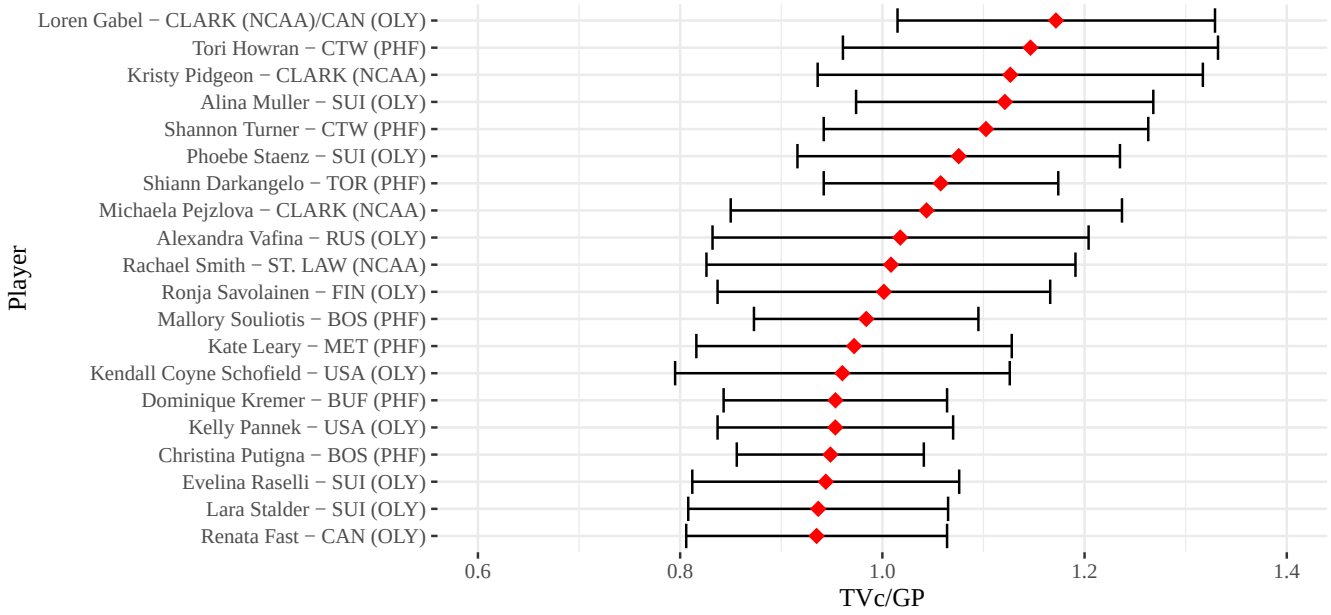
Note the wide distribution of shot values, as well as the significant right tails of puck recoveries and takeaways. This does not conflict with last year's finding that takeaways are only made important by the next play, thanks to the different baseline measure. Figures 3 and 4 repeat Figure 1, for TVax and TVc separately. For these, I use "per game" designations, to showcase which players generate the most overall value.

Figure 3: Top 20 Players in Per−Game TVax



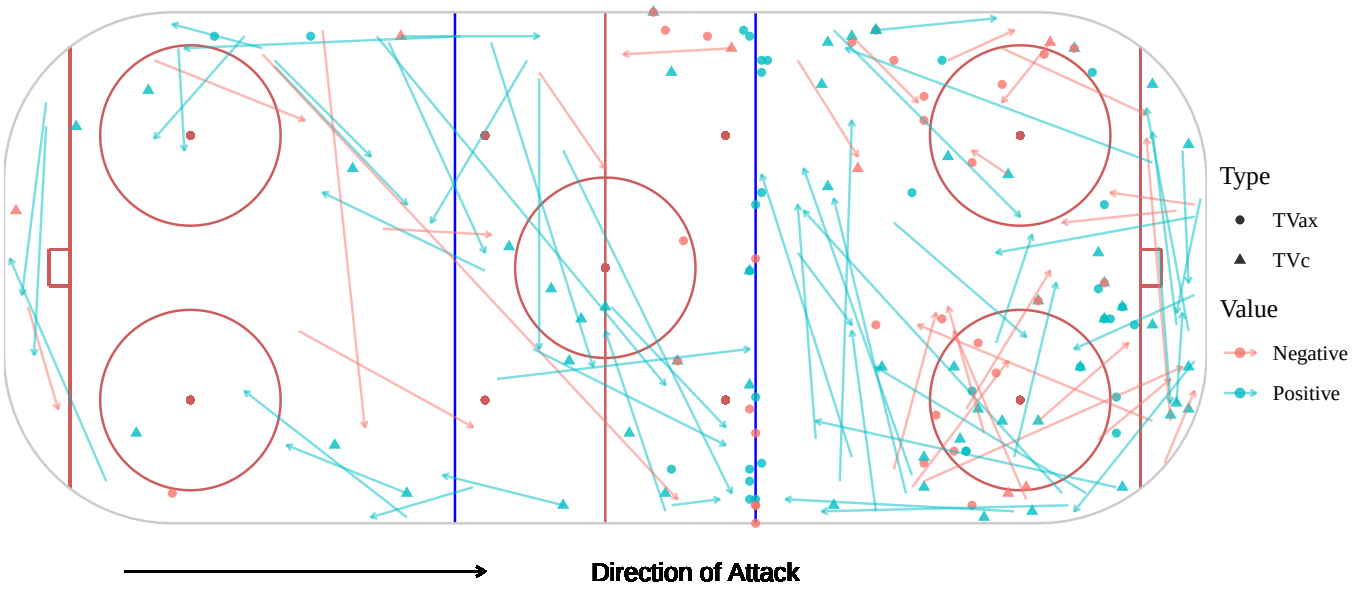
Thompson, Zandee-Hart, Savonina, Hiirikoski, Provorova, and Podvey are the six defenders in the top 20 Touch Value Above Expected producers. TVax is sufficiently position-independent, allowing defenders to be measured on the same scale as forwards. Monique Lamoureux appears in just 2 games in the data, with 69 total touches, just below the threshold to appear in Figure 1. She would be first overall in TTV/50 with any cutoff above 33.

Figure 4: Top 20 Players in Per−Game TVc



On the Touch Value Created leaderboard, Howran, Turner (Doyle), Savolainen, Souliotis, and Kremer make up the five defenders. This list is a who's-who of PHF talent, with the Boston Pride's reigning MVP leading the way:

Loren Gabel's 195 Puck Touches



Clearly, Gabel makes her living in the offensive zone. She funnels pucks to the center point from the wall, attempts and completes a fair complement of passes into the slot, and excels at retrieving the puck in the offensive zone. The recoveries make up the bulk of her success in the TVc metric. Most of her shots are lower-value, providing slightly negative TVax but she has a handful of very high danger shots near the netfront which creates a net positive.

Table 1: Loren Gabel Touch Values (4 GP)

Event	Zone	N	Touch Value	Per 10	Per Game
Puck Recovery	O	35	2.991	0.855	0.748
Puck Recovery	N	14	1.038	0.741	0.260
Pass	O	46	0.838	0.182	0.210
Shot	O	34	0.666	0.196	0.166
Zone Entry	N	20	0.550	0.275	0.138
Puck Recovery	D	8	0.438	0.547	0.110
Pass	N	15	0.187	0.125	0.047
Dump In/Out	D	3	0.043	0.143	0.011
Takeaway	O	2	0.039	0.195	0.010
Pass	D	19	0.011	0.006	0.003
Dump In/Out	N	6	-0.070	-0.117	-0.018
Dump In/Out	O	1	-0.106	-1.060	-0.026

Gabel's efficiency in creating value from offensive zone puck recoveries is the key to her success in this evaluation. Among players with 15 offensive zone recoveries in the data, Gabel ranks ninth in value per recovery and second behind Abby Roque (2 GP) in total offensive recoveries per game. Thanks to this combination of efficiency and volume, Gabel has a wide (0.056) lead over Kristy Pidgeon's second place in terms of value per game.

This is, of course, merely a sample of the kinds of analysis available with these metrics. Gabel's game has evolved in recent years, as many players' do. The games in this data are fairly outdated. Compared to last year's entry to the Big Data Cup, I feel this approach has made real strides in usefulness and accuracy. Separating shot value from other touches in all cases is a prudent step to understanding the ability and style of a player.

Further Improvements

There are always gains to be made in modeling, both by using a better xG model and of course with the event models, particularly for shots. Building a Shiny app following this data would be an interesting project as well. Re-evaluating the scales used to measure non-decision events could also be a productive change.