Title: Using Clustering to Identify A Player's Passing Efficiency After A Takeaway

By: Ian Madray, **Category:** Scouting Data - Open

Abstract: Understanding how defenses transition into offense is an important aspect of hockey. Using a Poisson regression model, predicted takeaways and the response of passing the puck after a takeaway were computed. A k-means cluster was then formed to identify how players perform with achieving takeaways vs. how well they perform in advancing the puck, specifically passing, after a takeaway.

Introduction: In the scouting dataset provided in this competition, the data is explicitly focused on who has possession of the puck. As a result, with a few caveats, such as complete or incomplete passes, the data lacks positional knowledge of players without the puck. An obvious assessment to make with the data is to focus on the derivation of shot attempts or goals, both offensive attempts. However, how can defensive efforts be represented? For some of the winning entries of the Big Data Bowl 2021, evaluating the defense is focused towards separating plays into discrete sequences of events, to evaluate individual contributions [1]-[2].

In the data, Takeaways – steals, won battles for the puck, or pass interceptions – stand out as a type of event that can give context of defensive plays related to shots or goals. Given that defense is a reaction to offense, understanding how a player performs in transition after a takeaway can evaluate how players react in certain situations, depending on position on an ice rink.

The proposed method to evaluate takeaways will be centered around creating a model to predict the total amount of takeaways per player and how efficient a player is at not committing an incomplete pass after a takeaway. To relate the two features, the players will be clustered by their performances in both categories.

Why are Takeaways Notable?: In the 40-game sample, 293 goals occurred. Within 5 plays after a takeaway, roughly 18% of goals were scored. In addition, nearly 11% of goals were scored within 5 seconds of a takeaway. Thus, understanding how takeaways can set a team's offense is important.

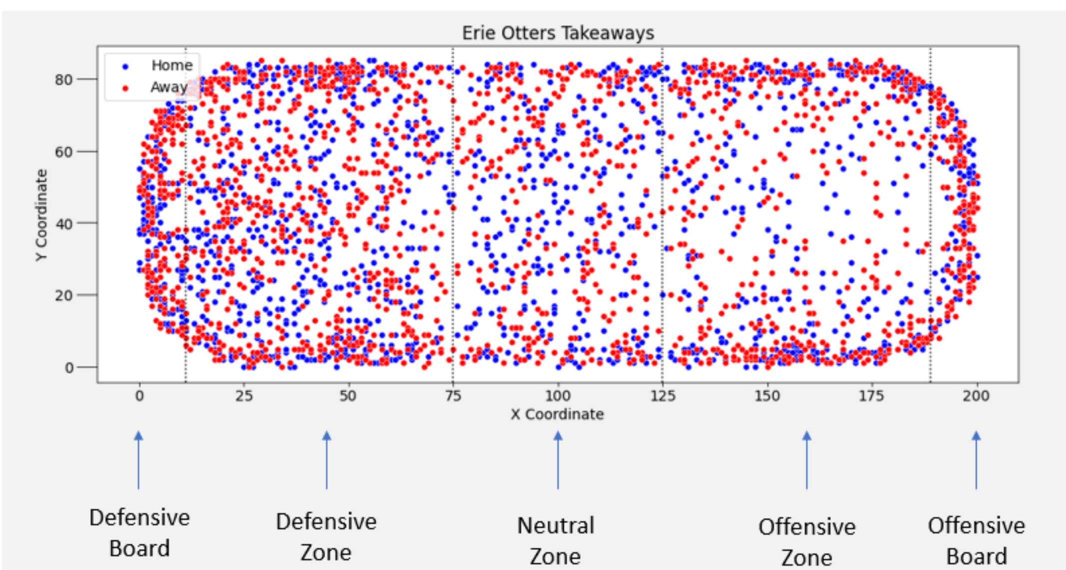


Fig. 1: Takeaways By Zone

The Erie Otters had 1,625 takeaways in 40 games. Each takeaway can be located by its x-y coordinate and zone. Fig. 1 shows where takeaways occur by zone and by home-away status. At first glance, takeaways occur predominately around the boards, and around the defensive area. About 62.3% of the Otters takeaways occur from a recovery from an incomplete pass. When creating a method to evaluate players, understanding the ratio of complete to incomplete passes after a takeaway will be important to highlight. If a player has a lower completion of passes, an opposing team could score off their own takeaway.

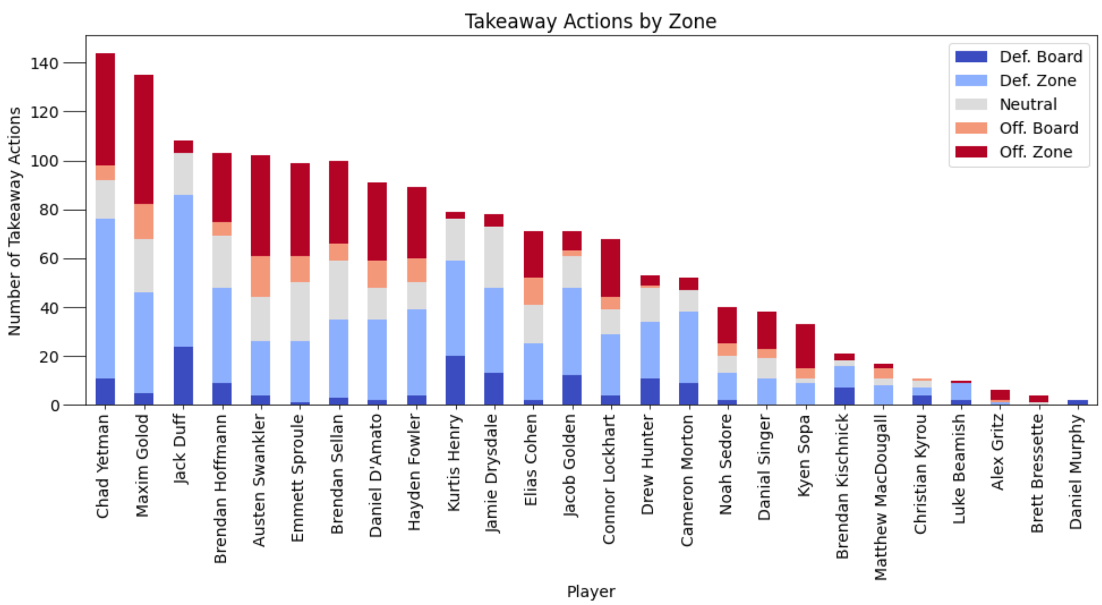


Fig. 2: Takeaway Actions Per Player By Zone

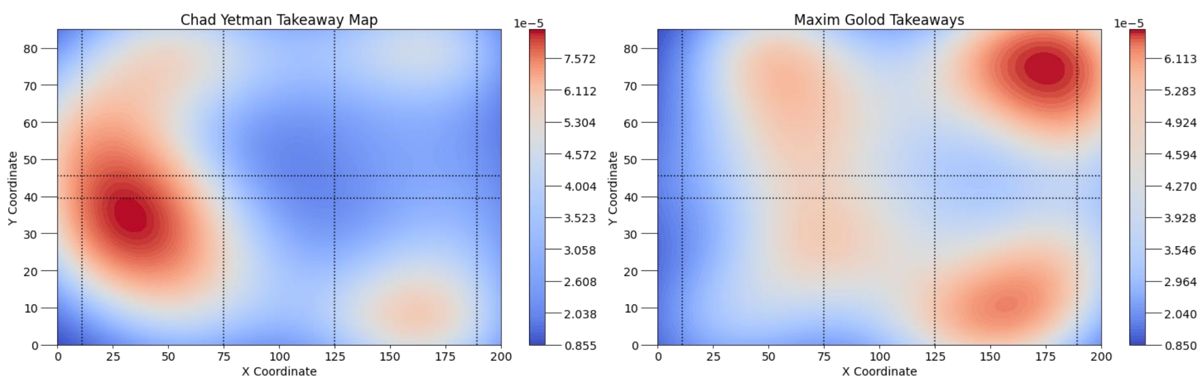


Fig. 3: Heatmaps of Takeaways – The X-Coordinate from 0 -> 200 represents Defensive -> Offensive Zones. The lines from Y = 45.5 and 39.5 and X = 11 and 189 represent the goal line.

Predicting Takeaways: Fig. 2 shows the number of takeaways each Erie Otters player has been credited across 40 games. Each takeaway is broken down into the following regions, with the same logic as Fig. 1's regions: Defensive Board (Def. Board), Defensive Zone (Def. Zone), Neutral, Offensive Zone (Off. Zone), and Offensive Board (Off. Board).

Using the bar plot, different players have takeaways at different zones. For example, Chad Yetman (Center) has more takeaways in the Defensive Zone compared to Maxim Golod (Left Wing) who has more takeaways in the Offensive Zone. With this information in mind, a heatmap can be plotted to determine where a player's takeaways are centered, as shown in Fig. 3.

To model the expected number of takeaways, a Poisson regression model using Statsmodels' generalized linear model class was implemented. The players are filtered by those who have 70 or more takeaways. The features are:

- Position Index (1 = Center, 2 = Left Wing, 3 = Right Wing, 4 = Defenseman),
- Number of takeaways from each individual zone (Def. Board to Off. Board), and
- Mean position of takeaways at x and y (Avgx and Avgy).

Aside from the mean coordinate positions, all the features are count-based data. The output is the actual number of takeaways, or the summation of each takeaway from each zone. Fig. 4 shows the predictions from the model and the differences between the actual takeaways and predicted takeaways.

	Player	Position	Player Number	Position Index	Avgx	Avgy	Def. Board	Def. Zone	Neutral	Off. Zone	Off. Board	Actual Takeaways	Predicted Takeaways	Takeaway Difference
0	Chad Yetman	Center	29	1	89	43	11	65	16	46	6	144	140.0	4.0
1	Maxim Golod	Left Winger	77	2	116	45	5	41	22	53	14	135	133.0	2.0
2	Jack Duff	Defenseman	34	4	42	39	24	62	17	5	0	108	113.0	-5.0
3	Brendan Hoffmann	Right Winger	91	3	91	47	9	39	21	28	6	103	116.0	-13.0
4	Austen Swankler	Center	21	1	126	39	4	22	18	41	17	102	104.0	-2.0
5	Emmett Sproule	Left Winger	93	2	122	41	1	25	24	38	11	99	97.0	2.0
6	Brendan Sellan	Left Winger	44	2	109	46	3	32	24	34	7	100	102.0	-2.0
7	Daniel D'Amato	Right Winger	17	3	110	34	2	33	13	32	11	91	84.0	7.0
8	Hayden Fowler	Right Winger	14	3	105	37	4	35	11	29	10	89	95.0	-6.0
9	Kurtis Henry	Defenseman	41	4	47	36	20	39	17	3	0	79	73.0	6.0
10	Jamie Drysdale	Defenseman	4	4	58	54	13	35	25	5	0	78	75.0	3.0
11	Elias Cohen	Center	13	1	109	44	2	23	16	19	11	71	69.0	2.0
12	Jacob Golden	Defenseman	5	4	57	51	12	36	13	8	2	71	69.0	2.0

Fig. 4: Poisson Regression Model Results: Predicted Takeaways By Player

CI Ratio: Like predicting the number of takeaways per player, an analysis needs to be conducted on what occurs after a takeaway occurs. For example, in Fig. 3, you can see where Chad Yetman and Maxim Golod's positioning based off a takeaway. Fig. 5 shows their activity one event after a takeaway. Despite lack of positional context from surrounding players, inferences can be made. In Chad Yetman's case, since we know incomplete passes are predominately the event that occurs before a takeaway, in the defensive zone, a breakout is more likely to occur and Yetman can dump the puck to either wing or back to a defenseman. Since Maxim Golod is in the offensive region, he can pass it to the center to keep the attacking threat alive.

To show how well a player can keep an offensive possession, the ratio between complete and incomplete passes, or CI Ratio, is also predicted through a Poisson regression model. As a note, other types of events can occur after a takeaway, such as a Dump In/Out, Zone Entry, Puck Recovery. In some of these cases, there is some ambiguity as to the context in which some of these events occur (i.e., did the Otters need to get rid of the puck for a line change). To simplify the model, the focus is only on whether a completion or incompletion occurs.

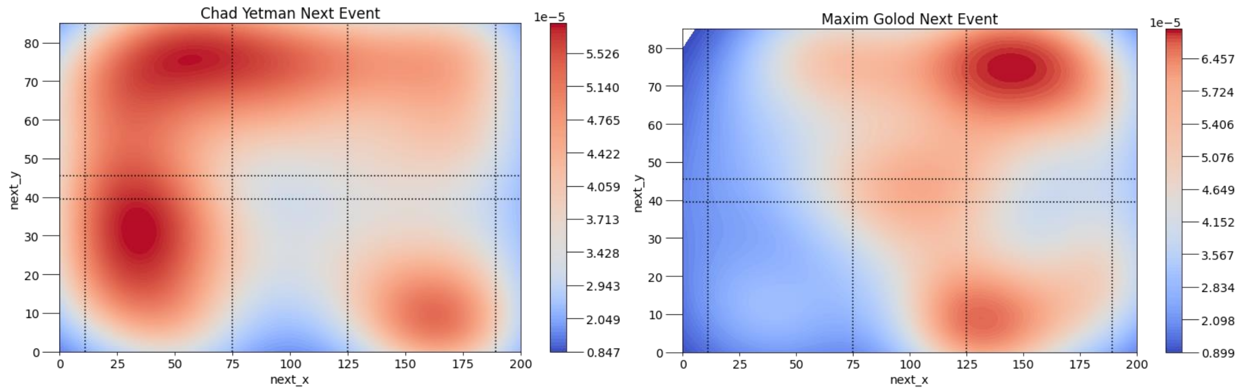


Fig. 5: Heatmaps of Events After Takeaways – The X-Coordinate from 0 -> 200 represents Defensive -> Offensive Zones. The lines from Y = 45.5 and 39.5 and X = 11 and 189 represent the goal line.

The features of the model are:

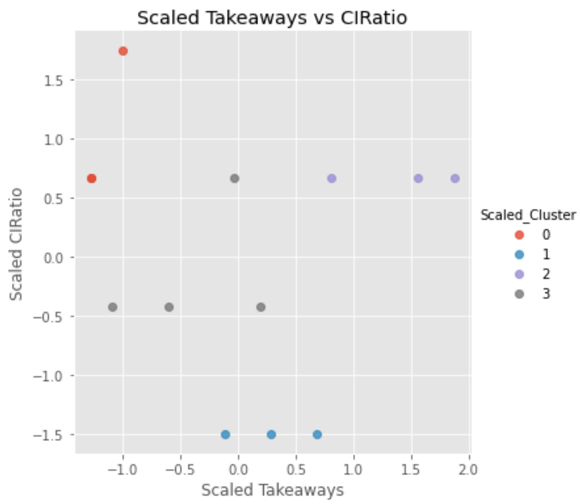
- Position Index,
- Mean coordinate based on event position (x,y – Denoted as Avg_nx and Avg_ny),
- Number of incompletions and completions,

The output of the model is the CI Ratio, defined as the number of complete passes divided by the number of incomplete passes. The results are given in Fig. 6.

Player	Position	Player Number	Position Index	Avgnx	Avgny	Incompletions	Completions	CIRatio	Predicted CIRatio	CIRatio Difference
Chad Yetman	Center	29	1	95	45	24	76	3.2	3.0	0.2
Maxim Golod	Left Winger	77	2	115	45	24	63	2.6	3.0	-0.4
Jack Duff	Defenseman	34	4	52	41	27	46	1.7	1.0	0.7
Brendan Hoffmann	Right Winger	91	3	93	48	17	44	2.6	3.0	-0.4
Austen Swankler	Center	21	1	127	39	27	44	1.6	1.0	0.6
Emmett Sproule	Left Winger	93	2	124	39	17	44	2.6	3.0	-0.4
Brendan Sellan	Left Winger	44	2	110	46	16	34	2.1	2.0	0.1
Daniel D'Amato	Right Winger	17	3	112	39	23	36	1.6	2.0	-0.4
Hayden Fowler	Right Winger	14	3	113	39	22	26	1.2	1.0	0.2
Kurtis Henry	Defenseman	41	4	72	35	15	35	2.3	2.0	0.3
Jamie Drysdale	Defenseman	4	4	70	53	10	41	4.1	4.0	0.1
Elias Cohen	Center	13	1	107	40	12	39	3.3	3.0	0.3
Jacob Golden	Defenseman	5	4	66	48	16	41	2.6	3.0	-0.4

Fig. 6: Poisson Regression Model Results: Predicted CI Ratio By Player

Clustering: Using the Predicted Takeaways and CI Ratio as features, the players are grouped together using K-Means Clustering. The predicted values were standardized using the scaling method through scikit-learn. Fig. 7 shows the graph of the clustered groups. The more positive the score for either prediction, the better.



Player	Position	Cluster
Jamie Drysdale	Defenseman	0
Elias Cohen	Center	0
Jacob Golden	Defenseman	0
Jack Duff	Defenseman	1
Austen Swankler	Center	1
Hayden Fowler	Right Wing	1
Chad Yetman	Center	2
Maxim Golod	Left Wing	2
Brendan Hoffmann	Right Wing	2
Emmett Sproule	Left Wing	3
Brendan Sellan	Left Wing	3
Daniel D’Amato	Right Wing	3
Kurtis Henry	Defenseman	3

Fig. 7: Clustering Results

Based off the grouping of the data, the players can be divided into four groups:

- Cluster 0: Low takeaways, High CI Ratio – Lower opportunities for takeaways, but have a high number of passing completions.
- Cluster 1: Average amount of takeaways, Low CI Ratio – Have the worst passing completions after a takeaway.
- Cluster 2: High takeaways, High CI Ratio – Efficient at getting takeaways and converting it into completed passes.
- Cluster 3: Average to low takeaways, Average to low CI Ratio – The players that do not excel in either stat is grouped here. Emmett Sproule has a high CI Ratio, but average amount of takeaways.

Conclusion: Through this analysis, players can be classified for taking away the puck from the opponent’s offense and generating it into a breakout into their offensive end, or to keep pressure on an opponent’s defensive half. Quantifying a player’s response in a transition state from defense to offense is important to know to be able to plan accordingly. Since takeaways lead to a significant impact of goals, through the data given, knowing how to plan around a player’s tendencies, such as the heatmaps shown, combined with the clustering data, can give teams an idea on how to counter their opponent’s tendencies. Utilizing count data in this fashion can be applied to different events, outside of takeaways, such as analyzing how a player immediately reacts after a shot or faceoff win occurs.

Acknowledgements: I would like to thank Brian Macdonald for guidance during the Office Hours sessions.

References:

[1] Subbaiah M., Chu D., Reyers M., Wu L. (2021). *Illuminating the Defense.* NFL Big Data Bowl 2021. Retrieved from https://www.kaggle.com/msubbaiah/illuminating-the-defense.

[2] Jill Reiner. (2021) *Evaluating and Clustering Coverage Skill.* NFL Big Data Bowl 2021. Retrieved from https://www.kaggle.com/reinerj1/evaluating-and-clustering-coverage-skill

[3] American Soccer Analysis. (2017, June 15). *Mapping Defensive Actions: A Spatial Analysis of Where Teams Focus Their Efforts*. Retrieved from https://www.americansocceranalysis.com/home/2017/6/14/mapping-defensive-actions-a-spatial-analysis-of-where-teams-focus-their-efforts

[4] Carpenter, C. (2021, Feb 25). *Where Goals Come From: Passing In The Final Third.* Retrieved from https://www.americansocceranalysis.com/home/2021/2/24/where-goals-come-from-passing-in-the-final-third

Github: Link