MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 3 - 4, 2017 HYNES CONVENTION CENTER

Apples-to-Apples: Clustering and Ranking NHL Players Using Location Information and Scoring Impact

Oliver Schulte, Zeyu Zhao
School of Computing Science, Simon Fraser University, Vancouver, Canada

Mehrsan Javan, Philippe Desaulniers
SPORTLOGiQ, Montreal, Canada

Paper Track: Other Sports
Paper ID: 1625

Abstract

Using new game events and location data, we introduce a player performance assessment system that supports drafting, trading, and coaching decisions in the NHL. Players who tend to play in similar locations are clustered together using machine learning techniques, which capture similarity in styles and roles. Clustering players avoids apples-to-oranges comparisons, like comparing offensive and defensive players. Within each cluster, players are ranked according to how much their actions impact their team's chance of scoring the next goal. Our player ranking is based on assigning location-dependent values to actions. A high-resolution Markov model also pinpoints the game situations and rink locations in which players tend to do actions with exceptionally high/low values.

1. Introduction

Player comparison and ranking is a very difficult task that requires deep domain knowledge. The difficulty is not only in defining appropriate key metrics for players, but also in finding a group of players who have similar playing styles. From the scouting perspective, a scout has to watch multiple games of a player to come up with a conclusion about the skills and style of a young talent. However, it is not possible for the scouts to watch all the games from all the leagues across the world. In this study, we propose a purely data-driven approach to simulate the way that the management and scouts evaluate players. This makes it possible, in principle, to apply our algorithm across multiple leagues (e.g. other professional or junior leagues) and find undervalued players. The proposed approach follows the intuitive insights to group players and develops advanced game models to assess players' skills. More specifically, this work develops statistical machine learning models to leverage the location information of a detailed dataset of NHL games in order to cluster and rank players based on their style and impact on the game outcome.

A new dataset provided by SPORTLOGiQ contains a rich set of player actions and game events along with their precise locations (*x-y* coordinates). In this paper, we have used the location information to first generate clusters of players who are similar and then compute a *Scoring Impact (SI)* metric based on the players' actions at different locations. The rationale behind clustering players before ranking them is intuitive; for example no one ever compares a defenseman to a forward; for the

purpose of this exercise, a forward should be compared to a forward while a defenseman should be compared to a defenseman. There are also some studies that point out this issue; for example it is suggested in [1] that player performance metrics are mostly meaningful for comparing similar players. Although this is trivial for anyone who knows hockey, building a purely data-driven approach to generate clusters of players without using any prior information is not an easy task. To build the player clusters, we use the location pattern of the players, i.e., where they tend to play. This generates clusters in an unsupervised fashion, which results in groups of players with similar styles and roles.

Once the clusters are formed any metric can be developed to rank the players and evaluate their impact on the game outcome. Here we focus on measuring how much a player's actions contribute to the outcome of the game at a given location and time. This is performed by assigning a value to each action depending on where and when the action takes place using a Markov decision process model. For example, the value of a pass depends on where it is taken and it has to be rewarded if it ends up in maintaining the puck possession. Once the values for the actions and game events are assigned, players can be ranked according to the aggregate value of their actions, and compared to others in their cluster. In this study, the value of a player's action is measured as its impact on his team's chance of *scoring the next goal*; the resulting player metric is called his Scoring Impact. We have chosen the *scoring the next goal* as the end goal of a sequence of game events because it can be clearly defined as a measurable objective for a team. However, the developed model is not necessarily dependent on this outcome and any other event can be used as the end state of the Markov process.

The experimental results indicate that the Scoring Impact correlates with plausible alternative metrics such as a player's total points and salary. However, *SI* suggests some improvements over those metrics as it captures more information about the game events. We illustrate the results by identifying players that highly impact their team scoring chances, yet draw a low salary compared to others in their cluster. We discuss about the advantages and shortcomings of our modeling approach in the following sections.

2. Markov Game Models: Previous Work and Our Approach

The Markov model-based approach to valuing decisions and ranking players was developed for basketball by Cervone et al. [4], who note that their approach extends to any continuous-flow sport. The details of our NHL model are quite different from their NBA model, mainly because the NBA tracking data from SportVU include the positions of all players. The SPORTLOGiQ data used in this work include the location of the puck events. In an earlier work on hockey game models, Routley and Schulte [5] developed a Markov model based on the publicly available NHL data, using the zone of an action as the only location information. Other NHL Markov models assessed player performance based on goals and penalties only [1, 6]. Depending on the outcome of interest, a Markov model can be used to assess the impact of a player's actions on outcomes of interest other than goals, such as wins [1, 7] and penalties [5]. Identifying player types by spatial action patterns was inspired by the work of Miller et al. [8] on NBA player types. Their work was developed solely based on shot locations and applied matrix factorization with Poisson point processes rather than clustering with discrete-region heat maps.

In the current work, player clustering is done by the use of the affinity propagation algorithm [2]. It groups players by clustering heat maps that represent their location patterns. To compute the probability that a team scores the next goal given the current state of the game, a Markov Decision Process is developed to model hockey games [3]. A Markov model is a powerful representation of game dynamics that has recently been shown to be effective for assigning values to actions and evaluating player performance [1, 4, 5]. The model defines the probability of a game continuation, given a current game state. For instance, given a current game state, it assigns a probability to a set of trajectories that result in an ending state, such as scoring the next goal. Our Markov game model consists of over 100,000 transition probability parameters. As opposed to approaches for player performance assessment that are based on using aggregate action counts as features, our model-based method has several advantages, including:

- Capturing the game context: the *states* in the model capture the *context* of actions in a game. For example, a goal is more valuable in a tied-game situation than when the scorer's team is already four goals ahead [1].
- Look-ahead and medium-term values: modeling game trajectories provides *look-ahead* to the medium-term consequences of an action. Looking ahead to the medium-term consequences allows us to assign a value to *every* action. This is especially important in continuous-flow games like hockey because evident rewards like goals occur infrequently [4]. For example, if a player receives a penalty, the likelihood increases that the opposing team will score a goal during the power play at some point, but this does not mean that they will score immediately.
- Player and team impact: The aggregate impact of a player can be broken down into his average *impact at specific game states*. Since game states include a high-level of context detail, the model can be used to find the game situations in which a player's decisions carry especially high or low values, compared to other players in his cluster. This kind of drill-down analysis explains, and goes beyond, a player's overall ranking. We provide what to our knowledge are the first examples of drill-down analysis for two players. (Taylor Hall and Erik Karlsson, who are the most highly ranked in their cluster). While this paper focuses on players, the same approach can be used to cluster and analyze the performance of lines and teams.

3. Hockey Dataset

We make use of a new proprietary dataset from SPORTLOGiQ. The data are generated from the videos using computer vision techniques including player tracking and activity recognition. Table 1 shows the dataset statistics for the 2015-2016 NHL season. The dataset consists of play-by-play information of game events and player actions for the entire season. Every event is marked with a continuous time stamp, the *x-y* location, and the player that carries out the action of the event. The play-by-play event data records 13 general action types (shown as Table A- 1 in the appendix). Table 2 shows an example play-by-play dataset. The table utilizes *adjusted spatial coordinates* where negative numbers refer to the defensive zone of the acting player, positive numbers to his offensive zone. To illustrate, Figure 1 shows a schematic layout of the ice hockey rink. The units are feet. Adjusted X-coordinates run from -100 to +100, and Y-coordinates from -42.5 at the top to 42.5 at the bottom, and the origin is at the ice center.

4. Location-Based Player Clustering

Hockey is a fast-paced game where players of all roles act in all parts of the ice hockey rink. Our player clustering method is based on each player's distribution of action locations across the rink. To represent the action location pattern of a player, we divide the rink into a fixed number of regions, as shown in Figure 1. This division uses four horizontal and three vertical regions, corresponding to the traditional center, left and right wings. For each player, the *region frequency* is the total number

Table 1. Dataset statistics for the 2015-2016 season.

Number of Teams	30
Number of Players	2,233
Number of Games	1,140
Number of Events	3.3M

Table 2. Sample play-by-play data

gameId	playerId	Period	teamId	xCoord	yCoord	Manpower	Action Type
849	402	1	15	-9.5	1.5	Even	Lpr
849	402	1	15	-24.5	-17	Even	Carry
849	417	1	16	-75.5	-21.5	Even	Check
849	402	1	15	-79	-19.5	Even	Puckprot
849	413	1	16	-92	-32.5	Even	Lpr
849	413	1	16	-92	-32.5	Even	Pass
849	389	1	15	-70	42	Even	Block

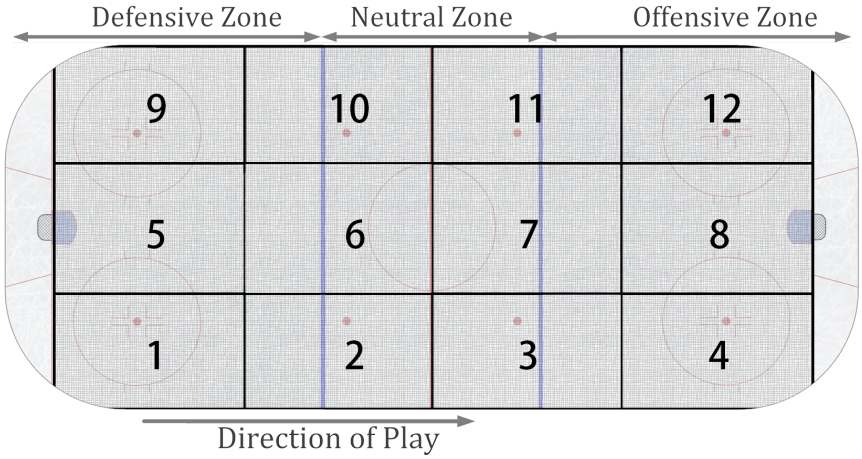

Figure 1. Rink divided into 12 regions for player heat maps.

of actions he has performed in a region, divided by the total number of his actions. Converting counts to frequencies avoids conflating the level of a player's activity with the location of his actions. We apply the well-known affinity propagation algorithm [2] to the player frequency vectors to obtain a player clustering. The appendix provides technical details on the affinity propagation algorithm. Affinity propagation produced nine player clusters which seem to be correct: four clusters of forwards, four clusters of defensemen, and one cluster of goalies. It is interesting to note that the clustering is an unsupervised process.

Figure 2 shows the 12-region activity heat map for Taylor Hall and Figure 3 represents the heat map for the cluster he belongs to. (Best viewed on-screen for color; darker red represents higher frequency, darker blue lower frequency.) Similarly, Figure 4 shows the heat map for Erik Karlsson and Figure 5 depicts the average heat map for Karlsson's cluster. The average heat map represents the average frequency of the game events which are happening in that region, over all players in the cluster. The heat maps show that Karlsson and other players in his cluster tend to play a defensive role on the right wing, whereas Hall and other players in his cluster play a more offensive role, mostly on the left wing.

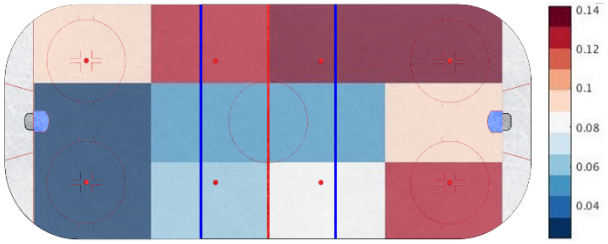

Figure 2. Taylor Hall's activity heat map.

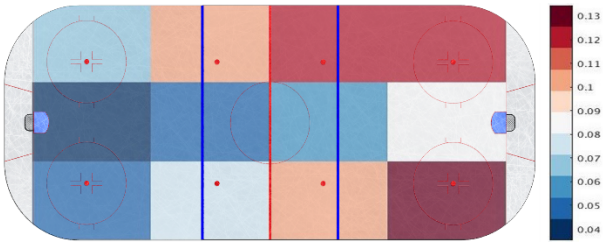

Figure 3. Activity heat map for Hall's cluster.

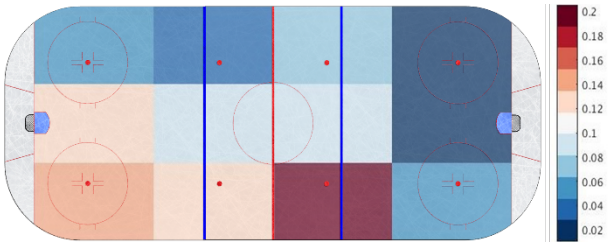

Figure 4. Erik Karlsson's activity heat map.

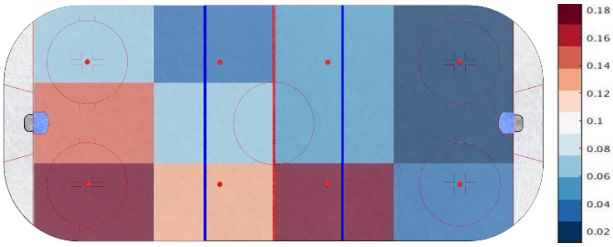

Figure 5. Activity heat map for Karlsson's cluster.

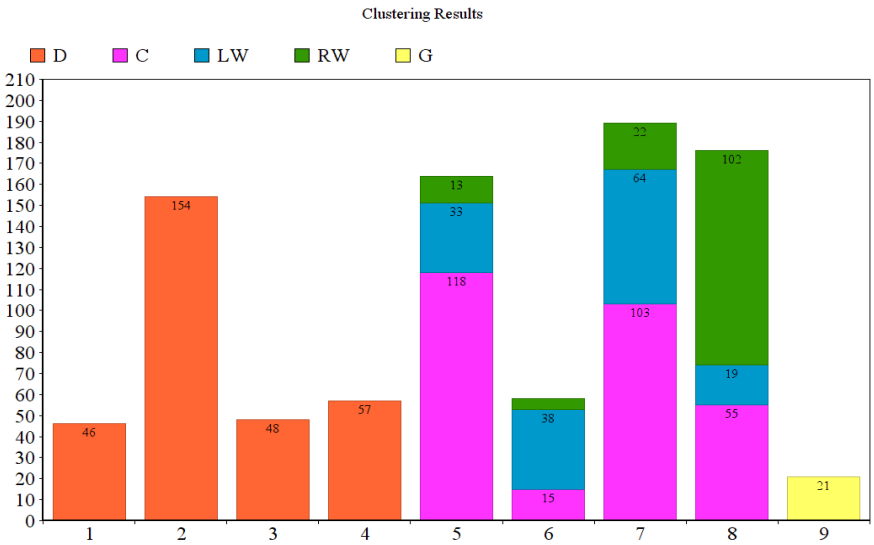

Figure 6. The learned clusters match the player categories of forward and defenseman.

Table 3. Top 4 Players in Taylor Hall's and Eric Karlsson's cluster, sorted by scoring impact per 20 minutes played

Cluster	Cluster Average Scoring Impact	Name	SI	GP	Goals	Assists	Passes	TOI.pg	Salary ($M)
7	2.851	Taylor Hall	4.775	81	26	39	320	19.204	6
		Pavel Datsyuk	4.675	60	14	33	159	19.655	7
		Evgeni Malkin	4.536	57	27	31	190	19.369	9.5
		Sidney Crosby	4.475	80	36	49	277	20.469	12
4	3.181	Eric Karlsson	6.093	77	15	66	303	28.975	7
		Kris Letang	4.888	71	15	51	168	26.945	7.25
		Alex Pietrangelo	4.831	73	7	30	202	26.305	6.5
		Tyson Barrie	4.696	78	14	36	163	23.200	3.2

It is important to compare the learned clusters with the known player types. Figure 6 shows that *the clusters match the basic grouping into defensive players and forwards.* We emphasize that the algorithm discovers this grouping only from the game event location information without being given any prior or explicit information about the player's official position. Forwards are commonly divided into centers, left wing players and right wing players. The learned forward clusters match this division to some extent. For instance, cluster 5 and 7 contain mainly but not only centers, cluster 6 contains mainly but not only left-wingers, and cluster 8 contains mainly but not only right-wingers. It indicates that not only the clusters match the conventional player positions, but also they provide information beyond those predefined positions. As an example, Table 3 shows the top four players in Hall's cluster and in Karlsson's cluster along with their scoring impact (described below) and some standard metrics.

5. The Markov Game Model

A Markov model is a dynamic model that represents how a hockey game moves from one game state to the next. A sequence of state transitions constitutes a trajectory. The parameters of a (homogeneous) Markov chain are transition probabilities $P(s'|s)$ where s is the current state and s' the next state. Previous Markov chain models for ice hockey have included goal differential and/or manpower differential in the state space [1, 6, 9]. Then the transition probabilities represent how goal scoring and penalty drawing rates depend on the current goal and manpower differentials. This approach can measure the impact of those actions that directly affect the state variables, i.e., goals and penalties. Markov decision processes and Markov game models include both states *and actions*, which allows us to measure the impact of *all* actions. The parameters of our Markov game model are *state-action transition probabilities* of the form $P(s', a'|s, a)$ where a is the current action and a' is the next action. The model therefore describes state-action trajectories as illustrated in Figure 9.

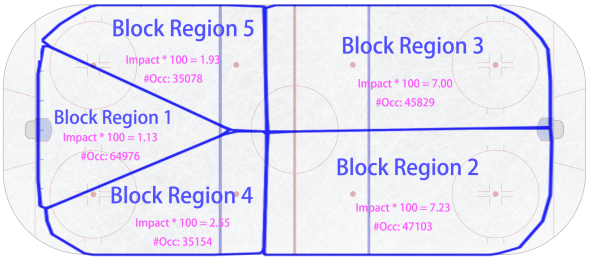

Figure 7. The learned regions for "Block" events

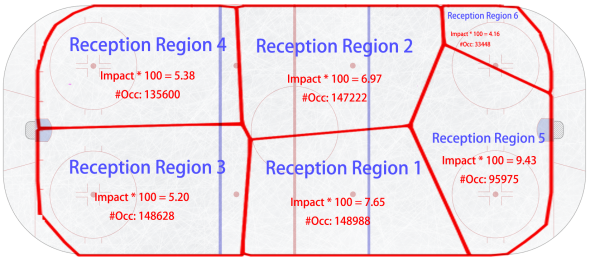

Figure 8. The learned regions for "Reception" events

5.1. Spatial Discretization

Our Markov model represents the probability that a given action occurs at a given location on the rink. To model the action occurrence probability, we discretize the rink space into a discrete set of regions. One option for generating discrete regions is to use a fixed grid, such as the one shown in Figure 1. However, the problem with a fixed grid is that different types of actions tend to be distributed in different locations. For example, shots hardly ever occur in the defensive zone, whereas blocks often do. Therefore, using the same grid for shots and blocks is neither statistically nor computationally efficient. Instead, we learned from the data *a separate discretization tailored to each action*, by applying affinity propagation to cluster the locations of occurrences of a given action type. Figure 7 shows the resulting regions for Blocks, and Figure 8 for Receptions. In each figure,

the cluster mean is shown with an occurrence label indicating how many actions are happening in each region. The figures also show the impact of the actions on scoring the next goal for each region, averaged over the game contexts. Those numbers are derived from the developed Markov game model.

5.2. State and Action Spaces

A state is a vector of values, one for each *state* variable, as shown in Table 4. An action event, *a,* is one of the 13 action types combined with two attributes: which team performs the action (Home or Away) and the action location where the action takes place. For instance, *block(home,region3)* denotes the event that the home team blocks the puck in the block-region 3 (see Figure 7). Figure 9 shows a possible state-action trajectory, which Table 5 describes in play-by-play format. In our model the number of states is "17 x 3 x 4 = 204", and there are 63 action-region pairs (see Table A-1 in appendix) which can be carried out by either the home team or the away team, for a total number of "63 x 2 = 126" action events.

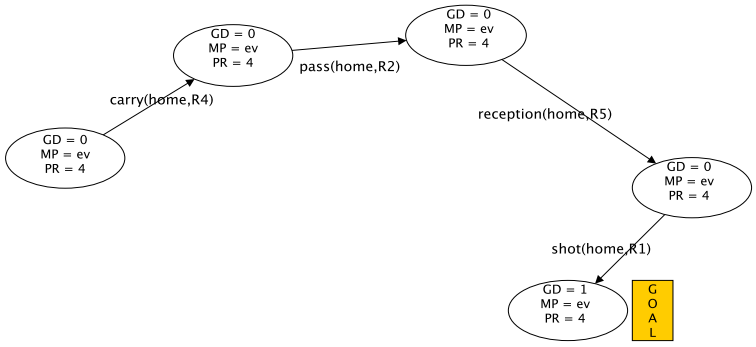

Figure 9. A possible state action trajectory in our Markov game model.

Table 4. State Variables and Values

Notation	Name	Range
GD	Goal Differential	[-8,8]
MD	Manpower Differential	EV, SH, PP
P	Period	[1,4]

Table 5. The state-action trajectory in play-by-play format. For quantities derived from the model see the text.

	State Variables			**Action Parameters**			**Quantities derived from Model**		
Event	Goal Differential	ManPower Differential	Period	Team	Action Type	Region	Transition Probability	Conditional Value (Home)	Impact
0	0	Even	4	Home	Carry	4	----	73%	----
1	0	Even	4	Home	Pass	2	21%	71%	-2%
2	0	Even	4	Home	Reception	5	3%	76%	5%
3	0	Even	4	Home	Shot	1	34%	86%	10%

5.3. Parameter Estimation

The key quantities in our model specify the joint *state-action distribution* $P(s', a'|s, a)$ that an action a occurs at the game state *s,* and is followed by game state *s'* and action *a'*. Because the distribution of the next action and its location depends on the most recent action and its location, the action distribution represents spatial and temporal dynamics. For example, the transition probability of 21% in the second row of Table 5 includes the probability that play moves from carry-region 4 to pass-region 2, given the current game context. We refer to a state-action pair as a *game context*, so *P(s',a'|s,a)* models a context transition probability. Decomposing this probability as $P(s', a'|s, a) = P(s'|a', s, a) \times P(a'|s, a)$, we see that it combines two quantities of interest: (1) the state transition probabilities $P(s'|a', s, a)$ that describe how game states evolve given players' actions. (2) The *action distribution* $P(a'|s, a)$ that describes how a random player acts in a given game context.

We estimate the action-state distribution using the *observed occurrence counts* $n(s', a', s, a)$, which record how often action a' and state s' follows state s and action a in our dataset. For simplicity we slightly abuse notation and use n also for marginal occurrence counts, for example $n(s, a) = \sum_{s',a'} n(s', a', s, a)$. The maximum likelihood estimates are computed as follows:

$$\hat{P}(s', a'|s, a) = \frac{n(s', a', a, s)}{n(s, a)} \quad (1)$$

The number of possible state-action quadruples is unmanageably large at over 600 million. However, the number of quadruples that occur more than zero times is only 112,590. The necessary computations for computing and storing the estimated values can be efficiently managed using appropriate data structures; for more details please refer to [7]. Our code is available on-line [10]. We next show how our Markov game model can measure the impact of all actions.

6. Action Values and Scoring Impact

In our model, the agents are a generic home team and a generic away team, not individual players, similar to previous Markov game models [1] for hockey. This is appropriate for the goal of assigning generic values to all action events. In this paper we use the Markov model to quantify how a player's action, given a game context, affects the probability that his team scores the next goal. A similar approach can be followed to quantify the impact of actions on other outcomes of interest, such as winning the game [1, 7] and penalties [5]. A key feature of a Markov model is that it quantifies not only the immediate but also the medium-term impact of an action.

For $T = home$ or $away$, let $P(T\ scores\ next|s, a)$ denote the probability derived from the model, that after action a, the team T scores the next goal, before the opposing team $\overline{T}$. In the appendix we show how this probability can be computed using the dynamic programming algorithm. For a point in a game, it is possible that a play sequence ends with neither team scoring. Therefore another quantity of interest is the conditional probability that a team scores *given* that one of the two teams scores next. We refer to this as the *conditional value* of a game context for team T.

$$CV_T(s, a) = \frac{P(T\ scores\ next\ |s, a)}{P(T\ scores\ next\ |s, a) + P(\overline{T}\ scores\ next\ |s, a)} \quad (2)$$

The conditional value is an appropriate quantity for evaluating actions since the goal of an action is to improve a team's position relative to their opponent. The *impact of an action* is defined as the extent to which the action changes the conditional value of the acting player's team at a state. Figure 7 shows the impact of a "Block" by region, averaged over game contexts; ditto Figure 8 for "Receptions". The scoring impact metric for a player is defined as their total impact over all their actions and formulated as follows:

$$Impact(a'; s, a) = \sum_{s'} CV_T(s', a') \times P(s'|a', s, a) - CV_T(s, a) \quad (3)$$

$$SI_i = \sum_{a',s,a} n_i(a', s, a) \times Impact(a'; s, a) = \sum_{s,a} n_i(s, a) \times \sum_{a'} Impact(a'; s, a) \times P_i(a'|s, a) \quad (4)$$

where $P_i(a'|s,a) = \frac{n_i(a',s,a)}{n_i(s,a)}$ is the action distribution for player *i.* The occurrence counts $n_i(a', s', s, a)$ record how many times the game reaches the state *s'* and player *i* takes action *a'* after state *s* and a player (not necessarily *i*) took action *a*. The second expression for the scoring impact shows that the SI metric can be interpreted as the expected impact of a player given a game context *(s,a)*, weighted by how often the player reaches the context.

6.1. Correlation Analysis

The *SI* metric shows a strong correlation with other important metrics, such as points, time on ice, and salary. This correlation increases by computing the metric for comparable players rather than all players. Table 6 shows the correlation between SI and time on ice (per game). For example, the correlation between *SI* and time on ice is 0.83 overall, and increases to 0.89 and 0.92 for the clusters shown in the table. The *SI* is also temporally consistent [1], i.e., a player's SI metric in the first half of the season correlates strongly with his SI metric in the second half (ρ = 0.77).

Table 6. Correlation between SI and TOI (per 20 min played)

Cluster #	All	1	2	3	4	5	6	7	8
Correlation Coefficient	0.83	0.89	0.89	0.92	0.89	0.92	0.82	0.92	0.90

6.2. Case Studies

For our example clusters, we discuss the top-ranked player and some undervalued players. The appendix shows metrics for all players discussed.

Taylor Hall's cluster. This cluster, cluster number 7, comprises forwards only. Table 3 shows the top 4 players by scoring impact: Taylor Hall, Pavel Datsyuk, Evgeni Malkin, and Sidney Crosby, who all are known as excellent offensive players. Taylor Hall is recognized as a high calibre forward, placing him highly in the NHL fantasy rankings [11]. His goals-per-game metric is 0.32, which is excellent but behind for instance Malkin's at 0.47. This shows how our ranking is correlated with goals but also takes into account the value of other actions by the player. For instance, our ranking reflects that the total number of Hall's passes is 320, substantially more than Malkin's 190 passes.

The highly ranked players with low salary in cluster 7 are Aleksander Barkov (rank 6, salary $0.925M) and Jack Eichel (rank 8, salary $0.925M). Both players are junior (first NHL season in 2011 for Barkov, 2012 for Eichel). Barkov is viewed as having played a successful season and received from the Florida Panthers a six-year contract extension for $35.4M, a six-fold salary increase [12], which is consistent with our ranking. Eichel is a rising star [13]. Being in the same cluster as Hall suggests that Barkov and Eichel are strong prospects for replacing this senior forward.

Erik Karlsson's Cluster. This cluster comprises defense players only. Table 3 shows the top 4 players by scoring impact. The top player in cluster 4 is Erik Karlsson. He has won the Norris Trophy twice for best all-round defenseman in the NHL. The NHL ranks him as the top defenseman for fantasy play in the 2016 season [14]. According to our SI ranking, John Klingberg draws a low salary in this cluster. Although he signed a contract with the Dallas Stars in 2011, he did not play a full NHL season until 2014-2015. After this season, he was recognized by an invitation to the NHL all-rookie team, which is consistent with our ranking. Being in the same cluster as Karlsson suggests that he is a strong prospect for replacing this senior defenseman.

6.3. Explaining the Rankings: Drill-Down Analysis

In this section we illustrate how a player's ranking can be explained by how he performs in specific game contexts. This breakdown makes the ranking interpretable because it explains the specific observations that led to the rating and pinpoints where a player's effectiveness deviates from comparable players. Our basic approach is to find the game contexts in which a player's expected impact differs the most from a random player in his cluster. We refer to this metric as the player's *added impact,* $\Delta_i(s,a)$, computed as follows:

$$\mathrm{E}_j(s,a) = \sum_{a'} Impact(a';s,a) \times P_j(a'|s,a) \tag{5}$$

$$\Delta_i(s,a) = \mathrm{E}_i(s,a) - \sum_{j \epsilon C} \frac{n_j(s,a)}{n_c(s,a)} \mathrm{E}_j(s,a) \tag{6}$$

where C is the cluster of player i and $n_C(s,a) = \sum_{j \epsilon C} n_j(s,a)$. Drill-down analysis looks for game contexts where the player shows an unusually high or low added impact. For Taylor Hall, his highest added impact is in the first period, with even score and manpower, after his team has managed a reception in region 1. Among action types, the highest added impact stems from Block. Figure 10 compares Hall's region distribution for Blocks with those of a random player from his cluster. In the specified game context, a Block has the most scoring impact in the left-wing offensive region 3. For this game context, 50% of Taylor Hall's Blocks occur in this high-impact region, compared to only 19.6% of Blocks for a random player from his cluster.

For Erik Karlsson, highest added impact is in the third period, with even score and manpower, after his team has managed a pass in region 4. Figure 11 compares Karlsson's region distribution for Receptions with those of a random player from his cluster. In the specified game context, a Reception has the most scoring impact in the right-wing offensive region 1. For this game context, 37.5% of Karlsson's Receptions occur in this high-impact region, compared to only 14.6% of Receptions for a random player from his cluster.

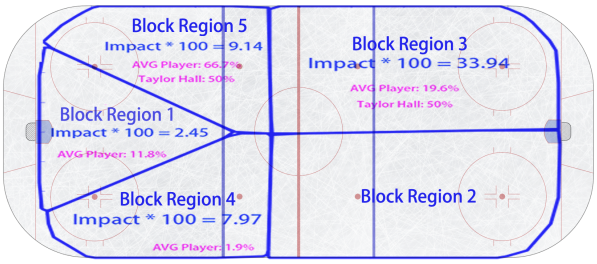

Figure 10. Drill-Down Analysis for Taylor Hall.
He manages high-impact Blocks after a Reception by his team, in period 1 with even score/manpower.

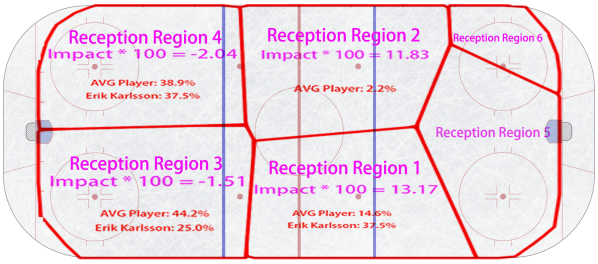

Figure 11. Drill-Down Analysis for Erik Karlsson.
He manages high-impact Receptions after a Pass by his team, in period 3 with even score/manpower.

7. Discussions

In this paper we proposed a pure data-driven approach based on clustering and Markov decision process to support the way that scouts and managers evaluate players. This study showed how location information of the game events and player actions can be used to identify players with

similar styles and roles and rank them based on their impact on scoring the next goal. Our work supports apples-to-apples comparisons of similar players. Once the clusters are formed, a high-resolution large-scale Markov game model quantifies the impact of *all* events on scoring the next goal. The aggregate impact of an action provides a principled effective way to assess player performance. Breaking down the aggregate impact allows the analyst to pinpoint the exact situations in which a player's decisions tends to deviate-positively or negatively-from comparable players. Statistical modelling could further enhance drill-down analysis by identifying which features of the game context and of a player's actions predict a high added-impact.

Although considering the Scoring Impact as a metric for player performance assessment is coherent with the overall objective of the team, we have to emphasize that it cannot completely capture players' strengths and skills. Scoring the next goal is a sophisticated process and is affected by many other factors such as the opposing goalie and the shot quality (which depends on where other players are when a shot happens). Therefore, the immediate extension of the model is to improve the shot quality model by first, incorporating the location of other players in the Markov process, and second, putting the goalie into the equation. In future work we will also research Markov decision processes for ice hockey that represent continuous spatial-temporal processes, rather than discretized time and space. Another important direction is modelling action distributions and state transitions at the level of individual teams, lines, and players.

Acknowledgements

We thank SPORTLOGiQ for providing hockey data. This research was supported by Discovery and Engage grants from Canada's Natural Sciences and Engineering Research Council. We are grateful for constructive discussions in Simon Fraser University's Sport Analytics Group.

References

[1] Pettigrew, S. (2015), Assessing the offensive productivity of NHL players using in-game win probabilities, in '9th Annual MIT Sloan Sports Analytics Conference'.
[2] Frey, B. J. & Dueck, D. (2007), 'Clustering by passing messages' Frey, B. J. & Dueck, D. (2007), Science 315(5814), 972-976.
[3] Puterman, M. L. (1994), Markov Decision Processes: Discrete Stochastic Dynamic Programming, Wiley, New York, NY, USA.
[4] Cervone, D.; D'Amour, A.; Bornn, L. & Goldsberry, K. (2014), POINTWISE: Predicting points and valuing decisions in real time with NBA optical tracking data, in '8th Annual MIT Sloan Sports Analytics Conference'.
[5] Routley, K. & Schulte, O. (2015), A Markov Game Model for Valuing Player Actions in Ice Hockey, in 'Uncertainty in Artificial Intelligence (UAI)', pp. 782--791.
[6] Thomas, A.; Ventura, S.; Jensen, S. & Ma, S. (2013), 'Competing Process Hazard Function Models for Player Ratings in Ice Hockey', The Annals of Applied Statistics 7(3), 1497-1524.
[7] Routley, K. (2015), 'A Markov Game Model for Valuing Player Actions in Ice Hockey', Master's thesis, Simon Fraser University.
[8] Miller, A.; Bornn, L.; Adams, R. & Goldsberry, K. (2014), Factorized Point Process Intensities: A Spatial Analysis of Professional Basketball., in 'ICML', pp. 235--243.
[9] Kaplan, E. H.; Mongeon, K. & Ryan, J. T. (2014), 'A Markov model for hockey: Manpower differential and win probability added', INFOR: Information Systems and Operational Research 52(2), 39--50.
[10] Schulte (2016). http://www.cs.sfu.ca/~oschulte/sports/

[11] https://www.nhl.com/news/2016-17-left-wing-fantasy-rankings/c-281561644
[12] https://en.wikipedia.org/wiki/Aleksander_Barkov_Jr
[13] https://en.wikipedia.org/wiki/Jack_Eichel
[14] https://www.nhl.com/news/fantasy-hockey-top-60-defensemen-rankings-update/ c-798669. Retrieved December 2016.
[15] Sutton, R. S. & Barto, A. G. (1998), *Reinforcement learning: An Introduction* , MIT Press.

Appendix

A-1: Game Events and SPORTLOGiQ Dataset

Table A- 1 lists all action types used in our study. For each action type, we list how many regions were learned by the clustering algorithms, and how often events of this type occurred. In the full dataset, the action types are classified further for a total of 43 different types. For example, for each dump-in, the data distinguish a chip-in from an actual dump-in. We used only the 13 main types, to reduce the number of parameters of the Markov model. Fewer parameters reduce the computational complexity, and can be more reliably estimated.

Table A- 1 Action Types Recorded in the 2015-2016 Season Data

Action Types	Description	#Regions	#Occurrences
Block	A block attempt on the puck's trajectory	5	228,140
Carry	Controlled carry over a blue line or the red center line	8	257,312
Check	The player attempts to use his body to gain possession	7	79,321
Dump-in	The player sends the puck into the offensive zone	3	87,454
Dump-out	The defending player dumps the puck up the boards	3	97,951
Goal	The player scores a goal	1	6,061
Lpr	Loose puck recovery. The player recovered a free puck.	6	699,189
Offside	The player is over the offensive blue line ahead of the puck	3	7,059
Pass	The player attempts a pass to a teammate	7	926,012
Puckprotection	The player uses his body to protect the puck by the boards	7	107,270
Reception	The player receives a pass from a teammate	6	709,861
Shot	The player shoots on goal	4	140,872
Shotagainst	Shot was taken by the opposing team; attributed to goalie	3	35,627

A-2: Examples of player clusters

Table A- 2. Top Eight Players per Cluster by Scoring Impact, standardized by 20 minutes of game played.

Taylor Hall's Cluster

Name	SI	GP	Goals	Assists	Passes	TOI.pg	Salary($M)
Taylor Hall	4.775	81	26	39	320	19.204	6
Pavel Datsyuk	4.675	60	14	33	159	19.655	7
Evgeni Malkin	4.536	57	27	31	190	19.369	9.5
Sidney Crosby	4.475	80	36	49	277	20.469	12
Anze Kopitar	4.398	81	25	49	218	20.867	7.7
Aleksander Barkov	4.396	57	22	31	138	19.430	0.925
Ryan Getzlaf	4.394	67	12	50	261	19.506	9.25
Jack Eichel	4.335	71	21	32	241	19.122	0.925

Erik Karlsson's Cluster

Name	SI	GP	Goals	Assists	Passes	TOI.pg	Salary($M)
Erik Karlsson	6.093	77	15	66	303	28.975	7
Kris Letang	4.888	71	15	51	168	26.945	7.25
Alex Pietrangelo	4.831	73	7	30	202	26.305	6.5
Tyson Barrie	4.696	78	14	36	163	23.200	3.2
Brent Burns	4.637	75	25	48	204	25.864	5.76
Drew Doughty	4.499	82	14	37	168	28.018	7.1
John Klingberg	4.393	62	9	48	199	22.688	2.25
Dustin Byfuglien	4.375	81	19	34	177	25.203	6

A-3: Clustering Algorithm

Affinity propagation
Affinity propagation does not require specifying the number of clusters in advance. Instead, it automatically determines the number of clusters, based on a preference hyper-parameter *p(i)* for data point *i*; data points with higher preferences are more likely to be selected as cluster centers. Following the advice of Frey and Dueck [2007], we found that setting *p(i)* to 4 times the median of the similarity values for all data points led to a tractable number of clusters for both players and action regions.

Action Regions
It is possible to use three regions for the horizontal direction, corresponding to the defensive, neutral, and offensive zone. However, adding a 4th horizontal division led the clustering algorithm to produce more informative groupings, without producing too many clusters. Adding a 5th horizontal division produced essentially the same player clusters as the 3x4 division of Figure 1.

An alternative approach to discretizing locations would be to apply nonnegative matrix factorization to a matrix of location transition counts [4]. The latter has the advantage that the learned regions capture not only where actions occur, but also where the game tends to move next. The disadvantage is higher computational complexity, and that arguably the resulting regions are less straightforward to interpret.

A-4: Dynamic Programming
Dynamic Programming [3] can be applied to compute the probability of an event for *each* game context. The main insight is that this probability can be computed efficiently for a *fixed event horizon l*, called the *look-ahead*. The probability of the event is then given by limit of the fixed-horizon probability as the look-ahead increases without bounds. The fixed-event probability satisfies the following recurrence relation for each context:

$$P(T\ scores\ next\ within\ 0\ steps|s,a) = \begin{cases} 1 \text{ if } a = goal(T, goalreg) \\ 0 \text{ o.w.} \end{cases} \quad (7)$$

$$P(T\ scores\ next\ within\ l+1\ steps|s,a \text{ and } a \neq goal(T,goalreg)\ ) = \sum_{s'} P(T\ scores\ next\ within\ l+1\ steps|s',a') \times P(s',a'|s,a) \quad (8)$$

$$P(T\ scores\ next\ |s,a) = \lim_{l\to\infty} P(T\ scores\ next\ within\ l\ steps)$$

The probability that a team scores next can computed by applying the recurrence equations to each game context, for look-ahead $l = 1,2,\ldots$ until the conditional values in Equation (2) converge. Our convergence criterion was that the conditional value of no state changes by more than 1% from the previous look-ahead to the next. For the SPORTLOGiQ dataset, convergence occurred for $l = 14$. This means that looking ahead more than 14 steps in possible game trajectories does not change the probability estimate of which team is more likely to score next.

Dynamic programming is a key algorithm in the field of reinforcement learning [15], a subfield of machine learning. Our terminology relates to reinforcement learning concepts as follows. The probability of scoring the next goal is an instance of a "value function". An action distribution is an instance of a "policy".