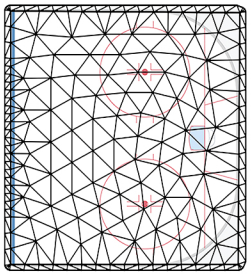
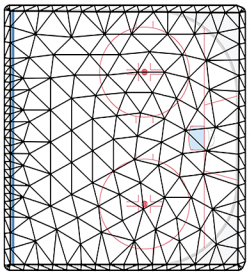
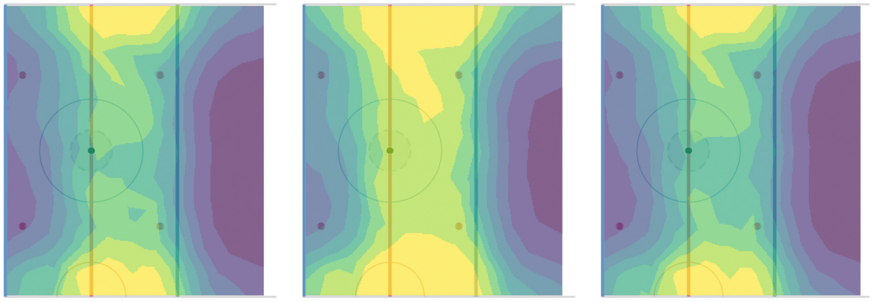


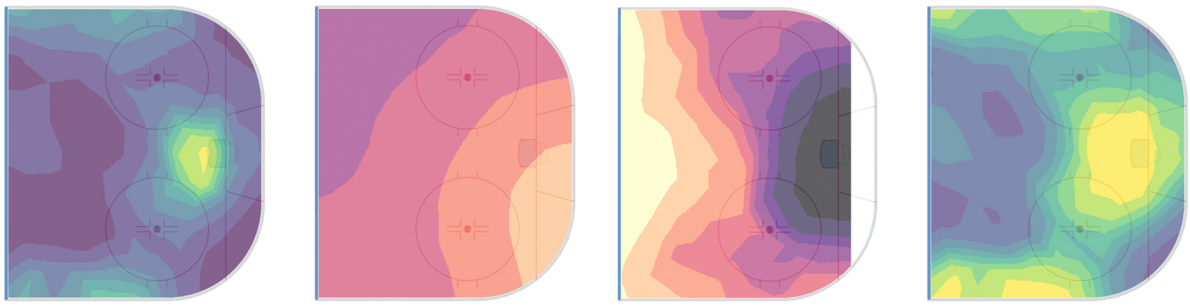
Big Data Cup 2021





Bayesian Space-Time Models for Expected Possession added Value

Brendan Kumagai, Mikael Nahabedian, Thibaud Châtel, Tyrel Stokes



1. Introduction

Hockey is a game of making the best possible decision in the shortest amount of time. Players need to react quickly to form a chain of plays to create valuable scoring chances. We build upon past work in basketball [1], soccer [2, 3], and hockey [4, 5] to quantify the value of player actions through the use of Bayesian statistical methods that can capture the fluid nature of offensive play beyond the current state of analysis that generally views plays in isolation.

Our approach adapts to match the resolution of the Stathletes data set, treating the observed play sequences as realizations from space-time stochastic processes with stopping times. This allows us to simulate realistic play sequences and among other things estimate and attribute the value of space and time in hockey.

Figure 1 provides an example of our 'Possession Added Value' (PAV) metric that we introduce in Section 2.2. to quantify the expected value of an event. Here we have an entry-to-exit sequence by the Sudbury Wolves. The PAV can be thought of as the boost in expected goals due to the decision made by the player. From a scouting perspective, analyzing each element of a sequence will help us determine if a puck touch improved or decreased the expected value of a sequence, which can be linked to individual player analysis. This project will allow scouts to move past isolated metrics and assess how much value a player is creating within sequences, on average, in a game and a season.

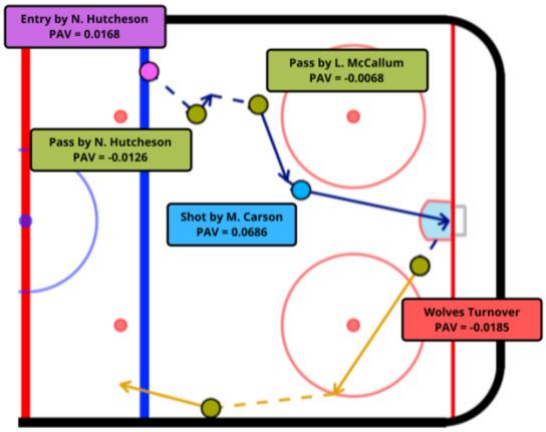


Figure 1: Entry-to-exit sequence example, Sudbury Wolves 2019-09-20.

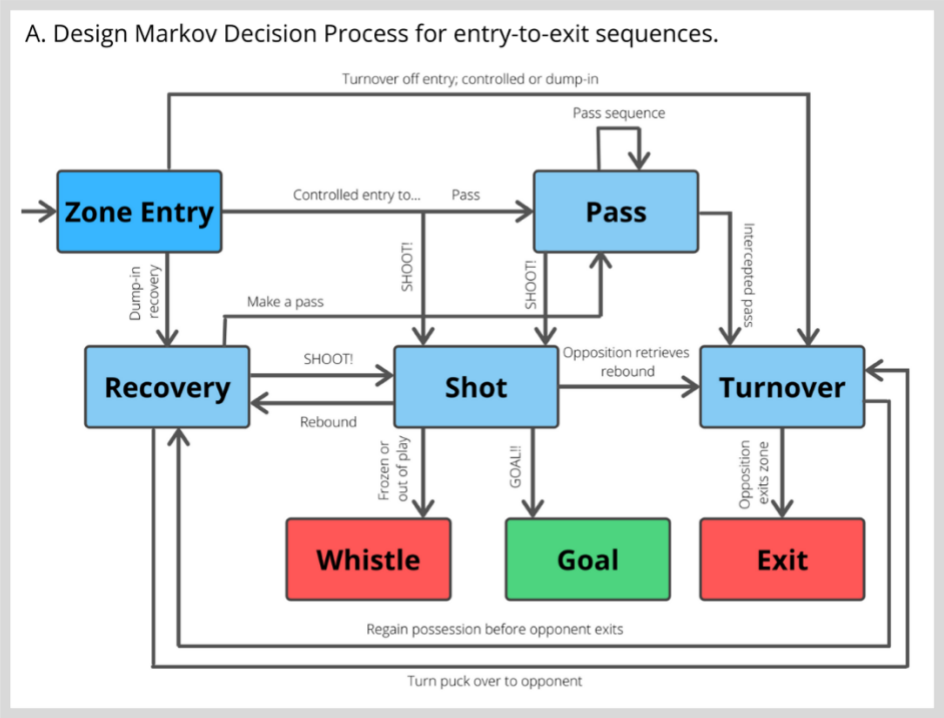
2. Methodology

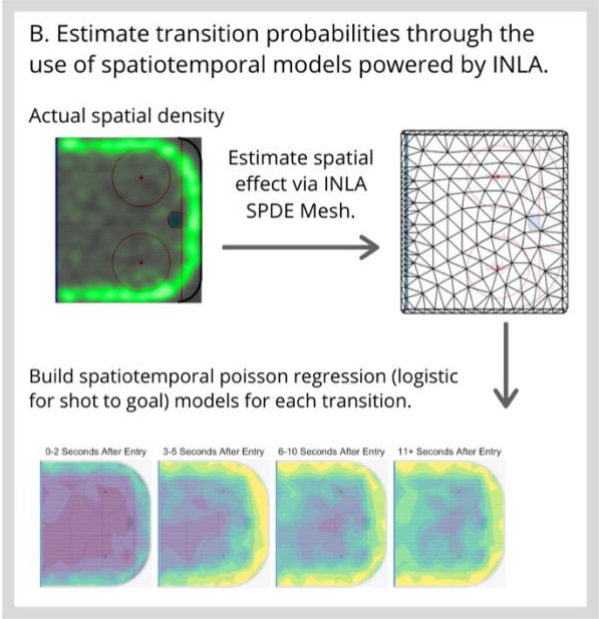
2.1. Model

In order to construct our player metrics and compare them fairly to a baseline, we need a model of expected value. In particular, given a location on the ice S, the time on the clock T, as well as game state variables such as score and strength state X, we want to model the expected value of the resulting offensive possession. Often, regression is the tool used for this task, however regression is not always able to capture the dynamic and sequential nature of plays and appropriately divide credit [1]. We construct a spatiotemporal Bayesian Markov transition model with stopping times at the resolution of the data. The full Markov transition model can be decomposed into 3 subcategories of models.

1) Action-Transition Model – (Similar to what Cervone et al. [1] call macro-transitions) Given the location in space, the time relative to key events (entry, last pass, last shot), the relevant covariates like score and strength state, and the previous action, these models predict the next action of the puck carrier. For example, if following a controlled entry, we find a player at the top of the left circle with the puck, this model predicts if the next action will be a Pass, a Shot, or a Turnover. We use the so-called "poisson trick" so that we may model the transition counts, rather than probabilities directly to take advantage of INLA for fitting spatiotemporal processes [6]. The count models are combined to approximately recover the posterior transition probabilities.

2) Movement and Time Models - Given the previous sequence and the most recent action transition, these models predict where the next action transition will happen in space and time. In the most elaborate example, if the previous action was a pass, this series of models predicts whether the pass is direct or indirect, then, given the pass is direct or indirect, we predict the location of the pass target. Subsequently, given the pass target, the location of the next event is predicted. Finally, given all of this movement, the amount of time elapsed between action transitions is then predicted. Time elapsed allows us to keep track of the score clock in our simulated events which serves two purposes. First, our transition models are modelled as functions of time since entry to reflect the fact that over time defensive structures change. Second, if there is no time left in the period at the end of a play, the play sequence is terminated. This helps correctly valuing plays made near the end of the period.





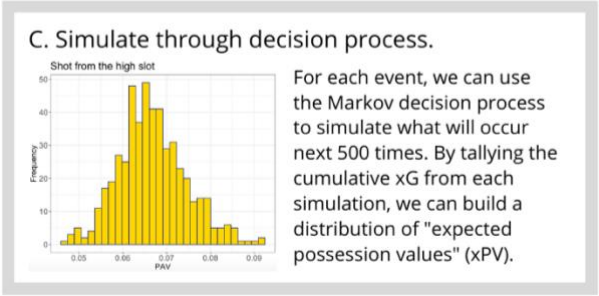


Figure 2: A brief overview of our modelling process

3) Expected Goal Models - Here we have two models which predict the expected goals from a shot. The models are split by whether or not there was a pass directly before the shot to capture the interaction between pass location, time, and shooting location.

Figure 3 illustrates a discrete representation of the expected goal model without pre-shot movement. Each panel represents a fixed time point. We see on the farthest left, immediately after entering the zone, the bright yellow is largest. This reflects the fact that immediately after entering the zone, the probability of a rush chance is high and defensive structures are likely to be the loosest and least set. At the far right, we see that after 10 seconds, the bright yellow is smaller, reflecting that defensive structure has likely been set up and dangerous shots are harder to generate.

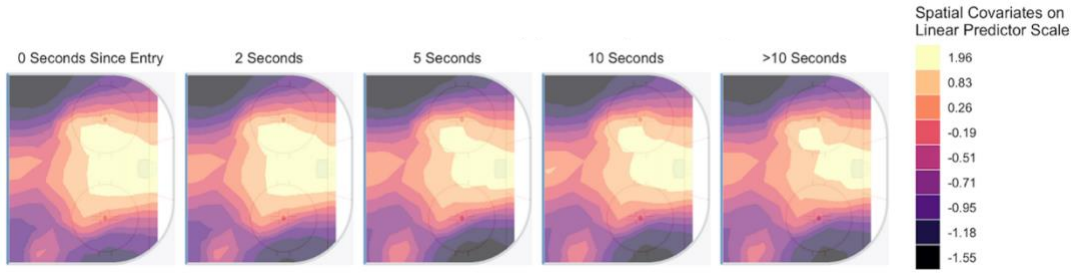


Figure 3: The spatiotemporal effect of our shot to goal without pre-shot movement model. Temporal effect represent time since entry.

All models, across all three categories belong generally to the family of space-time stochastic process models. The feature that we prioritized most is the continuous spatial nature of hockey, for example we expect the probability of a shot vs a pass vs a turnover to be different depending on the location of the ice, however we don't expect these probabilities to change abruptly across space. The typical solution to this problem in hockey has been to find a clever way to divide the ice up into sections which it is reasonable to believe are similar and then to use priors or penalties to pull the weights of nearby sections together. This, however, introduces a number of choices and trade-offs. Most importantly, by splitting the ice into small subsections we represent the continuous nature of hockey more accurately, but at the price of needing more information and often structure to estimate the coefficients precisely. The middle ground solution we opt for is a discrete approximation to a continuous process and priors which penalize model complexity described below.

1) Define a discrete Mesh - We chop the ice into unequal triangles. In areas where we expect the function to change more, we might make the triangles smaller. In our data, all locations are integer values which means that the minimum resolution of our data is 1 foot. We want to be careful not to try to fit functions which change faster than it is possible for our data to estimate. The discrete mesh defines basis functions and weights. Each point on the ice can then be represented using weighted combinations of the nearest nodes of the mesh. There is an analogous procedure to define time meshes when we fit continuous spatiotemporal models. An example of this discrete mesh can be found in Figure 2, Panel B.

2) Matérn Covariance Function - The Matérn covariance function defines how quickly or slowly the model changes over space (or time, or space-time) as well as how noisy it is. This structure allows us to share information across location and time which is crucial to our goal of accurately representing the changing nature of space in hockey with only 40 games of data.

With the discrete mesh and Matérn covariance function we can estimate a discrete projection of a continuous spatial process, essentially interpolating cleverly within the mesh. The resulting model has a finite number of parameters, but still allows us to estimate a continuous surface. For space-time models we additionally specify a process for the relationship of the spatial coefficients over time. Typically, this was done using a continuous AR(1) process.

3) Penalized Complexity Priors - The last step is putting reasonable priors on the parameters of the Matérn covariance function. The goal of the penalized complexity prior [7] is to avoid overfitting. We define a distance on the continuum between the most complicated possible model (i.e. if the probability of a pass say changed drastically for small movements on the ice), and the simplest nested model (i.e. space doesn't impact the probability of a pass at all). The priors we put say that without sufficient evidence, we will pull the model towards the simpler end of the continuum. This gives us flexibility when defining the mesh and allows the data to inform how complicated the model becomes in terms of effective number of parameters. This is extremely important if we want to treat both time and space as continuous to best reflect the fluid nature of hockey.

4) Fit the Model - All models were fit using Integrated Nested Laplace Approximation (INLA) [8] in R (R-INLA package). The advantage of using INLA is two-fold. First, they allow Bayesian sampling of a class of spatial models. Second, the approximations used are highly efficient and scale well. A full Bayesian set-up allows us to properly propagate the uncertainty across the different models with ease, as well as use the penalized complexity priors to avoid overfitting. The computational scalability of these models means that much of this framework could be implemented with data from even more games or with higher resolution data such as player tracking [1]. Scaling the posterior sampling of full play sequences beyond the needs of this project proved to be much more challenging, but we made progress with potentially suitable approximations with scaling capacity.

2.2. Possession Added Value (PAV)

For any observation, n, we get the expected goals given the associated space, time and covariates (xG_{S_n,T_n,X_n}) by sampling from our model posteriors and simulating play sequences subject to the stopping rules of whistles, zone exits, and the period ending. We use a Rao-Blackwellization step for goals to reduce sample variance. Sampling enough chains for a particular observation allows us to estimate the probability of a goal for any observation in the data set. In the spirit of previous work done in this area, we call the expected goals from our posterior samples as Expected Possession Value (xPV).

Using the xPV in the moment prior to an event as well as the xPV of the subsequent event, we can define a metric for the added value to a given possession, which we have dubbed Possession Added Value (PAV). The formula for PAV can be defined as:

$$PAV_n = xG_{S_n,T_n,X_n} + xPV_{S_{n+1},T_{n+1},X_{n+1}}\left(1 - xG_{S_n,T_n,X_n}\right) - xPV_{S_n,T_n,X_n}$$

Where xG_{S_n,T_n,X_n} is the expected goals given that the actual event is a shot or 0 otherwise, $xPV_{S_{n+1},T_{n+1},X_{n+1}}$ is the expected possession value of the subsequent event, and xPV_{S_n,T_n,X_n} is the expected possession value at the moment prior to the execution of the event. If a player is able to get the puck into more valuable space-time through their actions, then this value will be positive. Otherwise, it will be negative.

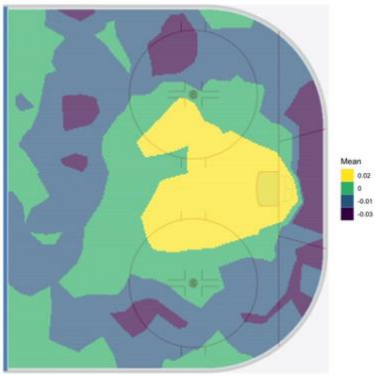
Adding in the $\left(1 - xG_{S_n,T_n,X_n}\right)$ multiplier to $xPV_{S_{n+1},T_{n+1},X_{n+1}}$ allows the formula to better reflect the total probability of scoring in the sequence by reducing the xPV to account for the chance that the shot actually goes in. This treats expected goals as a proper probability and accounts for the fact that the sequence only continues if no goals have been scored up until that part of the sequence.

For shots that end in whistles or goals, this step additionally requires simulating where the puck might have done in the event a goal or whistle did not occur and a sampling procedure to determine the values of those possible locations. These steps are done to preserve the interpretation of our metric as a probability and to be directly comparable to the posterior simulated sequences. This allows us to talk about the additional expected goals generated by different players.

You will notice that compared to current expected threat models in hockey [4], we are missing the downside risk part of value, that is the probability of the other team scoring. Given the short timeline of this project, we prioritized getting this model functioning for just the offensive zone, but the groundwork has been laid to incorporate defensive zone events and downside risk in future work. This model as it stands does not sufficiently value defensive play and tends to punish what appear to be conservative plays such as passes around the outside. This is partially a result of not explicitly modelling the defensive side of the puck.

2.3 General Findings

Figure 4: Average PAV by spatial coordinates



The PAV metric helps us assess a player's overall contribution within offensive sequences and break down his offensive impact for each of the following events: entries, passes, shots, turnovers and recoveries. Despite the fact that our dataset is limited in terms of the number of observations for players other than those of the Otters, it is interesting to see some general ideas emerge.

Figure 4 displays the spatial distribution of the average PAV in the offensive zone, serving as a good reminder about the location of high danger areas on the ice - highlighted in lighter colours - ensuring that our model is consistent with previous hockey research.

For the most part, players draw positive PAVs from zone entries. This is in line with the foundations of our model as a zone entry enables a team to move into the offensive zone, providing them, in most cases, with a more favourable position to generate offence. Even if a dump-in is historically half as productive as a controlled entry, it still moves the needle upwards, going from outside of the offensive zone to the possibility of a shot. Similarly, players generally draw positive PAVs from puck recoveries, as the recovery of the puck is tied to the possibility of generating offence. On the other hand, all players get a negative PAV from turnovers. Finally, about 95% of players receive a positive PAV from shots; these rare cases can be interpreted as a player taking an abundance of lower-danger shots when there are generally more beneficial options available.

Interestingly, only 15% of the players in our dataset have a positive PAV from passing plays. In general, traditional hockey fans assess the playmaking ability of players by focusing on a few game-changing passes. However, our model indicates that a pass is generally considered to be a lateral move in terms of increasing value. In fact, the value of a pass is slightly negative on average, likely due to the possibility of turning the puck over upon release.

Figure 5 shows that offensive zone passes are heavily concentrated along the boards. In both cases in our model where we 'transition' to a pass next, we see an extremely high density of perimeter events. In summary, our model treats the high volume of passes as less favourable events. Nevertheless, there might be some long-term benefits in prolonging puck possession by passing to less favourable positions, depending on the circumstances. However, without player movement data, our model can only capture proxies of the opposing team's defensive structure. Therefore, our model cannot fully differentiate a case where a player willingly passes the puck to a lower value part of the ice, yielding a certain

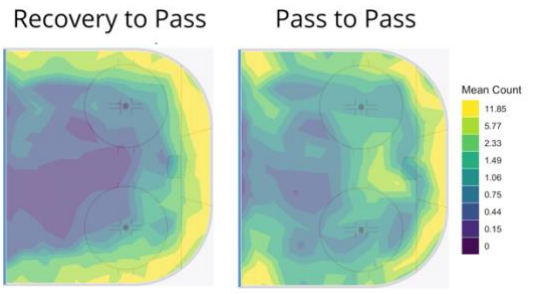
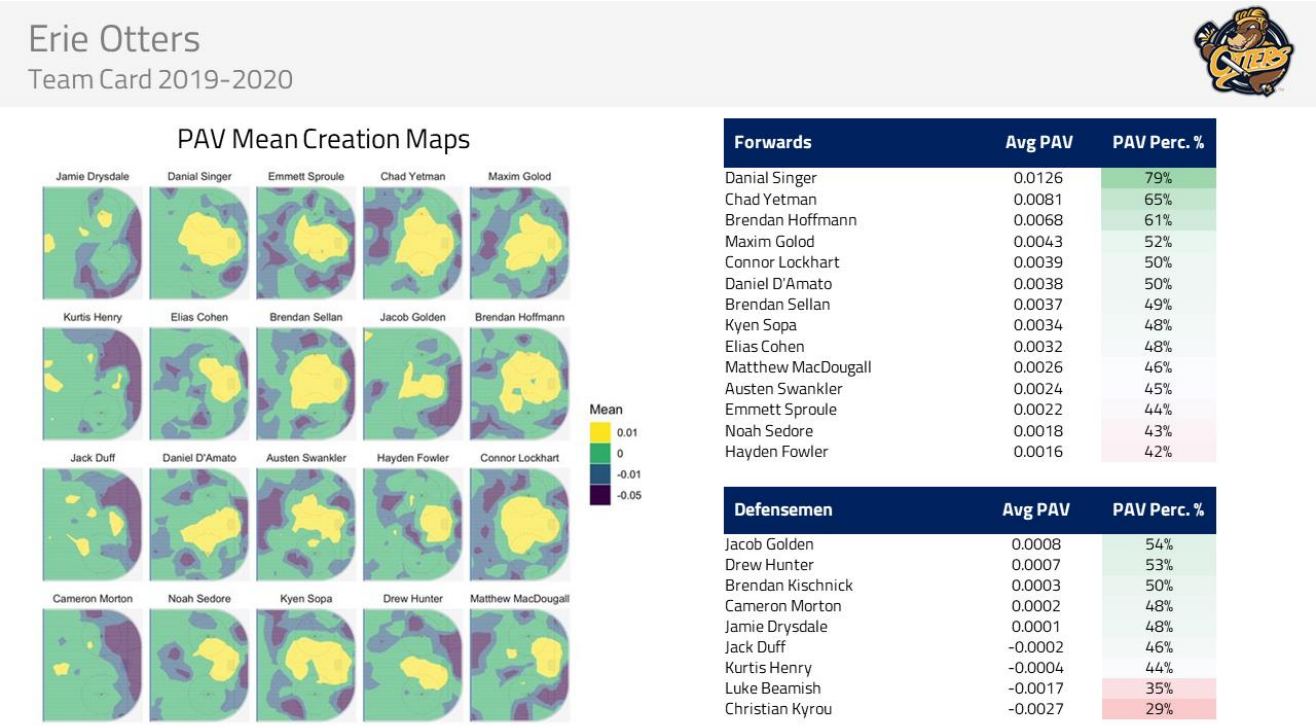


Figure 5: Pass reception density prior to making an additional pass at ES.

tactical advantage, from a situation where a pass truly hinders the ability of the team to generate offence given the poor decision of a player.

3. Applications of Possession Added Value (PAV) in OHL Scouting

3.1. Team-Level Evaluation: 2019-20 Erie Otters



Forwards	Avg PAV	PAV Perc. %
Danial Singer	0.0126	79%
Chad Yetman	0.0081	65%
Brendan Hoffmann	0.0068	61%
Maxim Golod	0.0043	52%
Connor Lockhart	0.0039	50%
Daniel D'Amato	0.0038	50%
Brendan Sellan	0.0037	49%
Kyen Sopa	0.0034	48%
Elias Cohen	0.0032	48%
Matthew MacDougall	0.0026	46%
Austen Swankler	0.0024	45%
Emmett Sproule	0.0022	44%
Noah Sedore	0.0018	43%
Hayden Fowler	0.0016	42%

Defensemen	Avg PAV	PAV Perc. %
Jacob Golden	0.0008	54%
Drew Hunter	0.0007	53%
Brendan Kischnick	0.0003	50%
Cameron Morton	0.0002	48%
Jamie Drysdale	0.0001	48%
Jack Duff	-0.0002	46%
Kurtis Henry	-0.0004	44%
Luke Beamish	-0.0017	35%
Christian Kyrou	-0.0027	29%

Figure 6: A team-level analysis PAV for the 2019-20 Erie Otters

At the team level, we can easily compare players by average PAV per event or convert this into percentile among players in the league. It is also possible to spatially depict the areas of the ice where each player is adding value to the sequence. As such, these graphs show that defencemen seem to be more prone to cumulating negative PAVs (about 70% of all defencemen in the dataset), due to the fact that they are generally limited to the point, which corresponds to lower value areas of the ice.

3.2. Sample Scouting Report using PAV: Connor Lockhart, 2021 Draft Eligible Prospect

From a scouting standpoint, breaking down PAV by event type could help better understand situations in which a prospect adds value to offensive sequences for his team. The following paragraphs exemplify how our metric can be utilized to analyze a prospect's game. Looking at Connor Lockhart's player card, we notice that this undersized right winger adds value to this team's possessions in 3 different ways.

With a 50% carry rate on zone entries, Lockhart allows his team to generate sustained offensive pressure by ensuring full control of the puck on zone entries, every second time. As a right-winger, he tends to enter the zone from the right side rather than the middle of the ice, making a trade-off between ensuring sustained offensive pressure and generating shots off the rush.

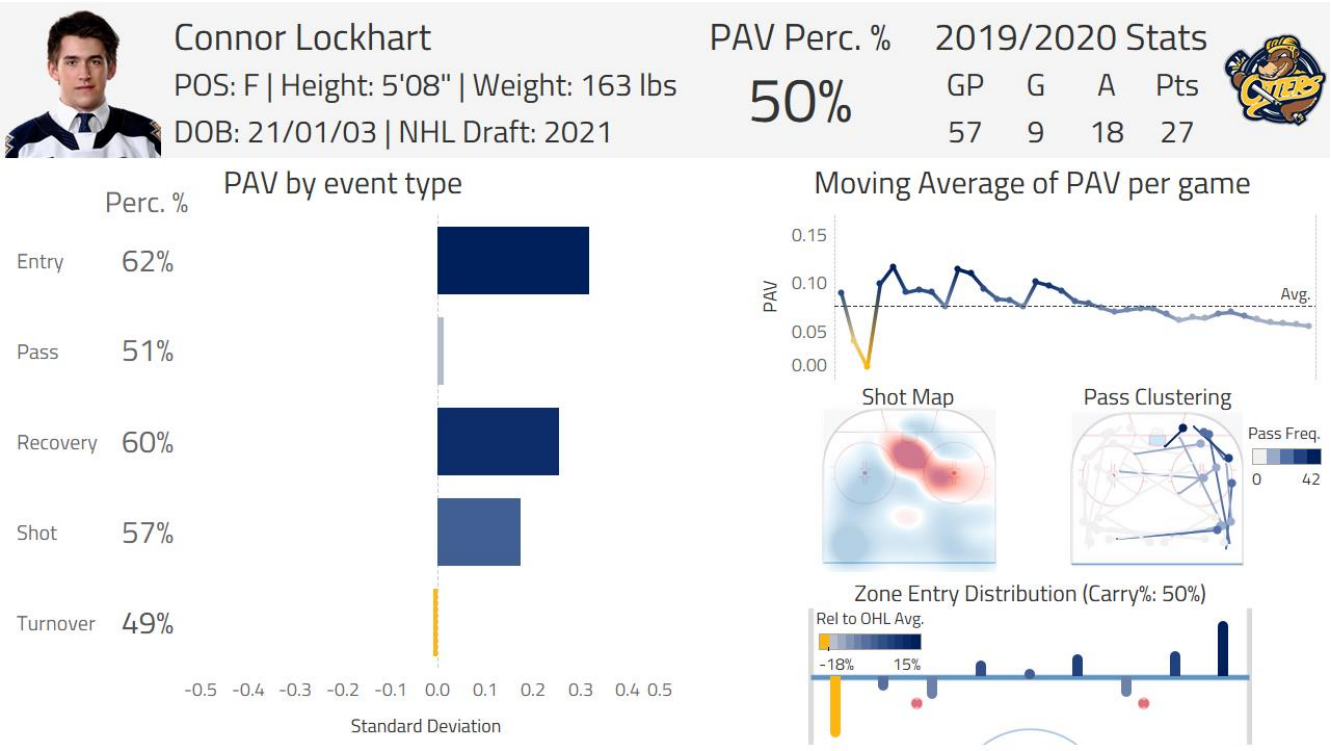


Figure 7: A player card for Connor Lockhart in the 2019-20 season.

In terms of recoveries, Lockhart's strengths in this area of the game are displayed by him ranking around the 60th percentile among league forwards in terms of PAV. Continuing to focus on smartly recovering the puck will help Lockhart sustain his strong track record in this category.

From a shooting perspective, most of his attempts are in the slot (high danger) area making him an offensive threat for the opposition. His shooting PAV, which is around the 57th percentile, helps highlight his ability to quickly release his shot in tight areas.

Lockhart is around league average in terms of PAV for passes and turnovers. Looking at his passing clusters, we notice that most of his passes don't add much value from an offensive standpoint (low to high passes, cycle passes, etc.). Ensuring a better first touch could help Lockhart bring his game to another level in both of these categories.

All in all, Lockhart ranked around the 50th percentile among OHL forwards in his rookie year, looking at average PAV per event, being the 5th best attacker on his team. He has shown some interesting signs of offensive upside, which could help convince an NHL team to take a chance on him in rounds 4-7 in the upcoming NHL draft.

5. Future Work

While our paper presented preliminary rankings of OHL players, there are many avenues to extend our work. There are two main branches to explore moving forward with the model we have presented.

First, we can continue to build upon our model and methodology. This can include expanding our model to the full ice, quantifying defensive contributions, accounting for quality of teammates, and extending to higher resolution tracking data.

Second, our paper is just scraping the surface of potential data analyses with the PAV metric we have developed. The robust nature of this metric would allow us to cluster players by play style, analyze the spatiotemporal changes in PAV over a season, and incorporate uncertainties into player and sequence evaluation.

Code and Figures

The code and spatial maps generated in this project can be found at https://github.com/brenkumi/BigDataCup2021

Acknowledgements

We would like to thank Stathletes and the Erie Otters for graciously organizing this competition and providing interesting and exciting data to work with. We would also like to thank Håvard Rue and the R-INLA team for creating an amazing computationally efficient method for computing Bayesian GMRF models with easy implementation in R. Without this package, our project would not be possible for this competition.

References

[1] Cervone, D., D'Amour, A., Bornn, L. & Goldsberry, K. (2016). A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes, Journal of the American Statistical Association, 111:514, 585-599, DOI: 10.1080/01621459.2016.1141685

[2] Fernandez, J., Bornn, L., & Cervone, D. (2020). A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. arXiv preprint arXiv:2011.09426.https://arxiv.org/abs/2011.09426

[3] Singh, K. (2019). Introducing Expected Threat (xT). Retrieved from https://karun.in/blog/expected-threat.html

[4] Yu, D., Piggott, P., Jung, C. & Forstner, S. (2020). A comprehensive analysis of pass difficulty, value and tendencies in hockey. ISOLHAC. Retrieved from https://www.youtube.com/watch?v=Q-kWb6Vshmo&t=2313s

[5] Chatel, T. (2020). Introducing Offensive Sequences and The Hockey Decision Tree. Retrieved from https://hockey-graphs.com/2020/03/26/introducing-offensive-sequences-and-the-hockey-decision-tree/

[6] Lee, J.Y.L., Green, P.J., Ryan, L.M. (2017). On "Poisson Trick" and its Extensions for Fitting Multinomial Regression Models. arXiv preprint arXiv:1707.08538

[7] Simpson, D.P., Rue, H., Martins, T.G., Riebler, A. & Sorbye, S.H. (2014). Penalising model component complexity: A principled, practical approach to constructing priors. arXiv preprint arXiv: 1403.4630

[8] Rue, H., Martino, S. & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71: 319-392. https://doi.org/10.1111/j.1467-9868.2008.00700.x