Visualizing Playstyle in Women's Hockey

Introduction

Is it possible to visualize a player's role using event tracking data? Using previous research from @samgoldbergTHFC (*https://www.americansocceranalysis.com/home/2020/3/3/clustering*) and @johnspacemuller (*https://spacespacespaceletter.com/the-seven-styles-of-soccer/)*, I used k-means clustering to classify playstyles based on several attributes.

Using the k-means clustering algorithm, players were classified into 6 groups based on the following statistics from even-strength events:

1. Shot Attempts
2. Average Shot Distance
3. Pass Completion Percentage
4. Average Pass Length
5. Total Recoveries and Takeaways

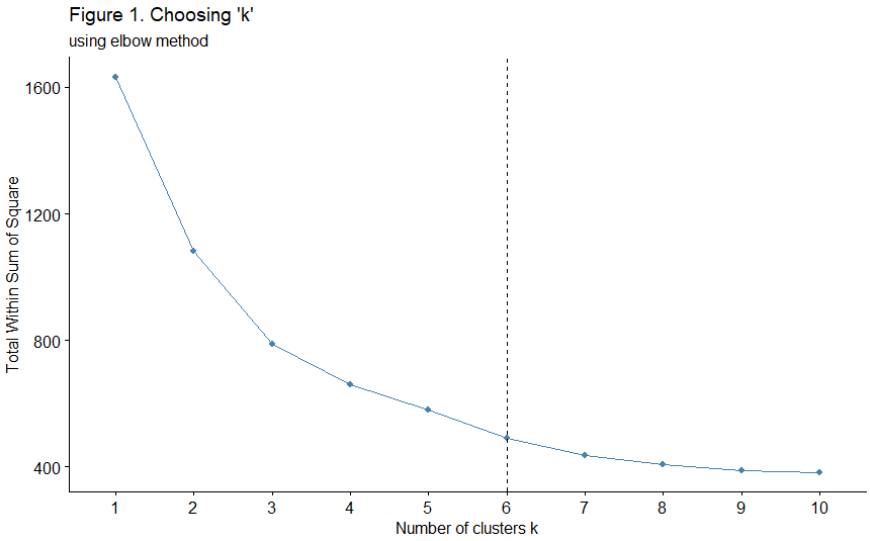
The algorithm classified players into the following groups, named based on the cluster's key characteristics:

Cluster	Description
1	Low Event Skaters
2	Low Event Goalies
3	Defensive Defenders
4	Offensive Defenders
5	Playmakers
6	Shooters

The algorithm used volume statistics and can be improved with the use of predictive metrics allowing general managers to identify similar players for contract negotiations.

K-Means Clustering Algorithm

I used a k-means clustering algorithm to classify players into *k* clusters such that players in each group are similar to each other based on the attributes listed above. I pre-specified the number of clusters in order to minimize the within-cluster variation. *Figure 1* shows the selection of *k* using the "elbow" method. The "elbow" is the point where increasing the number of clusters does not reduce variation.



Principal Component Analysis

In order to visualize clusters using more than 2 variables, I reduced the number of variables to 2 dimensions using a *Principal Component Analysis (PCA)*. The purpose of the PCA algorithm is to reduce the number of variables for simplicity while retaining as much information as possible. *Figure 2* shows the *k* = 6 clusters produced using the first 2 dimensions of the PCA algorithm. Note that 72% of the variability is explained by the first 2 dimensions.

Figure 2. Women's Hockey Playstyle Clusters

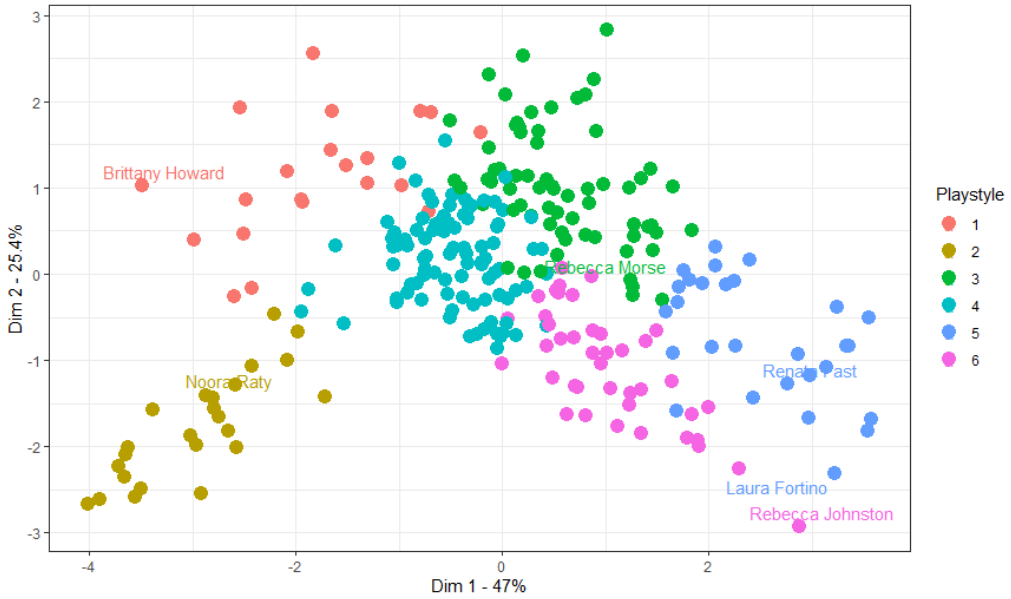


Table 1 shows the variable means for each cluster. Each cluster's average number of events recorded was included in the table, but not in the clustering algorithm. However, we see that players with more recorded events were placed in clusters 5 and 6. This indicates that volume statistics (Shot Attempts, Total Recoveries) contributed heavily to the algorithm's classification. In the future, I want to leverage data that is more predictive in nature (see Limitations). Players in clusters 5 and 6 have the puck more often than their counterparts, but that does not necessarily describe their true value to the team.

Table 1. Attribute Means

Cluster	Shot Attempts	Avg Shot Distance	Pass Completion %	Avg Pass Length	Total Recoveries	Total Events
1	1.5	9.4	0.4	27.0	9.0	24.5
2	0.0	0.3	0.9	13.9	23.8	31.5
3	6.0	50.5	0.6	39.3	54.9	118.4
4	6.9	28.6	0.7	33.8	39.1	101.8
5	19.8	50.3	0.7	39.8	175.1	355.0
6	24.2	27.7	0.6	32.3	105.4	271.0

Results

Clusters 1 and 2 were low event players or goalies who rarely touched the puck. Clusters 3 and 4 were involved in more events but were not close to clusters 5 and 6 in total events. Ignoring flaws in the model, here are some notable takeaways from each cluster.

Cluster 1 – Low Event Players

This cluster of players were involved in few events either because they did not play in enough games, or they were fourth liners. Team Canada's Brittany Howard was involved in only 3 events all at the same time, making it into cluster 1.

Cluster 2 - Goalies

Almost all players from this cluster are goalies. The only exception was Kalie Grant of the Clarkson Golden Knights. *Figure 3* shows the location of even strength events from players in cluster 2. If I removed Kalie Grant from the data, there would not be any points outside the blob of points by the defensive goal area.

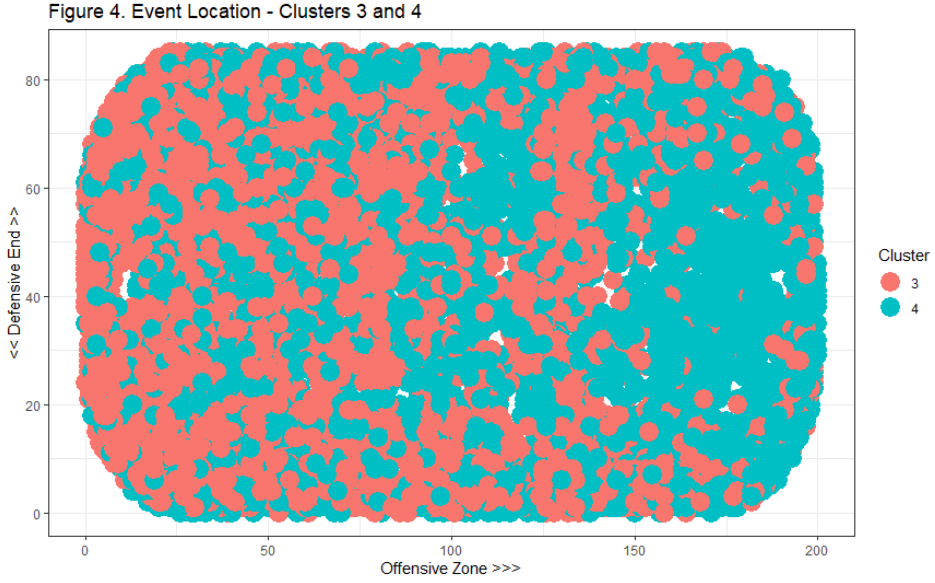
Figure 3. Cluster 2 Event Location

Cluster 3 and 4 – Defensive vs. Offensive

Players in cluster 3 attempted longer passes but completed a lower percentage of passes than players in cluster 4. Cluster 3 shot distance was significantly further from the net compared to cluster 4. Also, Cluster 3 averaged more takeaways/recoveries. These differences indicate that

cluster 3 players are more defensive-minded, while players from cluster 4 are offensive-minded. *Figure 4* shows Cluster 3's events take place in the defensive zone and around the edges of the offensive zone, while cluster 4 events tend to stay up the middle of the ice and towards the net.

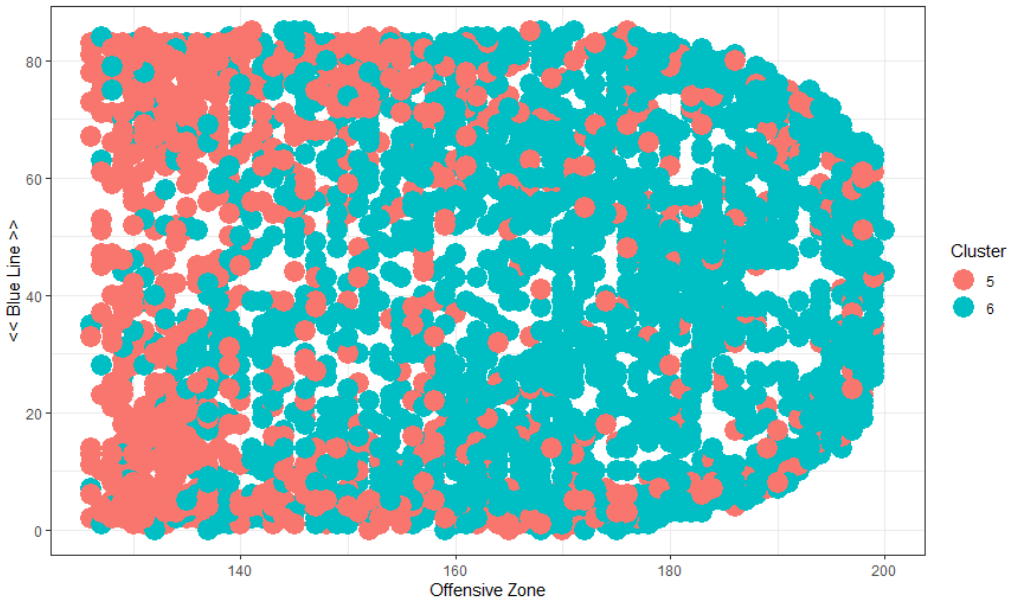
Figure 4. Event Location - Clusters 3 and 4



Cluster 5 and 6 – Playmakers vs. Shooters

Clusters 5 and 6 include players with the most events recorded by a wide margin. *Figure 5* shows pass release points and shot location for even-strength events in the offensive zone.

Figure 5. Pass and Shot Location - Clusters 5 and 6



These clusters are comparable because they both include high-event players. Cluster 6 (shooters) pass and shot release points are located close to the net, while cluster 5 events of this type tend to stay near the blue line and on the outer edges of the offensive zone.

Limitations

The most glaring limitation was that the clustering algorithm's classification was reliant on volume variables. Players with more recorded events were placed in clusters 5 and 6, while low-event skaters (Cluster 1) and goalies (Cluster 2) had the least number of recorded events on average. In the future, I would analyze a subset of players (ex: high-volume forwards). Then, the use of k-means clustering with *predictive* variables would enable high-volume forwards to be classified into more granular playstyles. I hypothesize that players who commit low-value and high-value events would be classified separately.

Conclusion

Although this model had limited predictive capabilities, I maintain that the use of k-means clustering is valuable for teams seeking to identify common players. Determining a player's intrinsic value is important when the time comes to sign new players *and* evaluate current players. However, to avoid a model that classifies based on playing time, it is necessary to include predictive data in the clustering algorithm. This makes it possible to identify 4th liners that have potential for greatness, or alert teams when a high-volume player is not providing value anymore.