Clustering SHL teams using Offensive Zone Metrics*

Ali Raza Zaidi, Fauzan Lodhi , Xikai Wu, Ryan Cao, and David Awosoga

University of Waterloo, Waterloo ON N2L 3G1, CAN

Abstract. Advance scouting is essential to adequately prepare teams for upcoming contests. Traditional process involve spending copious amounts of time reviewing video and visually identifying key items. To expedite this process and more efficiently guide film study, this work analyzes offensive zone actions to identify and categorize patterns in strategy.

Keywords: hockey analytics · spatiotemporal data · sports statistics.

1 Introduction

An archetypal invasion-type game, ice hockey is built upon strategies to generate offensive opportunities and limit the quantity and quality of chances for the opposing team. Given the storied past of the sport, numerous approaches have been developed by teams to maximize their per-possession efficiency given their roster construction. In particular, insights into offensive zone strategies quantify how teams formulate an attack against their opponent, how these strategies vary within the league and what combinations of strategies are often used in conjunction with one another. The following analysis clusters teams by their offensive zone strategies within the Swedish Hockey League to analyze the different play styles within the league. Similar approaches have been utilized on a team level in other invasion sports such as soccer [1], but hockey-specific analyses either analyzed different facets of the game [2], the events the leading up to an offensive possession [3], or on individual player characteristics [4]. Therefore, utilizing a team-centered approach focusing on archetypes for offensive zone tendencies distinguishes this work from any predecessors and provides exciting grounds for analysis and exploration. The framework for this analysis identifies key events based on a combination of accumulated and derived metrics that reveal team tendencies on offense.

2 Background

This analysis is performed on event-level data from one hundred and fifty-six games from the Swedish Hockey League's (SHL) 2023-24 season. Each event tracked several features including the game ID, team ID, elapsed game time, location on the ice, team in possession and event type. Possessions in the offensive zone (o-zone) were tracked to determine the o-zone time, and events occurring

* Supported by the University of Waterloo Analytics Group for Games and Sports.

within this o-zone shift were measured to determine their frequency. Within the o-zone, the attacking team uses their possession to generate scoring opportunities. Certain offensive zone strategies emphasize high-risk plays that provide better-scoring chances but risk losing possession earlier. Other strategies prefer low-risk events where puck possession is paramount. Higher risk strategies will generally have a lower average offensive zone time, but will counteract that with higher quality and more dangerous shifts.

3 Methods and Algorithms

O-zone plays occur when a team makes a play within the opposing team's blue line until the end boards within the same half. Offensive zone shifts are defined as consecutive possessions occurring within the offensive zone. They start when an attacking team enters the offensive zone, or gains possession of the puck while already in the offensive zone, and finish when the defensive team gets possession of the puck, or when the puck leaves the offensive zone. The following events are marked as favourable outcomes for the offensive team and are tracked accordingly: **passes**, **shots**, **penalties drawn**, **goals scored**, **loose puck retrievals**, and **icings drawn**. We also derive secondary metrics based on the offensive zone time between the selected events. To justify the metrics chosen, passes and loose puck retrievals were tracked as they indicate control of the play, while shots (and goals) indicate the 'end goal'for the shift. Opposing team icings and penalties drawn indicate a successful o-zone strategy that forces the defensive team to commit a mistake. Cohesively, these paint a picture of a successful o-zone shift.

With this in mind, the idea is to use these metrics to cluster teams using unsupervised learning algorithms. Given the small number of teams (n=14) within our dataset, it was hypothesized that clustering algorithms like k-means may not be the best fit. Instead, the approach taken was that of `Gaussian Mixture Model`, which is formed from several Gaussian models describing the underlying domain [5]. The model generated was trained on the aggregate statistics, listed in Table 1. Additionally, the data was cleaned to remove non-regulation-time and non-even-strength events to ensure the data was unbiased, and it was also then scaled.

Due to the small number of teams, we capped the number of clusters to a maximum of 6 to try and ensure that we wouldn't get many clusters of a single point. We also used Principal Component Analysis (PCA) to reduce the data's dimensionality, building visual intuition for how many clusters to expect [6]. Our PCA process was able to explain 96% of the variance with just 5 components. Then, based on manual inspection of the PCA graphs, present in Figure 1 below, it was determined that 5 clusters seemed to be the best case. We corroborated this using analysis of BIC scores [7], as in Figure 2 which when capped to 6 clusters, yielded its best result at $k = 5$ clusters, with the 5-cluster GMM almost exactly matching the PCA clusterings.

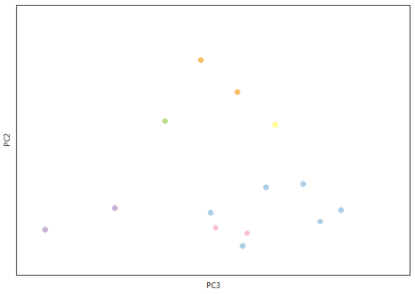

Fig. 1. PCA Clustering

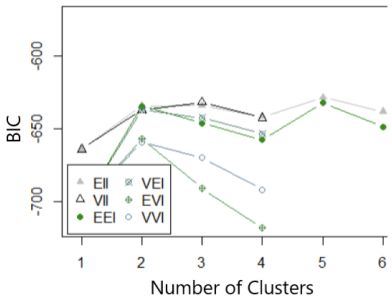

Fig. 2. BIC Clustering

4 Overview and Discussion of Findings

After training the `GMM` with 5 specified clusters, groupings emerged relating certain metrics to certain teams. We collected this data into a heatmap, labelled as Figure 3 below. Table 1 below highlights what each of the metrics highlighted in the analysis below are specifying. All derived metrics are reported in terms of their average values.

Table 1. Definitions of the Derived Metrics

Event	Description
time	time between shots
time_mod	time between shots*
pass	passes between shots
pass_mod	passes between shots*
ozone_time	length of the o-zone shift
max_xG	max xG generated within the o-zone shift
cmltv_xG	cumulative xG generated within the o-zone shift
goals	goals scored per the o-zone shift
passes	passes recorded within the o-zone shift
slotpasses	passes made to teammates in the 'slot'within the o-zone shift
fwdtime	length of time spent by forwards in the o-zone during the shift
dmentime	length of time spent by defencemen in the o-zone during the shift
penalties	penalties drawn within the o-zone shift
puckretrievals	successful puck retrievals within the o-zone shift
icings	opposing team icings committed within the o-zone shift

*shots without a pass in between are excluded

`Opportunistic Offence`: Based on the heatmap in Figure 3, we can see that cluster 1 generally has lower metrics than any of the other clusters, bar its goals and shots. Their low o-zone shift length and xG stats indicate that they struggle to maintain dangerous o-zone pressure. However, their fair number of shots and above-average goals indicate that they are finishing well on the chances given. Thus this cluster has been labelled as an opportunistic offence given their propensity to capitalize on the few chances they get.

`Chip & Chase`: Cluster 2 is more varied in terms of metrics, with below-average o-zone time, opposing team icings and defensemen o-zone time but

Mean Metric Values for Each Cluster

Metric (mean)	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
time	8.475	9.173	9.187	10.072	10.133
time_mod	10.556	11.062	11.359	12.082	11.996
pass	2.155	2.425	2.264	2.681	2.724
pass_mod	2.684	2.923	2.799	3.217	3.225
ozone_time	6.516	7.024	6.577	7.808	7.385
max_xG	0.041	0.049	0.052	0.051	0.045
cmltv_xG	0.042	0.051	0.053	0.053	0.046
goals	0.052	0.048	0.054	0.042	0.059

Mean Metric Values for Each Cluster

Metric (mean)	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
shots	1.113	1.122	1.101	1.131	1.097
passes	1.561	1.830	1.579	1.984	1.963
slotpasses	0.239	0.310	0.271	0.317	0.255
fwdtime	3.813	4.212	3.767	4.518	4.178
dmentime	1.897	1.876	1.896	2.308	2.262
penalties	0.005	0.008	0.013	0.009	0.006
puckretrievals	0.893	0.952	0.876	0.964	0.928
icings	0.009	0.007	0.011	0.014	0.008

Fig. 3. Identified Clusters

above-average stats in xG, slot passes, and puck retrievals. The cluster struggles to keep the puck within the offensive zone, especially with a lack of a defenceman anchoring the offence but also creates high-danger chances close to the net. Their puck retrieval stats also hint at the chip-and-chase nature of this cluster, where quality net-front opportunities are created through forechecking, at the cost of stability and longevity within the o-zone.

`Off the Rush`: Cluster 3 seems to have below-average metrics in most categories, except for xG, goals and penalties drawn. This would be indicative of a team relying on chances off the rush where opportunities would often be shorter in length, but with absurd xG, goal and penalties drawn metrics, given the possible odd-man rushes.

`High-Pressure Cycling`: Cluster 4 has some of the highest metrics, particularly for o-zone time, slot passes, xG and puck retrievals, with their only below-average metrics being goals and penalties drawn. This indicates a possession-heavy game, and their command of the middle of the ice gives them high-quality scoring opportunities using slot passes. They also seem to cycle the puck and force opposing teams to commit mistakes like icings. Their below-average goals may indicate a lack of finishing ability.

`Point-driven Offence`: Cluster 5 has above-average metrics in goals, defencemen time, and passes, but below-average metrics for xG, slot passes and shots. This indicates a perimeter-driven offense as they pass the puck often between the defencemen anchoring the offence but not often to the slot. There point shots would be less frequent and would have a lower xG but the traffic created by the offence would be vital in scoring on more shots than expected.

Within each cluster, certain attributes display a significant amount of variance. This is most notable within cluster 2 for the 'goals'attribute which range from 0.034 up to 0.063. This highlights the clustering occurring based on the core playing style rather than team competency. Table 2 below lists the teams in

Table 2. Teams within each cluster

Cluster	1	2	3	4	5
Teams	795	855, 869, 814, 877, 792	524, 885	503, 686, 726	825, 634, 628

each cluster. The anomalous nature of team 795 is highlighted as it is the only team playing its weak playstyle, but they make it work.

These results show that several different playstyles, along with their strengths and weaknesses, can be seen within the 2023-24 SHL season, though the degree of effectiveness of any given strategy will depend on the team employing them.

5 Conclusion and Future Steps

This analysis determined the 5 main types of team playstyles that exist within the 2023-24 SHL season, and where the teams fit into said clusters. These clusters can be used to analytically determine a team's preferred play style, and plan for it. With more games of data, we might be able to flesh out the model with more information, and with data from other leagues like the NHL which contain more teams, or even previous SHL seasons' worth of data, we would be able to cluster with more teams per label. Additionally, with player-tracking data, it may be possible to further define playing styles based on the whole team's positioning.

6 Code Access Links

The code used in this project can be accessed here: `https://github.com/AliRZ-02/linhac2024`

References

1. Coutinho, D., Gonçalves, B., Laakso, T., & Travassos, B.: Clustering ball possession duration according to players' role in football small-sided games. PloS one, 17(8), e0273460. (2022) `https://doi.org/10.1371/journal.pone.0273460`
2. Radke, D., Brecht, T., & Radke, D.: Identifying Completed Pass Types and Improving Passing Lane Models. Linköping Electronic Conference Proceedings, 71–86. (2022) `https://doi.org/10.3384/ecp191008`
3. Olivestam, A., Rosendahl, A., Hampus Svens, H., & Hellberg, L.: On the Attack: Using Analytics to Unlock the Secrets of Successful Zone Entries in Hockey. LINHAC 2023 (2023). `https://www.ida.liu.se/research/sportsanalytics/LINHAC/LINHAC23/papers/paper-students-LiU.pdf`
4. Stimson, R.: Identifying Playing Styles with Clustering. Hockey-Graphs (2017). `https://hockey-graphs.com/2017/04/04/identifying-player-types-with-clustering/`
5. Scikit-Learn: Gaussian Mixture Models `https://scikit-learn.org/stable/modules/mixture.html#gmm`
6. SciKit-Learn: Principal component analysis (PCA) `https://scikit-learn.org/stable/modules/decomposition.html#pca`
7. Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E.: mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models, 293 (2016) `https://journal.r-project.org/archive/2016/RJ-2016-021/RJ-2016-021.pdf`