Entering the Zone: Quantifying Zone Entries

Zeke Weng, Daniel Du, Alexander Khan, Samarth Sharma
University of Toronto Sports Analytics Student Group

March 8, 2024

1 Introduction

The primary elements in hockey Expected Goals (xG) models typically encompass the distance from the goal and the angle of the shot. Teams leverage this data to recognize which areas on the ice yield a higher likelihood of scoring from a shot. Advanced xG models may also take into account additional factors such as time remaining, player handedness, shot type, among others. However, these xG models tend to overlook the nature of zone entry or the impact an on-puck defender can have in enhancing offensive xG by inducing turnovers or forcing missed shots. This aspect is frequently underestimated, despite the fact that effective defense significantly affects the team's offensive capabilities.

Objective: Our primary objective was to analyze the quality of zone entries, focusing on the offensive capabilities of players on the attacking team and the effectiveness of defensive players in controlling the puck. Using our model, we also hoped to discern whether the data aligns with or diverges from previously established research.

1.1 Relevant Work

Possibly the most prominent research to assign quality to zone entries came from Eric Tulsky, Geoff Detweiler, Robert Spencer and Corey Sznajder's 2013 paper presented at the MIT Sloan Sports Analytics Conference [1]. This paper detailed that over 3000 NHL games, zone-entry dumps versus carries lead to half as many shots created and a third as many goals. Thibaud Chatel's 2022 paper analyzed more than 50,000 zone entries of the SNL to find that dump-ins (excluding dump and change) lead to less than 40% of shots. [2]

2 Model Creation

To analyze Zone Entries, we will start by quantifying the amount of value that a Zone Entry adds. To do this, we will use the assistance of an xG model. For each Zone Entry, the shots that follow that Zone Entry have an associated xG value. Each Zone Entry's xG will be the summation of the xG of all post-zone Entry shots. Then we average out the xG across the Zone Entries by groupings to find any distinct patterns or points of interest. The groupings will be based on Region (Location) and Type of Zone Entry. The steps are as follows.

2.1 xG Model

Utilizing both Big Data Cup 2021 and Big Data Cup 2024 play-by-play data, we created a simple xG Model utilizing a Logistic Regression with the features angle and distance.

Intuitively, the smaller the distance between an offensive player and the net, the higher the chance that the shot will result in a goal. Not only does distance increasing the difficulty of a shot, it also proxies other critical factors that may affect the chance of a successful shot. With an increased distance, the puck will have more time on the rink, allowing more time for the defense to react. Distance also impacts the number of defenders between the net and the shooter. A distance of 5 feet would usually mean one shooter vs one defensive player; however, a distance of 35 feet could mean one shooter vs multiple defensive players, increasing the likelihood of a blocked shot.

The other integral feature is the angle of the shot. A larger angle gives the shooter more net room to aim at. A player directly in front of the goal has more options to find an optimal part of the net to shoot at. However, an offensive player on the far right or far left of the rink has a much narrower window, reducing the chance of a goal.

Below are the Equations used to calculate distance and angle

$$d = \sqrt{(x-189)^2 + (y-42.5)^2} \tag{1}$$

$$\theta = \arccos(\frac{d_1^2 + d_2^2 - 36)}{(2d_1 d_2)} \tag{2}$$

$$d_1 = \sqrt{(x-189)^2 + (y-45.5)^2} \tag{3}$$

$$d_2 = \sqrt{(x-189)^2 + (y-39.5)^2} \tag{4}$$

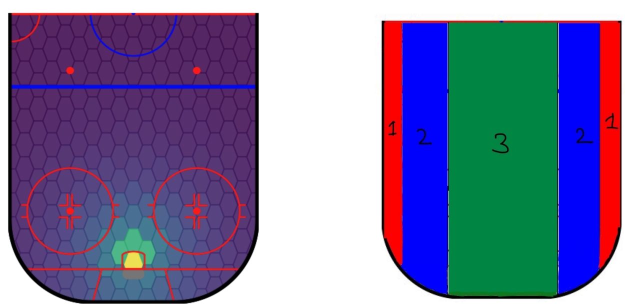
1. d is the distance between the player and the center of the net.
2. θ is the angle of the player to the net.
3. d_1, d_2 are the distances between the player and the left post and right post of the net.

With the angle and distance calculated, we combined the play by play data from BDC2021 and BDC2024 to create our xG Model, using a Logistic Regression Model. Below are the Coefficients for the Features.

	Coeffecient	(Std. Error)
Constant	-1.0276	0.2294
Angle	0.0105	0.0050
Distance	-0.0227	0.0068

Table 1: Coefficients of xG logistic regression model

As expected, when the angle increases, the xG increases which is indicated by the positive weight in the model. This is in line with our expectations that a bigger angle leads to a increased chance of a successful shot. Also, as expected, when the distance increases, the xG decreases which is indicated by the negative weight in the model. This is also in line with our expectations that a larger angle distance leads to a decreased chance of a successful shot. Below is a visualization for the xG Model on the Rink a), and also for Zone Regions b), discussed in 2.2.



(a) xG on Rink (b) Zone Regions

2.2 Categorizing and Quantifying Zone Entries

In theory, Zone Entries through the middle of the rink should yield better results as the attacker has more ice to work with, allowing for more strategy, mobility, and options in attacking the net. It keeps the defense guessing and they cannot lean towards a particular side. Zone Entries through the leftmost or rightmost channel narrow the rink as the boards limit the offensive player's decisions. As a result, in the long-run we expect Zone Entries through the middle to be more prolific than Zone Entries through the sides.

Another factor we saw important to categorize was the context of the Zone Entry or the 'Type.' Zone Entries generated on takeaways leading to breakaways would add great value to that possession. However, without tracking data, we are forced to proxy such events with 'Carried' Zone Entries. The types of Zone Entries that we consider are 'Carried,' 'Played,' and 'Dumped.'

We categorized each of the Zone Entries by Location as 'Region 1', 'Region 2', and 'Region 3'. Since the rink is symmetric, we evaluated the Regions symmetrically as well.

1. Region 1 corresponds to the outermost 7.5 feet of the rink $[0, 7.5) \cup (77.5, 85]$
2. Region 2 corresponds to the next 15 feet of the rink $[7.5, 22.5) \cup (62.5, 77.5]$

3. Region 3 corresponds to the outermost 7.5 feet of the rink $[22.5, 77.5]$

Below is the distribution of Zone Entries by Region and by Type across both the BDC2021 and BDC2024 datasets. See above for a visualization of the 3 regions.

	Carried	Played	Dumped	Total
Region 1	278	276	39	593
Region 2	222	102	44	368
Region 3	187	88	37	312
Total	687	466	120	1273

Table 2: Location x Type Distributions

With our Regions and Types defined, we proceeded to find the average xG associated with each label. The process of finding the average xG is as follows:

1. Sum the xG of all the attempted shots following a Zone Entry and assign that Zone Entry that summed xG.
2. Repeat this process for every single Zone Entry.
3. Group by label, whether Region or Type or any other grouping, and take the average of all the Zone Entry xGs.
4. Evaluate the results.

2.2.1 Zone Entries by Region and Zone Entries by Type

Sorting the Zone Entry Data into the respective Regions (1,2,3), we found the average xG associated by Region. Also, we sorted the Zone Entry Data into the respective Types ('Carried','Played','Dumped') and found the average xG associated by Type. The trends here match our expectations. The results are in Table 3 and Table 4.

Region 1	Region 2	Region 3
0.0163	0.0233	0.0248

Table 3: xG by Zone Entry region

Carried	Played	Dumped
0.0303	0.0210	0.0057

Table 4: xG by Zone Entry type

2.2.2 Zone Entries by Region minus Dumped

Since Dumped Zone Entries generate extremely low xG, for this part, we filtered out the 'Dumped' Zone Entries, leaving 'Carried' and 'Played'. Then, we sorted the Zone Entry Data into the respective Regions (1,2,3) and found the average xG associated by Region. The results are as follows:

Region 1	Region 2	Region 3
0.0250	0.0296	0.0336

Table 5: xG by Zone Entry type

Taking out the 'Dumped' Zone Entries leaves us with greater differences between the Regions, again aligning with our expectations for Region xGs.

2.2.3 Zone Entries by Region and Type

Lastly, we sorted by Region and Type, excluding 'Dumped' to stratify the differences further. We sorted the Zone Entry Data into the respective Regions and found the average xG associated by Region. The results are as follows

Across the board, the most compelling takeaway from these results in the differences in xG by Region specifically for the 'Carried' Zone Entries. The 'Played' Zone Entries do not perfectly meet this trend.

With the synthesized results of the groupings and resulting xGs, we will apply the averaged xG back into the BDC2024 Dataset to evaluate players. For Zone Entry of the Types 'Carried' and 'Played,' we will apply the specific averaged xG according to the 4th grouping where the average xG depends on the region in which the action is taking place. For 'Dumped' Zone Entries, we will assign the flat xG amount observed in the 2nd grouping across all regions.

Carried 1	Carried 2	Carried 3	Played 1	Played 2	Played 3
0.0258	0.0301	0.0371	0.0188	0.0274	0.0157

Table 6: xG by Zone Entry Region and Type

2.3 Player xG and Defensive xG (dxG) Assignment

After assigning Zone Entry Values based on the averages from the last section, we will use those xG values to assess players. Through Zone Entry, we assign values to players for both one's ability to create a Zone Entry and a defensive unit's collective ability to Stop Zone Entries. On the offensive end, the player that create the Zone Entry is given full credit of the Zone Entry's xG. On the defensive end, all the defenders are assigned the negative value of the Zone Entry's xG split across all the defenders as 'dxG.'

3 Proof of Concept

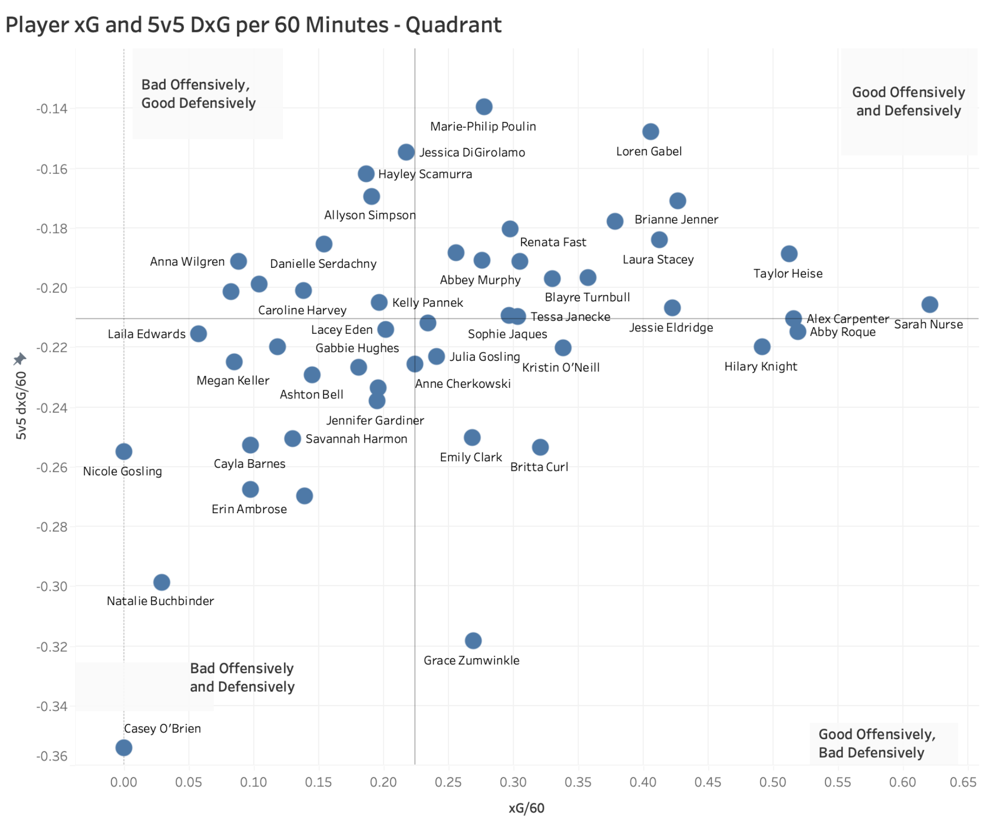
Our group went on to study these factors for the four hockey games between Team Canada and Team USA in the dataset. This included player shift data, so we could record all of the skaters on the ice. This became extremely useful as a replacement for player tracking.

The xG model included two factors, distance and angle, and also represented the defensive distance and angles as negative xG for the attacking team. The negative xG of a defensive player was also counted as positive dxG for such a defender. For our analysis metrics, we only counted a player's defensive xG in 5-on-5 and power play situations, since it would be unfair to punish a player's inability to stop offensive zone entries while shorthanded.

4 Results & Discussion

4.1 Individual Player Comparison: Offensive and Defensive xG Contribution

The graph below compares the attacking and defensive xGs from all players' zone entries in the first 4 games of the 2023-24 Rivalry Series between the Canadian and American Women's National Teams. Skaters are divided into quadrants, with the top right quadrant featuring players having an xG/60 and 5v5 defensive xG/60 above the median. Some of the world's top players such as Marie-Philip Poulin, Brianne Jenner, and Taylor Heise land in this quadrant [3].



4.2 Relationship between Regions and Attacking xG

By analyzing the attacking xG of players compared with the number of zone entries they make in a certain region, we can see the quality of chances created through zone entries in that region. From the graphs below, we can see that there is a low correlation between a player's attacking xG and their Region 1 zone entries, and a stronger correlation for Regions 2 and 3.

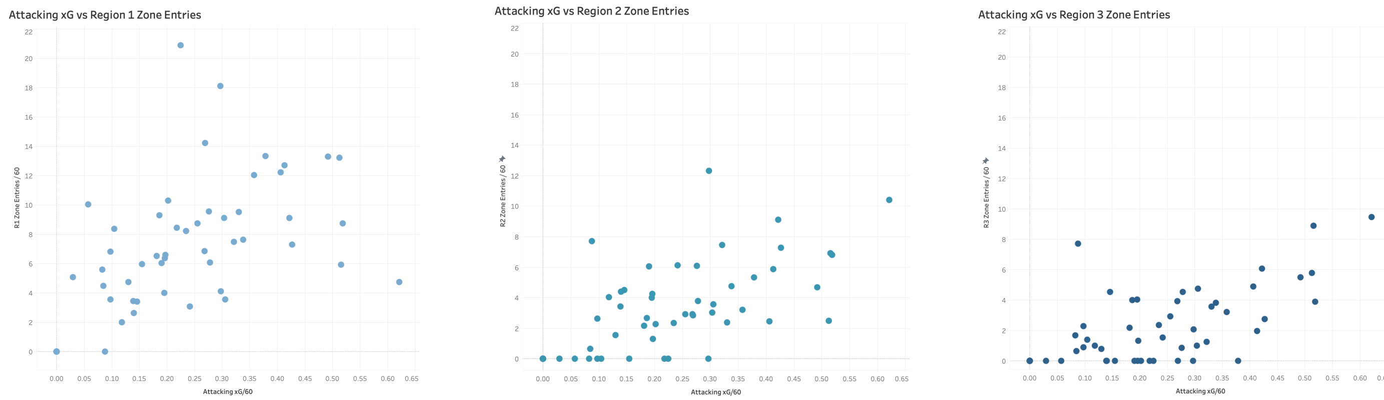


Figure 2: Attacking xG vs Region 1,2 and 3

4.3 Player Archetypes

To analyze a player beyond their surface-level stats, we can use our 3-region model to determine where their zone entries most frequently occur. Ranked below are the players whose zone entries created the most attacking xG, and which regions their zone entries occurred in:

Players' Most Frequent Region of Attack

Name	Attacking xG/60	% of xG from R1	% of xG from R2	% of xG from R3	R1 Entries/60	R2 Entries/60	R3 Entries/60
Sarah Nurse	0.621	13.6	46.0	40.5	4.7	10.4	9.5
Abby Roque	0.519	39.7	38.4	21.9	8.7	6.8	3.9
Alex Carpenter	0.516	14.3	25.8	59.9	5.9	6.9	8.9
Taylor Heise	0.512	50.4	14.6	35.0	13.2	2.5	5.8
Hilary Knight	0.491	51.6	27.1	21.4	13.3	4.7	5.5
Brianne Jenner	0.427	31.2	45.0	23.8	7.3	7.3	2.7
Jessie Eldridge	0.422	41.3	27.9	30.8	9.1	9.1	6.1
Laura Stacey	0.413	46.1	36.3	17.6	12.7	5.9	2.0
Loren Gabel	0.406	37.2	18.1	44.7	12.2	2.4	4.9
Jamie Lee Rattray	0.378	57.6	42.4	0.0	13.3	5.3	0.0
Blayre Turnbull	0.358	53.8	27.0	19.2	12.0	3.2	3.2
Kristin O'Neill	0.338	37.4	35.6	27.0	7.6	4.8	3.8

Players with the most zone entries per 60 minutes from each of the three regions are shown below, along with their attacking xG and the percentage of their attacking xG generated from that region. These tables display the regions in which players enter the zone and the ability of the player to create xG.

Name	Attacking xG/60	% of xG from R1	R1 Entries/60
Anne Cherkowski	0.224	100.0	20.9
Sophie Jaques	0.296	100.0	18.1
Grace Zumwinkle	0.269	94.0	14.2
Jamie Lee Rattray	0.378	57.6	13.3
Hilary Knight	0.491	51.6	13.3
Taylor Heise	0.512	50.4	13.2
Laura Stacey	0.413	46.1	12.7
Loren Gabel	0.406	37.2	12.2
Blayre Turnbull	0.358	53.8	12.0
Lacey Eden	0.202	93.5	10.3

Name	Attacking xG/60	% of xG from R2	R2 Entries/60
Sarah Fillier	0.297	74.25	12.34
Sarah Nurse	0.621	45.96	10.42
Jessie Eldridge	0.422	27.94	9.11
Anna Wilgren	0.088	50.00	7.71
Britta Curl	0.321	59.49	7.47
Brianne Jenner	0.427	45.04	7.29
Alex Carpenter	0.516	25.84	6.92
Abby Roque	0.519	38.40	6.80
Julia Gosling	0.241	76.38	6.12
Abbey Murphy	0.276	34.84	6.09

Name	Attacking xG/60	% of xG from R3	R3 Entries/60
Sarah Nurse	0.621	40.48	9.47
Alex Carpenter	0.516	59.89	8.90
Anna Wilgren	0.088	50.00	7.71
Jessie Eldridge	0.422	30.79	6.08
Taylor Heise	0.512	35.00	5.79
Hilary Knight	0.491	21.37	5.48
Loren Gabel	0.406	44.69	4.89
Renata Fast	0.305	51.52	4.74
Marie-Philip Poulin	0.278	31.92	4.55
Ashton Bell	0.145	49.97	4.52

By looking further into zone entry data, we can gather information about a player's playing style and archetype. Comparing players with a similar xG but different zone entry region frequencies can reveal insights, even with a small data sample. Consider the players below:

Surface-level attributes, such as a player's position, are visible from the data. Canadians Sarah Fillier and Sophie Jaques have an almost identical attacking xG, but the centre Fillier generates her xG from R2 (outer-middle region) entries, while the defender Jaques's zone entries come entirely along the boards, which is intuitive considering their positions.

Players' Most Frequent Region of Attack

Name	Attacking xG/60	% of xG from R1	% of xG from R2	% of xG from R3
Alex Carpenter	0.516	14.3	25.8	59.9
Taylor Heise	0.512	50.4	14.6	35.0
Sarah Fillier	0.297	21.8	74.3	3.9
Sophie Jaques	0.296	100.0	0.0	0.0
Marie-Philip Poulin	0.278	34.4	33.7	31.9
Grace Zumwinkle	0.269	94.0	6.0	0.0

Marie-Philip Poulin and Grace Zumwinkle have a similar xG, but their regions of attack reveal much about their playing archetype. Renowned for her smarts and controlling play on the ice, the centre Poulin's chances are created equally through all three regions. Zumwinkle, a right winger, creates her xG almost entirely from a high number of Region 1 zone entries.

Americans Alex Carpenter and Taylor Heise both have an impressive attacking xG, but their contributions come in different ways. Carpenter has the highest % of xG from R3 among players analyzed, and her 8.9 R3 Entries/60 rank her best among the Americans, while over half of Heise's xG is from R1 entries. In the 4 game sample, Carpenter had 3 goals and 1 assist and Heise had 1 goal and 2 assists. we can hypothesize that Carpenter gets into good shooting positions following R3 entries, while Heise's xG involvements from R1 entries are more pass-related [4].

5 Limitations

A significant constraint our project faced was the lack of tracking data of players on the ice. Given tracking data, we would have been able to factor in, say, if players on the defending team off the puck had stood in the way of the goal blocking a shot attempt. This is the idea of such pre-shot models researched prior. Factoring in off-puck movement could more accurately explain xG summaries from Zone Entries onto each team. [5]

Importantly, this dataset only consisted of four games with 1,237 zone entries including dump-ins. If a larger dataset were available with more games and zone entries, it is possible that our model would have higher variation of outcomes.

The performance of any team in any sport also depends on factors of the game setting itself. This includes factoring in the amount of rest of each team, whether or not teams have injured players that would have otherwise made a large impact. Other, non-quantifiable factors in the game setting include the stakes of the game, player discipline, and differences of coaching between two teams.

6 Conclusion

We implemented a simple expected goals (xG) model based on distance and angle to analyze the quality of zone entries. We categorized zone entries by their type and specific region, highlighting individual and team offensive and defensive patterns. This approach identified statistical tendencies, and also helped us analyze a player's value and contributions, as well as a classification of the player's archetype. Moreover, our results align with prior research, confirming the consistency in the quality of zone entries. The model advanced our knowledge on optimizing team performance through strategic zone entries and enhancing defensive strategies against them.

References

[1] Using Zone Entry Data To Separate Offensive, Neutral, And Defensive Zone Performance

[2] https://thibaudchatel.substack.com/p/assessing-the-risk-management-of?s=r

[3] https://www.tsn.ca/nhl/poulin-fillier-top-tsn-s-top-25-players-in-women-s-hockey-1.1858065

[4] https://web.archive.org/web/20240130025358/https://www.hockeycanada.ca/en-ca/team-canada/women/national/2023-24/rivalry-series/stats/player-stats

[5] https://hockey-graphs.com/2019/08/15/expected-goals-model-with-pre-shot- movement-part-4-variable-importance/

Code, Dataset, Figures, Equations