Identifying Player Roles in Ice Hockey

Rasmus Säfvenberg Niklas Carlsson Patrick Lambrix

Linköping University, Sweden

Abstract. Understanding the role of a particular player, or set of players, in a team is an important tool for players, scouts, and managers, as it can improve training, game adjustments and team construction. In this paper, we propose a probabilistic method for quantifying player roles in ice hockey that allows for a player to belong to different roles with some probability. Using data from the 2021-2022 NHL season, we analyze and group players into clusters. We show the use of the clusters by an examination of the relationship between player role and contract, as well as between role distribution in a team and team success in terms of reaching the playoffs.

1 Introduction

Ice hockey is a fast-paced team sport that emphasizes both physical prowess and technical ability [10]. However, the expectations and responsibilities of players vary, not only based on playing position, but also on the role of the player. The three traditional groups of positions in ice hockey are goaltenders, defenders, and forwards [21], where the latter two positions are referred to as skaters. However, the roles of the players are not always that clear cut. For instance, while defenders are typically given the highest responsibility for preventing the opposition from scoring, there are defenders who specialize in offensive contribution [21].

The benefits of categorizing players into roles are multi-fold. For team staff it will allow the choices and design of rosters and line-ups to be more effective in-game. Additionally, the construction of team rosters is also constrained from an economic standpoint. In the National Hockey League (NHL), the salary cap prevents a team from having salary expenditure above a fixed amount [6]. On an individual level, if there is a disagreement in expectations of the player's role between a player and a team, the development of the player may be hampered and the likelihood of attaining success is lowered for both parties [14].

Work on player roles has been performed in different sports (e.g. [1, 19]). Prior work regarding player roles in ice hockey has typically utilized methods that assign each player into a distinct cluster, e.g., using k-means, and used a limited set of performance metrics, e.g., points, plus-minus, and penalty minutes, which may leave some roles or role nuances undiscovered [21, 6]. In comparison, the aim of this paper is to identify different player roles for skaters in ice hockey using performance metrics that span more aspects of the game than previous work, as well as a wider basis for discovering different player roles. Further, players can be assigned to different roles to different degrees.

The contributions of this paper are as follows. First, we identify player roles by using a larger set of performance metrics than previous work, allowing us to discover new roles and/or key components in understanding a role, and by using fuzzy clustering, allowing each player to belong to a role to some degree, rather than assigning each player to a distinct role. Further, we show applications to team constructions in terms of player contract comparison and team composition for successful and less successful teams. Our findings have value to players, scouts, and managers.

The remainder of the paper is organized as follows. Sect. 2 describes the data used for the analysis while Sect. 3 introduces the method including preprocessing, principal component analysis and fuzzy clustering. Sect. 4 presents and contextualizes the results. Further, we show two applications. In Sect. 5 we compare players to players with similar roles with respect to their salaries and in Sect. 6 we investigate the relationship between team composition based on roles and reaching the playoffs. Limitations of the study are addressed and concluding remarks are drawn in Sect. 7.

2 Data

We use data from 2021-2022 NHL regular season obtained from the official website of the NHL[1] and their public API, as well as salary data from CapFriendly[2]. The data combines play-by-play data with shift data. From this data, a set of 46 variables was derived. Variables regarding goals, assists, and expected goals are used to evaluate offensive quality and frequency among players. Plus-minus (+/−), xGF, xGF%, and xGF% Relative serve as proxies for team performance while the player is on the ice. Giveaways gives some measure of puck control, while takeaways and blocks represent defensive contributions. Hits, net hits, penalties, net penalties, penalty minutes, and number of penalties per group all portray player aggression and physical play. The number of penalties per group variables are also split into the penalties the player is given as well as the penalties that are drawn, to distinguish between players who are the instigator and the receiver. Offensive zone starts can depict if a player is more offensively or defensively orientated. The time on ice variables capture how much ice time the player has, while the coordinate variable describes where the player typically is when performing each event. Finally, weight characterizes player physique, which is used to gauge the physical dimension of players and its impact on player role. Furthermore, the variables xGF%, and xGF% Relative serve as a proxy for puck possession, and, as [22] explains, can negate the weaknesses of the traditional plus-minus metric.

3 Method

The analysis consisted of preprocessing, dimensionality reduction and clustering.

[1] `https://www.NHL.com`
[2] `https://www.capfriendly.com`

In the **Preprocessing** step, a threshold was used of 200 minutes for minimum number of minutes played during the season to exclude players who had insufficient playing time. The number of defenders satisfying this requirement was 263 (out of 345), and the number of forwards 485 (out of 659). We then split the data into two subsets, one for defenders and one for forwards, to take positional variations into consideration. Further, all[3] performance-related variables with counts, i.e., not variables with percentages, were then standardized by dividing by total time on the ice (TOI) and multiplied by 60 (number of minutes in a game in regulation time). The variables were also normalized by subtracting the variable's mean and dividing by its corresponding standard deviation.

Next, principal component analysis (PCA) was utilized to perform **dimensionality reduction** on the data, as clustering in high dimensions tends to become ineffective [2]. The selection of the number of principal components was primarily based on parallel analysis [13, 12] to reduce the probability that too many components are kept. This selection method was shown to be among the best performing in [17]. Based on experiments regarding the robustness for the method [12] we ran 100,000 iterations using the 95^{th} percentile as the basis for selecting the number of components.

As we wanted to model that players can take on different roles to certain degrees and that roles tend to have overlapping elements, we opted to use a fuzzy **clustering** algorithm rather than the crisp clustering algorithms used in previous work. In fuzzy clustering, the objects are assigned a probability of belonging to a given cluster, where the probabilities of cluster membership of an object sum to one. In this paper, we used the fuzzy c-means algorithm [9, 3]. The objective in fuzzy c-means algorithm is to create k fuzzy partitions among a set of n objects from a data vector $\mathbf{x}$ by solving (1) until convergence.

$$\min_{\mathbf{U},\mathbf{C}} J_m = \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^m d^2(\mathbf{x}_i, \mathbf{c}_j) \;\; s.t. \;\; u_{ij} \in [0,1], \sum_{j=1}^{k} u_{ij} = 1 \tag{1}$$

In (1), d denotes the distance between object i and the j:th cluster centroid c_j. Moreover, u_{ij} is the degree of membership for object i to cluster j. The hyperparameter m controls the degree of fuzziness, where a higher m leads to a fuzzier solution [4]. It can also be shown that the fuzzy solution converges to the crisp solution as $m \to 1$ [15] and as $m \to \infty$ then $u_{ij} \to \frac{1}{k}$.

There is no optimal m that suits all cases [4]. However, $m \in [1.5, 3.0]$ tends to give satisfactory results in general [4] or are typical values [24], $m = 2$ results in compact and well separated clusters [9], but can also negatively affect the clustering [8, 20]. The formula that we use for deciding m was proposed in [20].

$$f(n,p) = 1 + \left(\frac{1418}{n} + 22.05\right) d^{-2} \left(\frac{12.33}{n} + 0.243\right) d^{-0.0406\log(n) - 0.1134}, \tag{2}$$

which only depends on the number of objects n and the dimensions p.

[3] Except xGF, which used 5 on 5 TOI

Similar to its crisp clustering counterpart, k-means, the fuzzy c-means algorithm also requires that the number of clusters k are specified in advance. A popular method for deciding how many clusters should be formed is to compare a set of candidate k by considering one or more cluster validity indices [16]. In [23] a large set of different fuzzy cluster validity indices are compared. Although no singular validity index is optimal for all data, the modified partition coefficient (MPC) was one of the indices that partitioned many of the investigated data sets into the best number of clusters. The MPC is an extension of the partition coefficient (PC) [3]: $\mathrm{V}_{PC} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^2 \in \left[\frac{1}{k}, 1\right]$, PC exhibits a monotonic evolution as k increases. As a result, [7] proposed the MPC, which is defined as:

$$V_{MPC} = 1 - \frac{k}{k-1}(1 - V_{PC}) \in [0, 1]. \tag{3}$$

Similarly to the PC, MPC quantifies the extent of sharing between fuzzy subsets. For both indices, the optimal number of clusters is given by the k that maximizes the index [23].

Although some cluster validity indices are typically only considered for fuzzy clustering, extensions of crisp clustering validity indices have also been proposed, an example being a fuzzy extension of the silhouette width criterion, which is frequently used in crisp clustering [5]. The silhouette of an object i is computed by $\mathrm{s}_i = \frac{a_i - b_i}{\max\{a_i, b_i\}} \in [-1, 1]$, where a_i describes the average dissimilarity for object i to all other objects belonging to the same crisp cluster, while b_i is the minimum average dissimilarity to the clusters where object i is not assigned [18]. Moreover, the crisp silhouette is defined as the average of the silhouette over all objects. However, in the context of fuzzy clustering, the crisp silhouette does not account for information regarding the degree of cluster overlap between two clusters. To generalize this criterion to fuzzy clustering, the fuzzy silhouette (FS) is defined by:

$$V_{FS} = \frac{\sum_{i=1}^{n}(u_{ig} - u_{ig'})^{\alpha} s_i}{\sum_{i=1}^{n}(u_{ig} - u_{ig'})^{\alpha}}, \tag{4}$$

where u_{ig} and $u_{ig'}$ represent the two largest elements from $\mathbf{U}_i$ [5] while $\alpha \geq 0$ is a weighting coefficient that is commonly set to 1 [11]. One distinction between the crisp and fuzzy silhouette is that the latter computes the weight for each term, based on the two fuzzy clusters that are found to be the best match. The optimal number of k is obtained by maximizing the index [5].

4 Results

Principal component analysis. Figure 1 shows the proportion of variance explained by each component generated by the PCA for defenders and forwards, respectively. The first six components are responsible for the majority of the variance in the data, by providing an explanation of at least 50% of the variance. However, the proportion of explained variance decreases rapidly, and in order to explain, e.g., 90% of the variance, at least 26 and 27 components are required for

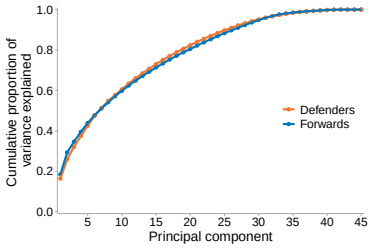

Fig. 1: Proportion of variance explained for defenders and forwards

defenders and forwards, respectively. We used parallel analysis to determine the set of components, which resulted in the selection of the first eight for defenders and nine components for forwards. These choices explain approximately 55-57% of the variance for each respective position group.

Fuzzy clustering. For obtaining values for m in fuzzy c-means we used Eq. 2 which resulted in 2.407 for defenders and 2.179 for forwards. For k we used Eqs. 3 and 4 to evaluate cluster cohesion. $k = 2$ produced the most distinct clusters for defenders and forwards and less cohesive clusters were observed for values larger than $k = 3$ (defenders) and $k = 4$ (forwards). In addition to these metrics, the cluster assignment of players was also compared to domain knowledge to guide the final choice. More specifically, the players who belong to the same cluster should share the same style of play, regardless of if other possible roles, i.e., clusters, have some overlap. As a result, $k = 3$ and $k = 4$ were chosen for defenders and forwards, respectively, as they provide more cohesive clusters while also allowing the number of roles to be as descriptive as possible.

Figure 2 shows the distribution of the probabilities representing cluster membership. We note that the densities of probabilities for defenders are more similar than forwards, where cluster F4 has a high peak close to zero. Furthermore, clusters D1 and D3 among defenders have similar distributions while cluster D2 is more centered. For forwards, both clusters F1 and F3 appear to span the entire range of possible values between zero and one, while clusters F2 and F4 have a somewhat smaller range. Except for cluster D2 among defenders, the densities reach a peak between 0 and 0.25.

To explore the variables characterizing each cluster, we retrieve the cluster centroids, which are expressed in terms of principal components. We then obtain approximate centroids corresponding to the original variables by computing an inverse transform of the centroids and the selected principal components. Moreover, since the data was standardized to have unit variance prior to conducting PCA we also invert this procedure by multiplying by the standard deviation and adding the mean for each variable. Using this method, we then obtain approximate centroids on the original variable scales, expressed in per 60 minutes of ice time (Table 1).

Among defenders, a higher probability to belong to cluster D1 is connected to the most offensively skilled defenders, who assist their team's attacking presence

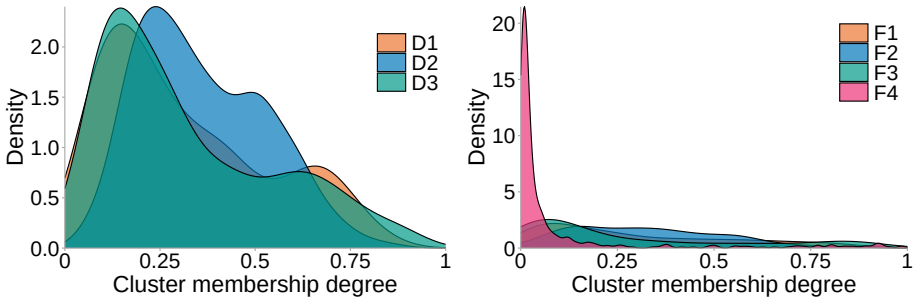

Fig. 2: Cluster membership degrees for defenders (left) and forwards (right).

by contributing more goals, assists, xG, and takeaways while also playing closer to the opposition's net. They also start in the offensive zone and in powerplay situations, more often than defenders in the other clusters. In cluster D3 we find the most physical defenders, as the number of fights, hits, and penalties are the highest, alongside the largest average weight. Their offensive contribution, with respect to xG, assist, and OZS% rank is lowest. They also garner the highest share of time played while shorthanded, but lowest in overtime and while shorthanded. Finally, the third cluster, D2, contains the defensive specialists, where goals, penalty minutes, and hits are at their lowest, in combination with playing closer to their own net. They are also the shortest and lightest.

Regarding forwards, the players in cluster F4 can be described as physical players, with the highest weight in combination with most fights, hits, and penalty minutes. The offensive production is second lowest, and they tend to be preferred in defensive situations. The offensive specialists can be found in cluster F3, where goals, assists, and xG are at their highest, which also can be seen in their play closer to the opposition's goal. Moreover, players in F3 also draw the most non-physical penalties. These players also block the most shots. A lower, but still second highest, offensive proficiency characterizes the players in F2, alongside positive xGF% values. Finally, the two-way forwards reside in F1, with skating that covers the entirety of the rink. Additionally, the lowest xG and goals are created by these players, alongside the lowest +/− and xGF On. However, they rank the highest for time played while shorthanded and percentage of starts in the defensive zone.

Table 1: Approximate cluster centroids of the original variables per 60 minutes.

Variable	F1[a]	F2[b]	F3[c]	F4[d]	D1[e]	D2[f]	D3[g]
Goals	0.536	0.784	1.189	0.596	0.304	0.172	0.165
Assists	0.782	1.057	1.535	0.663	1.032	0.701	0.592
xG	0.649	0.823	1.023	0.711	0.271	0.187	0.174
xG difference	-0.113	-0.039	0.166	-0.116	0.033	-0.015	-0.010
S%	0.090	0.108	0.143	0.093	0.059	0.041	0.040
+/−	-0.532	-0.221	0.274	-0.370	0.164	-0.180	-0.097
xGF%	0.469	0.497	0.525	0.471	0.510	0.487	0.486
xGF%Rel	-0.024	-0.001	0.024	-0.025	0.010	-0.008	-0.014
Giveaways	1.270	1.513	1.918	1.463	1.872	1.740	1.793
Takeaways	1.512	1.618	1.848	1.277	1.125	0.846	0.818
Blocks	2.137	1.747	1.441	2.220	3.737	4.214	4.672
Hits	6.122	4.668	3.023	12.749	3.427	4.229	6.279
Net hits	1.006	-0.202	-0.976	6.655	-0.874	-1.008	0.822
Fights	0.058	0.024	0.019	0.460	0.020	0.027	0.107
Penalties	0.691	0.661	0.642	2.007	0.629	0.573	0.929
Net penalties	-0.029	-0.080	-0.207	0.456	0.207	0.187	0.398
Penalty minutes	3.591	3.299	3.368	11.621	2.310	2.145	3.796
Physical penalties drawn	0.145	0.120	0.127	0.833	0.094	0.101	0.228
Physical penalties on	0.148	0.117	0.119	0.921	0.091	0.093	0.244
Restraining penalties drawn	0.395	0.425	0.488	0.459	0.208	0.176	0.170
Restraining penalties on	0.330	0.333	0.309	0.529	0.340	0.321	0.440
Stick penalties drawn	0.162	0.182	0.217	0.189	0.113	0.102	0.119
Stick penalties on	0.151	0.157	0.159	0.318	0.148	0.117	0.180
OZS%	0.502	0.501	0.500	0.499	0.500	0.500	0.498
PP%	0.161	0.303	0.522	0.087	0.340	0.149	0.060
SH%	0.205	0.144	0.140	0.124	0.287	0.281	0.359
OT%	0.121	0.225	0.464	0.035	0.423	0.192	0.120
Median X Blocker	-61.418	-61.600	-62.203	-60.421	-71.531	-71.816	-72.058
Median X Giveaway	-9.247	7.064	22.694	-12.545	-54.876	-62.295	-64.705
Median X Hit taken	46.848	52.974	56.480	42.968	-79.173	-82.887	-83.735
Median X Hitter	50.035	53.585	45.437	63.619	-74.427	-75.385	-72.842
Median X Penalty drawn	38.681	48.730	55.660	29.017	-44.261	-51.050	-52.665
Median X Penalty	-1.786	7.576	2.976	16.351	-66.892	-67.643	-68.462
Median X Shooter	69.933	70.348	69.813	70.270	51.044	49.973	48.561
Median X Takeaway	4.481	12.296	8.304	7.197	-43.224	-49.087	-48.726
Median Y Blocker	3.240	3.550	3.366	4.741	2.374	2.665	2.824
Median Y Giveaway	13.743	11.202	8.965	15.791	11.802	12.917	14.816
Median Y Hit taken	14.749	13.006	11.718	17.486	19.306	19.300	21.594
Median Y Hitter	12.296	12.299	12.062	11.374	19.586	20.391	23.384
Median Y Penalty drawn	10.509	8.171	4.680	6.694	11.419	12.341	10.447
Median Y Penalty	10.644	9.845	7.487	6.765	6.644	9.580	7.123
Median Y Shooter	1.855	1.680	1.872	1.931	8.742	11.059	13.364
Median Y Takeaway	8.543	7.226	5.514	11.394	14.730	16.196	16.270
Height (inches)	72.832	72.776	72.648	74.357	73.291	73.318	74.797
Weight (pounds)	196.294	195.941	195.925	211.659	199.538	198.207	208.955

[a] *Examples*: Nick Bonino, Colton Sissons, Barclay Goodrow
[b] *Examples*: Anze Kopitar, Jamie Benn, Tyler Seguin
[c] *Examples*: Sidney Crosby, Auston Matthews, Connor McDavid
[d] *Examples*: Tanner Jeannot, Ryan Reaves, Pat Maroon
[e] *Examples*: Adam Larsson, Rasmus Ristolainen, Radko Gudas
[f] *Examples*: Roman Josi, Victor Hedman, Cale Makar
[g] *Examples*: Ivan Provorov, Christopher Tanev, Brian Dumoulin

5 Comparing player salary

One use of the clusters is to compare similar players regarding their roles with respect to salary. We compared each player's cap hit to their ten nearest neighbors to determine how the player's cap hit compares to their peers. The definition of neighbor in this context is based on selecting ten players with whom a given player has the lowest Euclidean distance, where the distance is measured by considering the fuzzy cluster membership probabilities. Due to the very right-skewed distributions of cap hit a logarithmic transformation was used. After computing the difference in cap hit and distance for a pair of players, the difference is then divided by the player's own cap hit and then considered as the basis for determining if a player is underpaid or overpaid when considering the cap hit of their neighbors. A summary of the fifteen most underpaid and overpaid players, relative to their neighbors, per position can be seen in Table 2. In general, a negative value of average difference indicates that a player is earning less than similarly performing players, while a positive value suggests the opposite. The value should not be interpreted objectively to determine whether a player has a good or bad contract, but rather how their contract stands in relation to players whose role was similar during the 2021-2022 season.

The results indicate that the most overpaid players, relative to their role, are Oliver Ekman-Larsson, Sean Monahan, and Milan Lucic, while the players who are deemed to be most underpaid are Oliver Kylington, Trevor Zegras, and Mason Marchment. A shared attribute among many underpaid players is that they are still on their entry-level contract, which is the first contract they sign when entering the league. As such, their true value may not (yet) be seen in their contract. However, some of the players that are suggested to be underpaid have after the season signed more lucrative contracts, including e.g., Jason Robertson, Mason Marchment, Jack Hughes, and Adam Fox[4]. Moreover, some of the overpaid players have since signed smaller contracts (Anton Strålman), retired (Duncan Keith and P.K. Subban), or are no longer in the league (Alexander Radulov and Danny DeKeyser). Interestingly, players who had a more unique distribution of cluster membership degrees, such as Brady Tkachuk, were more difficult to evaluate, as they can be quite distant to their nearest neighbors. Consequently, they may be overrated by the model while in reality the contract is not as bad as the model describes.

6 Team composition

Team composition can have a substantial impact on team performance [6]. Therefore, we also investigate if there are any patterns between player roles and team success for the given season. We first compute the minutes played for all players per team and retain the 18 players with the highest playing time. The choice of 18 players is based on the roster size in the NHL, where 20 players, of whom 18 are skaters and 2 goaltenders, are allowed to be used in any given game.

[4] `https://www.capfriendly.com/transactions`

Table 2: Most underpaid and overpaid players, relative to players with similar cluster membership probabilities for each position.

(a) Underpaid defenders.

Rank	Player	Avg. Rel. Diff.
1	Oliver Kylington	-1.060
2	Evan Bouchard	-0.812
3	Adam Fox	-0.750
4	Adam Boqvist	-0.726
5	Erik Gustafsson	-0.688
6	Anthony DeAngelo	-0.663
7	Moritz Seider	-0.654
8	Bowen Byram	-0.636
9	Noah Dobson	-0.636
10	Alexandre Carrier	-0.584
11	Rasmus Sandin	-0.565
12	Kale Clague	-0.551
13	Calle Rosen	-0.548
14	Gabriel Carlsson	-0.541
15	Jaycob Megna	-0.536

(b) Overpaid defenders.

Rank	Player	Avg. Rel. Diff.
1	Oliver Ekman-Larsson	0.484
2	Esa Lindell	0.416
3	Ryan McDonagh	0.410
4	Marc-Edouard Vlasic	0.402
5	Jeff Petry	0.382
6	Anton Strålman	0.380
7	Nick Leddy	0.375
8	T.J. Brodie	0.374
9	Darnell Nurse	0.367
10	Danny DeKeyser	0.357
11	P.K. Subban	0.352
12	Duncan Keith	0.352
13	Tyler Myers	0.348
14	Rasmus Ristolainen	0.348
15	Ryan Pulock	0.345

(c) Underpaid forwards.

Rank	Player	Avg. Rel. Diff.
1	Trevor Zegras	-0.927
2	Mason Marchment	-0.889
3	Jason Robertson	-0.869
4	Joshua Norris	-0.837
5	Jack Hughes	-0.800
6	Matthew Boldy	-0.795
7	Anton Lundell	-0.779
8	Martin Necas	-0.759
9	Tim Stützle	-0.744
10	Michael Bunting	-0.717
11	Carter Verhaeghe	-0.715
12	Nathan Walker	-0.703
13	Nick Suzuki	-0.673
14	Cole Caufield	-0.672
15	Lucas Raymond	-0.655

(d) Overpaid forwards.

Rank	Player	Avg. Rel. Diff.
1	Sean Monahan	0.483
2	Milan Lucic	0.438
3	Brady Tkachuk	0.418
4	Jonathan Drouin	0.418
5	Antohy Beauvillier	0.397
6	Tyler Johnson	0.394
7	Jamie Benn	0.392
8	Kevin Hayes	0.387
9	Andrew Ladd	0.385
10	Alexander Radulov	0.374
11	Dustin Brown	0.373
12	Colton Sissons	0.371
13	Christian Dvorak	0.356
14	Nick Foligno	0.356
15	Niklas Bäckstrom	0.353

Thus, these 18 players can then represent a possible composition of players for any given team and game. Except for the San Jose Sharks (8D / 10F) and the Florida Panthers (5D / 13F), the team compositions either consisted of 7 defenders and 11 forwards or 6 defenders and 12 forwards. Next, for a given team we then sum the cluster probabilities among all players in each position group (defenders and forwards) to obtain an estimate of how many players they have in each role, which is then divided by the total number of players for the given position to find the proportion of roles each team has. Thus, the sum of all forward clusters sums to one and likewise for defenders. This is illustrated in Fig. 3, where a hierarchical clustering using Ward's linkage method and Euclidean distance groups the playoff and non-playoff teams by team composition.

An observation from the clustering is that a distinction between playoff and non-playoff teams is apparent, as 14 out of 16 playoff teams were grouped together with the two exceptions being the New York Rangers and the Dallas Stars. Similarly, the Anaheim Ducks and Vegas Golden Knights, who both missed the playoffs, were clustered with the other playoff teams. In general, the playoff team

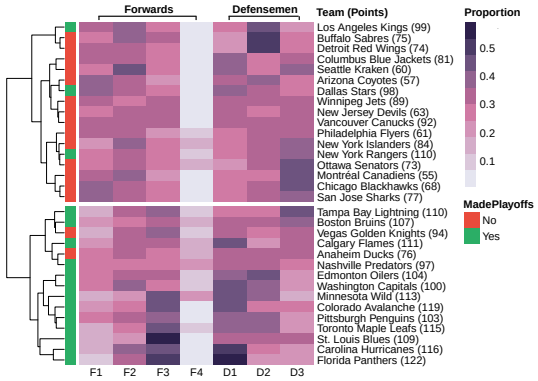

Fig. 3: Team composition and playoffs.

cluster had a higher proportion of forwards from role F3, while the proportions of F1 and F2 were lower than the corresponding roles for the non-playoff cluster. There did not seem to be any noticeable differences between the two hierarchical clusters with respect to F4, as most teams had few players in this role. Among defenders, the playoff teams tended to have higher proportions of role D1 and fewer players of role D2 and D3. Conversely, D2 and D3 were more common among the non-playoff teams, which consequently implies a lower proportion of D1. For the incorrectly clustered teams some additional information may shed light on how they were clustered. In particular, by contrasting the Dallas Stars and Vegas Golden Knights we note a point differential of 4 in favor of Dallas, while Dallas scored 29 fewer goals than Vegas. This could indicate that Vegas was more offensively capable but less consistent. Both teams concede the same number of goals. For the New York Rangers the offensive capabilities were league average, as their goals scored ranked 16^{th} out of 32 teams but they had 2^{nd} fewest goals conceded, which can be attributed to their goaltender Igor Shesterkin who was voted the top goalie during the season. Finally, the Anaheim Ducks had an even distribution of roles and thus may be closer in distance to many teams.

7 Conclusion

In this paper we have proposed a novel method for quantifying player roles in ice hockey from a large set of performance indicators and player data using fuzzy c-means. We also investigated the application of comparative contract evaluation for the comparison of salary and player role, which can be used as a component in decision-making regarding contract negotiation and player acquisition. Moreover, an investigation of the relation between player roles and team success gave insight into what roles may provide additional success for a team.

Some limitations are worth mentioning. The data upon which this study is based is not bias-free and does not cover all events that occur in an ice hockey

game. This is particularly evident for evaluating the defensive contribution of players. In regard to the contract evaluation, it is dependent on the chosen distance metric and number of neighbors. For instance, by choosing the maximum number of neighbors the league's highest paid players are deemed the most overrated. In addition, the highest paid players in the league cannot be underpaid, as there is nobody or very few paid more than them. Lastly, there is the possibility that the team that a player plays for may be a latent factor unaccounted for in this analysis, since a player's style of play may differ between teams and their performance may also be affected.

An extension of this work could be to include variables not available in the data used here that can further distinguish between player roles, e.g., passes and zone entries. Our method could easily be extended to capture these new variables. Additionally, by analyzing multiple seasons the results would also highlight changes in performance and role over a player's career. This method can also be generalized and applied to other leagues around the world, as the style of play may differ between leagues.

References

1. Aalbers, B., Haaren, J.V.: Distinguishing between roles of football players in play-by-play match event data. In: Proceedings of the 5th Workshop on Machine Learning and Data Mining for Sports Analytics co-located with 2018 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2018), Dublin, Ireland, September 10th, 2018. CEUR Workshop Proceedings, vol. 2284, pp. 31–41. CEUR-WS.org (2018). https://doi.org/10.1007/978-3-030-17274-9_3
2. Assent, I.: Clustering high dimensional data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **2**(4), 340–350 (2012). https://doi.org/10.1002/widm.1062
3. Bezdek, J.: Pattern Recognition With Fuzzy Objective Function Algorithms. Springer New York, NY (1981). https://doi.org/10.1007/978-1-4757-0450-1
4. Bezdek, J.C., Ehrlich, R., Full, W.: FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences **10**(2-3), 191–203 (1984). https://doi.org/10.1016/0098-3004(84)90020-7
5. Campello, R.J., Hruschka, E.R.: A fuzzy extension of the silhouette width criterion for cluster analysis. Fuzzy Sets and Systems **157**(21), 2858–2875 (2006). https://doi.org/10.1016/j.fss.2006.07.006
6. Chan, T.C., Cho, J.A., Novati, D.C.: Quantifying the contribution of NHL player types to team performance. Interfaces **42**(2), 131–145 (2012). https://doi.org/10.1287/inte.1110.0612
7. Dave, R.N.: Validating fuzzy partitions obtained through c-shells clustering. Pattern Recognition Letters **17**(6), 613–623 (1996). https://doi.org/10.1016/0167-8655(96)00026-8
8. Dembele, D., Kastner, P.: Fuzzy c-means method for clustering microarray data. Bioinformatics **19**(8), 973–980 (2003). https://doi.org/10.1093/bioinformatics/btg119
9. Dunn, J.: A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. Journal of Cybernetics **3**(3), 32–57 (1973). https://doi.org/10.1080/01969727308546046

10. Felmet, G.: Ice hockey. In: Krutsch, W., Mayr, H.O., Musahl, V., Della Villa, F., Tscholl, P.M., Jones, H. (eds.) Injury and Health Risk Management in Sports: A Guide to Decision Making, pp. 485–489. Springer Berlin Heidelberg (2020). https://doi.org/10.1007/978-3-662-60752-7_74
11. Ferraro, M.B., Giordani, P.: A toolbox for fuzzy clustering using the R programming language. Fuzzy Sets and Systems **279**, 1–16 (2015). https://doi.org/10.1016/j.fss.2015.05.001
12. Glorfeld, L.W.: An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. Educational and Psychological Measurement **55**(3), 377–393 (1995). https://doi.org/10.1177/0013164495055003
13. Horn, J.L.: A rationale and test for the number of factors in factor analysis. Psychometrika **30**(2), 179–185 (1965). https://doi.org/10.1007/BF02289447
14. Lefebvre, J.S., Martin, L.J., Côté, J., Cowburn, I.: Investigating the process through which National Hockey League player development coaches 'develop'athletes: An exploratory qualitative analysis. Journal of Applied Sport Psychology **34**(1), 47–66 (2022). https://doi.org/10.1080/10413200.2019.1688893
15. Miyamoto, S., Ichihashi, H., Honda, K.: Algorithms for Fuzzy Clustering. Springer Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78737-2
16. Pakhira, M.K., Bandyopadhyay, S., Maulik, U.: Validity index for crisp and fuzzy clusters. Pattern Recognition **37**(3), 487–501 (2004). https://doi.org/10.1016/j.patcog.2003.06.005
17. Peres-Neto, P.R., Jackson, D.A., Somers, K.M.: How many principal components? stopping rules for determining the number of non-trivial axes revisited. Computational Statistics & Data Analysis **49**(4), 974–997 (2005). https://doi.org/10.1016/j.csda.2004.06.015
18. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53–65 (1987). https://doi.org/10.1016/0377-0427(87)90125-7
19. Sattari, A., Johansson, U., Wilderoth, E., Jakupovic, J., Larsson-Green, P.: The interpretable representation of football player roles based on passing/receiving patterns. In: Brefeld, U., Davis, J., Van Haaren, J., Zimmermann, A. (eds.) Machine Learning and Data Mining for Sports Analytics. pp. 62–76. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-02044-5_6
20. Schwämmle, V., Jensen, O.N.: A simple and fast method to determine the parameters for fuzzy c–means cluster analysis. Bioinformatics **26**(22), 2841–2848 (2010). https://doi.org/10.1093/bioinformatics/btq534
21. Vincent, C.B., Eastman, B.: Defining the style of play in the nhl: An application of cluster analysis. Journal of Quantitative Analysis in Sports **5**(1) (2009). https://doi.org/10.2202/1559-0410.1133
22. Vollman, R.: Hockey Abstract Presents... Stat Shot: The Ultimate Guide to Hockey Analytics. ECW Press (2016)
23. Wang, W., Zhang, Y.: On fuzzy cluster validity indices. Fuzzy Sets and Systems **158**(19), 2095–2117 (2007). https://doi.org/10.1016/j.fss.2007.03.004
24. Wierzchoń, S.T., Kłopotek, M.A.: Modern Algorithms of Cluster Analysis. Springer Cham (2018). https://doi.org/10.1007/978-3-319-69308-8