Data Mining the Play-By-Play

Assessing and Applying NHL Performance Metrics

Using Statistical Methods

Alexandr Diaz-Papkovich

A thesis submitted to the Faculty of Graduate and

Postdoctoral Affairs in partial fulfillment of the

requirements for the degree of

Master of Science

in

Probability and Statistics

Carleton University

Ottawa, Ontario

©2015

Alexandr Diaz-Papkovich

Acknowledgements

I wish to thank my supervisor Shirley Mills for her guidance and encouragement in this topic, my classmate David Wilson for suggesting I pursue a thesis, and the online hockey analytics community, whose years of efforts formed the basis for my work.

I'd also like to thank my friends: Jake, Warren, Jon, Tyler, Phil, Brian, Dave L., Dave A., and Kayla for their equally indispensable programming advice and patience with my rants in IRC; Dylan, Dave H., Leo, Ashley, Ann Marie, Katie, and Trish for keeping me informed of the world outside my office.

My friends and colleagues at Statistics Canada were invaluable, particularly Mark for tolerating my endless theory questions.

I would like to thank my fiancée Anna for everything she has done. When Ottawa was colder than Mars, you ventured outside so I wouldn't have to.

This thesis is dedicated to Andres, Marina, and Andrew.

Abstract

Hockey statistics are traditionally simple with limited analytical value, especially for individual assessment. The National Hockey League provides Play-By-Play and Time-On-Ice files for every game, recording game events and players present during them. These are combined to derive Corsi, Fenwick, and shot differentials, now commonly used in the hockey analytics world. Using regression and data visualization we examined their value in explaining win percentage and goal percentage, demonstrating the value of full-strength tied Corsi particularly. We also examined the effects of score on shot attempts, and the effects of special teams and home advantage on win percentage. Finally, we created a binary transaction database of players and Corsi events and applied association rule learning to clearly measure individual performance and player chemistry, demonstrating significant potential for use in management and coaching.

Table of Contents

	Page number
Acknowledgements	ii
Abstract	iii
List of Tables	vii
List of Illustrations	ix
1 Background	1
1.1 Introduction	1
1.2 Introduction to hockey	2
1.3 Literature review	2
2 Methodology	8
2.1 Data collection	8
2.1.1 Data quality	11
2.2 Data organization	12
2.3 Statistical techniques	13
2.4 Software used	14

3	Application	15
3.1	Shot metrics	15
3.2	Goal percentage	17
3.2.1	Goal distribution	19
3.3	Score effect	22
3.4	Shot metrics, scoring, and winning	24
3.4.1	Uncontrolled	24
3.4.2	Controlling for situation and strength	29
3.4.3	Controlling for strength, situation, and venue	34
3.4.4	Conclusions on shot metrics	39
3.5	Special teams	40
3.6	Shot metrics and special teams	42
3.7	Save percentage	46
3.8	Data visualization	48
3.9	Player analysis via association rule learning	58
3.9.1	Description of association rule learning	58
3.9.2	Applying association rules to ice hockey	60
3.9.3	Example: 2013-14 Toronto Maple Leafs	61
3.10	Future applications of association rules	76
4	Conclusion	77
4.1	Team metrics	77
4.2	Association rules	78
	Appendix A	80
	References	86
	Bibliography	90

Software	93
Glossary	95

List of Tables

3.1	Regression results: Regulation win percentage on regulation goal percentage (2012-13 excluded)	17
3.2	Mean Corsi percentage by score difference, away teams	22
3.3	Mean Corsi percentage by score difference, home teams	22
3.4	Regression results: Regulation win percentage and regulation goal percentage on shot metrics	26
3.5	Regression results: Regulation win percentage on regulation shot metrics	27
3.6	Regression results: Regulation goal percentage on regulation shot metrics	28
3.7	Regression results: Regulation win percentage and regulation goal percentage on shot metrics, full-strength tied	29
3.8	Regression results: Regulation win percentage on full-strength tied regulation shot metrics	30
3.9	Regression results: Regulation goal percentage on full-strength tied regulation shot metrics	31
3.10	Regression results: Regulation win percentage and regulation goal percentage on shot metrics and home indicator, with interaction	35

3.11 Regression results: Regulation win percentage and regulation goal percentage on shot metrics and home indicator, no inter-action	36
3.12 Regression results: Regulation win percentage on venue and full-strength tied shot metrics	37
3.13 Regression results: Regulation goal percentage on venue and regulation full-strength tied shot metrics	38
3.14 Regression results: Regulation win percentage on special teams .	41
3.15 Regression results: Regulation win percentage on full-strength tied Corsi percentage and special teams	43
3.16 Regression results: Regulation win percentage on venue, full-strength tied Corsi percentage, and special teams	44
3.17 Regression results: Regulation save percentage on regulation Corsi percentage	46
3.18 A sample of transaction database entries	61
3.19 Rules for 2013-14 Toronto Maple Leafs	62
3.20 Forward pairs with lift	72
3.21 Kadri and Bozak lines, with rates/60 minutes	72
3.22 Defensemen, sorted by TOI (seconds)	74
3.23 Defense pairs, sorted by TOI (seconds)	74

List of Illustrations

3.1	Regression: Regulation win percentage vs Regulation goal percentage	18
3.2	Histogram of regulation goal percentages (2012-13 excluded) . . .	20
3.3	Q-Q Plot of regulation goal percentage (2012-13 excluded) vs Normal distribution	20
3.4	Overlaid histograms of regulation goal percentages, home and away teams (2012-13 excluded)	21
3.5	Corsi percentages by score differential and period (2012-13 excluded)	23
3.6	Regression: Regulation win percentage vs Full-strength tied Corsi percentage (2012-13 excluded)	32
3.7	Regression: Regulation goal percentage vs Full-strength tied Corsi percentage (2012-13 excluded)	33
3.8	Scatterplot of regulation Corsi percentage and regulation save percentage	47
3.9	Corsi against per minute vs Corsi percentage, Eastern Conference, away games	50
3.10	Corsi against per minute vs Corsi percentage, Western Conference, away games	51

3.11 Corsi against per minute vs Corsi percentage, Eastern Conference, home games	52
3.12 Corsi against per minute vs Corsi percentage, Western Conference, home games	53
3.13 Corsi for per minute vs Corsi percentage, Eastern Conference, away games	54
3.14 Corsi for per minute vs Corsi percentage, Western Conference, away games	55
3.15 Corsi for per minute vs Corsi percentage, Eastern Conference, home games	56
3.16 Corsi for per minute vs Corsi percentage, Western Conference, home games	57
3.17 Association rules for individuals, 2013-14 Toronto Maple Leafs .	65
3.18 Association rules for defensemen, 2013-14 Toronto Maple Leafs .	66
3.19 Association rules for defense pairs, 2013-14 Toronto Maple Leafs	67
3.20 Association rules for forwards, 2013-14 Toronto Maple Leafs . . .	68
3.21 Association rules for forward pairs, 2013-14 Toronto Maple Leafs	69
3.22 Association rules for forward trios, 2013-14 Toronto Maple Leafs	70

Chapter 1

Background

1.1 Introduction

This thesis is broadly split into two portions: one exploring team metrics, and the other player metrics. Team metrics concern macro-level statistics such as win percentage, goal percentage, shot statistics, special teams, etc. We will establish the relationships between them and demonstrate the usefulness of full-strength tied Corsi¹ in particular. We will also examine visualizations of statistics and identify notable patterns therein.

Player metrics are statistics that apply to an individual player or to groups of individual players. We will apply association rule learning to a database of player-event “transactions” as a method of examining player abilities within the context of those that play alongside them. We will use the 2013-14 Toronto Maple Leafs’ full-strength tied Corsi events as an example and will analyze the contributions of their forwards and defensemen.

¹Corsi and Fenwick events are different types of shot attempts. They are explained in greater detail in Chapter 3.

1.2 Introduction to hockey

Ice hockey is a sport played between two teams on an ice surface. Each team fields six players – typically five skaters and one goaltender – with the objective of outscoring the opposition before the end of the game. There are many professional and amateur leagues throughout the world; this thesis will concentrate on the National Hockey League (NHL). Games during the NHL regular season last 60 minutes and are usually divided into three equal periods, referred to as “regulation play”. In the event of a tie, there are five extra minutes of sudden death overtime play, followed by a shootout if necessary. Play begins with a faceoff between two players. Between periods, play is continuous except for stoppages for rule infractions, goals, injuries, or when the officials lose sight of the puck.

During a full season, each team plays 82 games divided equally between home and away. The 2012-13 season was shortened to 48 games due to a labour dispute. Many official statistics are kept, including goals, assists, shots, saves, and those on the efficacy of special teams. Goals are the most important statistic and as a result they are often the target of analysis. However, it is difficult to decouple an individual’s performance from that of their team; scorekeepers can track a player’s presence during goals through plus-minus statistics, but the infrequency of goals can complicate analysis.

1.3 Literature review

The use of analytics, also called advanced statistics, is now commonplace in professional sports. Hockey has quietly formed its own analytics culture, which began in humble areas such as the comment sections of blogs.[1] Significant contributions have come from statisticians, amateur and professional alike, who

post their findings online, and several have been recently hired by NHL franchises.[2]

Much of the available research is based on data culled from the NHL website, including the Play-By-Play (PBP) and Time-On-Ice (TOI) files, and less-detailed data such as box scores and game summaries. More rarely, research has used foreign or junior leagues' data. Unless otherwise stated, assume that any work referenced comes from NHL data.

Owing to its fluid and often chaotic nature, hockey statistics can be difficult to measure and meaningfully use. Skaters regularly reach 30km/h and fire shots exceeding 160km/h. They are constantly moving on the ice and play in shifts lasting from a few seconds to several minutes. If there are no stoppages, players change shifts as they skate by the bench, often called changing "on the fly". A simple bounce of the puck can change the direction of the play entirely, and it is not unusual to see long periods of end-to-end action. Adding a further challenge, skaters typically play alongside the same people for much of the season. Forwards are grouped into lines of three – a centre and two wingers – and defense players are paired. Consequently, statistical measurements of a player are confounded by teammates. Additionally, coaches will match their lines against the opposition based on skill and perceived advantages, so results must be interpreted in the context of the players that face each other. Going forward, it is imperative to keep these caveats in mind.

Traditional hockey statistics focus on discrete metrics such as goals, points, faceoff wins, hits, etc. Ultimately, teams want to win games, which requires that they outscore their opponents. However, goals are rare, occurring on about 10% of shots, which limits their analytic value. One commonly-used metric is plus-minus, which attempts to account for a player's presence during a goal: those on the ice for an even-strength or shorthanded goal-for are awarded +1,

and those for a goal-against are awarded -1. This can serve as a proxy for player contribution to goals, but it neglects to account for competition strength and team effects. Macdonald[4] proposed a least-squares adjusted plus-minus (APM) model for offensive and defensive contributions that accounted for the zone in which a play began (“zone starts”), a team’s playing strength, and the duration of a player’s shift. He refined the model[5][3] with ridge regression to account for collinearity between players, and found incorporating faceoff wins and possession statistics improved its accuracy.

Gramacy, Jenson, and Taddy[6] approached player contribution with a regularized logistic regression model that measured the goals scored given a player’s presence on the ice and concluded that most players’ impacts are indistinguishable from team effects. Schuckers, Lock, Wells, Knickerbocker, and Lock[7] studied whether a goal was scored within 10 seconds of a play given that a particular player was on the ice, and used this to create a rating for NHL skaters. Using a least-squares model, they measured the difference between goals and expected goals and ranked by sign and magnitude. At the time, the NHL’s play-by-play data did not include zone information or shot details. With Curro, Schuckers[8] expanded on his work with the Total Hockey Rating (THoR). This model used ridge regression and included zone starts and home-ice effects. Unlike the previous model, this one measured the probability of a goal within 20 seconds of a play, with a final metric of “wins created” relative to the average NHL player.

Spagnola’s thesis[9] examined the 2011-12 data for the Columbus Blue Jackets in an attempt to define the Complete Plus-Minus (CPM). He calculated the log-odds ratios of goals-for and goals-against for each player and defined the CPM as the difference in probabilities between the respective events. Higher values indicated a player with better performance. He presented the model

as a measure of how players perform within a team, and as a tool for roster management.

By finding an appropriate measurement of player value, stakeholders could build around team strengths and weaknesses and help determine player compensation.² Vincent and Eastman[10] attempted to categorize NHL players through k-means clustering on player weight and per-game points, penalty minutes, and plus-minus. After mathematically justifying the number of categories of players, forwards were split into “grinders”, “scorers”, and “enforcers”, while defensemen were categorized as “scorers” and “aggressors”. The study’s data predate the 2004-05 lockout, which the authors acknowledge as a dramatic turning point in how the game was played. Additionally, McIndoe[11] has shown that head injury concerns coupled with culture shifts in fighting and strategy have led to the slow disappearance of enforcers as a distinct role.

Researchers model team performance as well as player performance. Buttrey, Washburn, and Price[12] found that each team’s scoring is well-modelled as a Poisson process. They define the parameter of the distribution for each match as the product of a team’s base scoring rate (λ_0), an offensive factor(A_i), a defensive factor (B_j), and a home/away variable (D), so $X \sim POI(\lambda_{ij} = \lambda_0 A_i B_j D)$. They split games into strength levels and calculated the number of goals scored per 60 minutes. They tested their model after about 62 games had been played by each team and found a strong correlation between predicted and actual points awarded. Using the most recent n days proved more accurate than naively using an entire season, with $n = 20$ showing the highest correlation, $\rho = .688$.

Marek, Šediv, and Toupal[13] used 12 seasons of full-time results from the Czech Republic’s *Extraliga* to test several Poisson-based approaches. They denoted each match result by a pair of random variables (X_{ij}, Y_{ij}) represent-

²Since the 2005-06 NHL season the league has imposed a hard salary cap on both individual and team salaries. This is a considerable constraint on a team’s potential skill level.

ing goals scored by the home and away teams, respectively. From there, they researched scores as a bivariate Poisson result, as two independent Poisson variables, and as diagonal inflated versions of each. As with Buttrey et al, the parameters were the product of a base rate, a home effects variable, and offensive/defensive factors. They identified the diagonal inflated double Poisson model as the best and showed it was possible to make a profit against bookmakers in some situations.

Gill[14] assumed jointly independent Poisson models for games in the third period and found his model had excellent predictive value.

Scoring and game situations have also been represented as semi-Markov processes. In 2006, Thomas[23] used data from 18 National Collegiate Athletic Association (NCAA) games to split the game into 19 Markov states based on zone, puck control, and event (possession, faceoff, goal, takeaway, retreat). The holding times for the states were not exponentially distributed, suggesting a semi-Markov process over a typical continuous Markov process. After calculating the probability of a goal within 40 seconds of a state, he analyzed various team tactics given the location and possession status of the puck. He additionally noted the importance of defensive faceoffs for preventing goals. In later work[15], he explored similar semi-Markov processes to describe the non-exponential interarrival times between goals and estimated the value of a goal through its impact on the probability of winning.

Tulsky, Detweiler, Spencer, and Sznajder[16] examined puck-moving strategies in their work on zone entries. They tracked zone entries manually from 300 games in the 2011-12 NHL season and found entries with possession were more than twice as effective as dump-ins at generating shots, scoring chances, and goals. One notable finding was the lack of distinction between players once they had entered the offensive zone. They suggested that the talent component

becomes relevant in neutral zone play. More broadly, they also suggested teams should play more aggressively.

More universal results such as home advantage and the importance of faceoffs have been independently corroborated. The works of Pollard[17] in 2005 and Doyle and Leard[22] in 2012 confirms that home teams have consistently won more games than away teams. Doyle and Leard additionally noted that there was no significant difference between teams, but there was variation by season. They also found that faceoffs offered the home team an advantage (a result also found by Liardi and Carron)[18], but having the last line change made no difference. Away teams received on average 45 seconds of extra penalty time per game, suggesting referee bias may exist.

Tore Purdy[19] of the website Objective NHL studied the predictive values of shot differentials, Corsi, and Fenwick in 2011. During even-strength tied situations, he found that during the season Corsi was the best predictor for scoring, followed by Fenwick and shot differential. Over many seasons, Fenwick and shot differential predictions were very close, and better than Corsi. The author noted that it is possible that all three metrics are equally valid in the long run.

Chapter 2

Methodology

2.1 Data collection

The primary source of data is the NHL website, NHL.com. NHL games are indexed by season, season type, and a sequential integer starting at 1 for each season type. Seasons are referenced either in full or by the year in which they started. The three season types are preseason (code 01), regular season (code 02), and postseason/playoffs (code 03). A typical regular season has 1230 games with 82 for each team, evenly split between home and away. The 2012-13 season was shortened by a labour dispute and only featured 720 games. This analysis will use data from the regular season from the five most recent seasons, 2009-10 through 2013-14. It will also be limited to data from regulation play.

NHL games are tracked using the Real-Time Scoring System (RTSS). The RTSS is split into several pages containing different aspects of the game. We use the Play-By-Play (PBP) and player Time-On-Ice (TOI) pages for each team. Each page for each game was downloaded using a Python script, and the HTML was parsed using the Python package BeautifulSoup4. The PBP files are a

particularly rich data cache, offering a list of timestamped events for each game. The events tracked include:

- Beginning and end of official play
- Faceoffs
- Goals, shots on goal, blocked shots, missed shots, and penalty shots
- Penalties
- Hits
- Giveaways and takeaways
- Stoppages in play

Other events, such as penalty expirations and players entering or leaving play, are not recorded by the NHL. Adding these events improves the accuracy of time-based estimates, as it allows users to calculate the time the teams and players spend in specific situations, such as differences in score or strength. Penalty expirations were added through a program that calculated their times; there are rare ambiguous cases where the official rules require a captain to choose his team’s playing strength. These cases have minimal impact over the course of a season. Player shift changes are added by cross-referencing the TOI files with the time in-game. For each event, a player’s presence is indicated with a binary flag using his name as a column header.

The motivation for these dataset additions was twofold: first, to estimate how often teams were playing at even-strength; and second, to estimate the times that players spent together through different situations. We achieve this by measuring time between events, supplemented with penalty expirations and shift changes as implicit events. This breaks the game into smaller discrete slices of time allowing us to infer more than before. We can also control for the score

difference and the strength of either team. While being shorthanded or on the powerplay has obvious consequences on tactics, the effect of score difference is more subtle but has been researched.[24]

Though calculating penalty times is trivial on its surface, it requires careful work to ensure accuracy. Major penalties are easily identified and handled, since they always expire after five minutes. Minor penalties are simple in theory, but are fraught with special circumstances: they are often negated, stacked, or combined into a double-minor or cancelled out with coincidental penalties. Each of these cases must be identified and handled when calculating an individual penalty’s expiration time.

Adding player shift times is easier, but again requires special consideration. Naively matching times to events can result in players being assigned to the ice for events that occurred in a different play. For example, it is possible to have a shot, a save, and a faceoff all at the same time t . A player with a shift start time t would incorrectly be placed on the ice during the shot, while a player with a shift end time at t would be erroneously present during the faceoff.

Auxiliary datasets included team records, playoff finishes, and data such as team conferences and player-name conversions. Two sets of team records were created: the first was taken from the NHL website, while the second was calculated from PBP data to account for a few missing games. Unless otherwise stated, the team records used for analysis came from PBP data. In total, there were 5638 games captured in the PBP, out of a potential 5640. Player name conversions account for players changing the spellings of their names during the regular season (e.g. “Ilja Bryzgalov” to “Ilya Bryzgalov”, “Alex Steen” to “Alexander Steen”, “Michaël Bournival” to “Michael Bournival”, etc.).

2.1.1 Data quality

While every effort was made to ensure that the data were of the highest possible quality, there are examples of errors in the RTSS:

- 2013-14, game 0176: The opening faceoff was recorded at 0:15 of the first period, outside the neutral zone.
- 2012-13, game 0031, event 303: A faceoff was recorded between coincidental penalties, categorizing it as shorthanded rather than even-strength.
- 2012-13, game 0404, event 156: A shot was erroneously recorded with six skaters and a goalie for each team.

Additionally, the NHL's PBP files are not coded consistently, with some events falling outside the expected HTML table structure. In some cases, events at the end of a period or game – typically concurrent penalties as the consequence of a scrum – were not captured and ended up excluded from analysis. With approximately 1.5 million rows of data in the basic format of the dataset, these types of errors are inevitable. Thus, comparisons with other data sources will yield minor inconsistencies, though the conclusions drawn from this thesis are sound.

Wherever possible, we have patched known errors by using surrounding events for context to correct for aspects such as score, team strength, and player presence, keeping major statistics consistent with the NHL's records. When the data we sought to compare were not available from the NHL website, we consulted the websites Puckalytics and WAR On Ice.¹

¹Puckalytics and WAR On Ice can be found at <http://www.puckalytics.com/> and <http://war-on-ice.com/>, respectively. The NHL website does not offer shot metric statistics at the level of detail (e.g. by strength and score) required to make comparisons.

2.2 Data organization

Data are organized into two dataset types. One is league-wide, containing all events detailed above, but without players on the ice identified. The second type is created for each team for each season, in which players' names are column headers and their presence is indicated with a binary variable. Keeping players on the league-wide dataset would result in an unwieldy – and very sparse – file, so separating them allows us to carry out individual and team analyses without struggling for computing power, while leaving the option to combine them for more detailed study.

One of the primary motivations for keeping data in a binary format was its suitability for quick queries and for association-rule learning. A typical framework for association-rule learning requires a binary transactional database with the headers identifying individual items. In this situation, players and events are used as items, with the events being specified as consequents. Using a machine learning approach allows us to skip the painstaking task of tabulating results for each individual player and then devising comparison methods.

All events are identified by the situation and strength. The situation refers to a particular team's score differential, while the strength refers to the number of players on the ice relative to the opposition. Unless otherwise noted, all events used are from situations where teams are tied and have all six players on the ice. Data have been standardized where possible. The preferred method is to normalize values to per-minute or per-60-minute rates. Statistics are also represented through percentages. Combining these two approaches permits the study of intensity of play as well as control. While control of play will be demonstrated as a key factor, a team's intensity can add necessary context.

2.3 Statistical techniques

The team metrics used have been standardized and controlled as much as possible. In an effort to reduce uncertainty, only regular season games are considered and overtime and shootouts have been excluded from analysis. Regular season play limits team lineups to normal rosters and features at most two consecutive games between the same teams. Overtime involves tied teams playing four-on-four until either a team scores or five minutes of play have elapsed. Additionally, once a team has reached overtime play, they have secured one point in the standings and the tactics of the rest of the game are conditional on the impact of that point. Shootouts may be useful for marginal analysis of goaltenders and scorers, but they are considered beyond the scope of this study.

Where possible, relationships between variables were tested using ordinary least squares linear regression. Owing to the nature of the game, many variables are tied to each other. Shot attempts, goals, and save percentage are intrinsically linked: a goal is automatically a shot attempt; shots are more likely to score than shot attempts; and goals are used in the calculation of save percentage. In an effort to extricate these variables, the preferred measure of success is regulation win percentage. This is defined as the percentage of games where a team has outscored its opponent at the end of regulation play. Effectively, it attempts to measure a particular metric's contribution to outscoring an opponent over 60 minutes of all types of play. Wins in overtime and the shootout are not considered.

To determine statistical significance of differences between means, we use the t-test, provided that the values being compared have the necessary distributions. Data visualization is also used frequently to highlight trends that may be less obvious when presented as charts and tables. Individual contributions were measured primarily using association-rule learning. A more detailed explanation

of the technique is given in Chapter 3.

In many cases, the shortened 2012-13 season was excluded from analysis because the season was not long enough for variables to regress to long-run values. The relationships between variables were still present when the season was included, though typically they were slightly looser. Unless otherwise stated, analysis includes empty-net events. Since the majority of analysis is carried out during situations where the game is tied and both teams are at full strength, the main reason for an empty net is a team pulling the goalie during a delayed penalty call.

2.4 Software used

All data scraping, parsing, and formatting was done in **Python 2.7**. Statistical analysis was carried out in **R 3.1.2**. Linear models were fitted using the standard **lm** procedure. Graphics were generated using **ggplot2**, unless otherwise stated. Association-rule learning was done via the library **arules**. Additional functions, such as aggregation, string parsing, and data manipulation were carried out with the **plyr**, **stringr**, and **reshape2** libraries, respectively. **L^AT_EX** tables were generated using the **stargazer** package.

Chapter 3

Application

Before investigating individuals, we analyzed teams. Team statistics proved invaluable for establishing relationships between variables such as goal-scoring, shot metrics, save percentage, and winning.

3.1 Shot metrics

Three shot metrics – alternately called possession metrics – are used: the Corsi (C), the Fenwick (F), and the simple shot on goal. Each of them is a measure of shot attempts. Let S represent shots on goal, M represent shots missed, and B represent shots blocked. The subscripts f and a indicate “for” and “against”, respectively. Actively blocking a shot is considered a Blocked Shot For; similarly, having a shot blocked is considered a Blocked Shot Against. Reversing the subscripts in the equations will give the accompanying “against” values.

$$C_f = S_f + M_f + B_a$$

$$F_f = S_f + M_f$$

These values can be expressed as percentages of total shot attempts: the Corsi, Fenwick, and shot percentages.

$$C\% = \frac{C_f}{C_f + C_a}, \quad F\% = \frac{F_f}{F_f + F_a}, \quad S\% = \frac{S_f}{S_f + S_a},$$

All individuals on the ice during a shot attempt are flagged; a player's shot metrics reflect his team while he is on the ice and not necessarily his own shot attempts. Shot metrics are used for a variety of purposes, ranging from proxies of time spent in an opponent's zone to measuring a player's contribution to offence. They may be combined with TOI data to generate per-60-minute statistics, such as the Corsi-for per 60 minutes (CF60) and Corsi-against per 60 minutes (CA60).

3.2 Goal percentage

Goal percentage is obviously one of the strongest predictors of success. A simple linear regression, pictured in Figure 3.1, confirms this. Let y_i be the regulation win percentage and x_i the goal percentage for team i during a full regular season. Our model is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

Diagnostic plots show that the residuals are independently and identically normally distributed with mean 0 and constant variance, justifying linear regression on a percentage value. (See Figures 1 – 4 in Appendix A)

Table 3.1: Regression results: Regulation win percentage on regulation goal percentage (2012-13 excluded)

<i>Dependent variable:</i>	
	Win%
Goal%	1.844***
(Std. Err.)	(0.078)
Constant	−0.545***
(Std. Err.)	(0.039)
Observations	120 (4 seasons x 30 teams)
R ²	0.825
Adjusted R ²	0.824
Residual Std. Error	0.035 (df = 118)
F Statistic	557.480*** (df = 1; 118)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Figure 3.1: Regression: Regulation win percentage vs Regulation goal percentage

3.2.1 Goal distribution

The percentage of goals scored is normally distributed. A Shapiro-Wilk normality test returns $W = 0.9962$ with p-value 0.9886, which does not reject the hypothesis that goal percentages are normal. A histogram and QQ plot (Figures 3.2 and 3.3) of the data provide confirmation.

Figure 3.2: Histogram of regulation goal percentages (2012-13 excluded)

Figure 3.3: Q-Q Plot of regulation goal percentage (2012-13 excluded) vs Normal distribution

Figure 3.4: Overlaid histograms of regulation goal percentages, home and away teams (2012-13 excluded)

Note: Colours are illustrative. The overlap is not deliberately highlighted.

Home ice advantage has been documented over the entire history of the sport. Home teams consistently win more than away teams.[17, 22] The last four full seasons of data corroborate this finding. Splitting the goal percentage data into home and away shows that the mean goal percentage for a visiting team is 47.31%, while for home teams it is 52.72%. Performing a two sample t-test returns a test statistic of $t = -9.144$ and a p-value of $2.2 * 10^{-16}$, rejecting the hypothesis that the difference between means is zero. The 95% confidence interval for the difference is $(-0.0658, -0.0425)$

3.3 Score effect

As the game’s score changes, teams become more defensive or more aggressive. In a tied situation, both teams are in a “neutral” state in which neither wishes to sacrifice a goal. However, once one team scores, the other must aggressively push for recovery. These results hold at home and away.

Table 3.2: Mean Corsi percentage by score difference, away teams

Season	Down 2+	Down 1	Tied	Up 1	Up 2+
20092010	0.560	0.531	0.481	0.440	0.397
20102011	0.565	0.533	0.488	0.435	0.404
20112012	0.565	0.535	0.482	0.431	0.401
20122013	0.566	0.536	0.485	0.449	0.394
20132014	0.548	0.535	0.483	0.437	0.411

Note: Column headers refer to the score difference for away teams. E.g. “Down 2+” refers to away teams losing by 2 or more goals.

Table 3.3: Mean Corsi percentage by score difference, home teams

Season	Down 2+	Down 1	Tied	Up 1	Up 2+
20092010	0.600	0.563	0.518	0.471	0.442
20102011	0.603	0.569	0.511	0.466	0.434
20112012	0.598	0.567	0.518	0.467	0.434
20122013	0.608	0.554	0.513	0.462	0.435
20132014	0.599	0.565	0.518	0.467	0.441

Note: Column headers refer to the score difference for home teams. E.g. “Down 2+” refers to home teams losing by 2 or more goals.

Similar effects take hold for other shot metrics. The score effect is particularly pronounced in the third period, where it intensifies as the period goes on. Combined with the obvious impact of team strength on strategy, we understand that even-strength tied metrics are likely the most neutral. Figure 3.5 presents the average team Corsi percentage over the last four full seasons, with the third period split into intervals.

Figure 3.5: Corsi percentages by score differential and period (2012-13 excluded)

3.4 Shot metrics, scoring, and winning

All three shot metrics have predictive value, though the contexts in which they are used matter. According to Purdy[19], within a season even-strength tied Corsi correlates most closely with goal ratio, while over a larger sample Fenwick and shot percentages are closer. We will examine the relationships using linear regression on regulation win percentage and goal percentage and see which metrics provide the best fits. All values used are from regulation play averaged over a full season, with 2012-13 excluded.

Regression diagnostics are provided as accompanying materials. Cited diagnostics are provided in appendices. Unless otherwise stated, residual diagnostics meet regression assumptions on normality, independence, and homoscedasticity. No significant correlation between seasons was found (see Figure 9 in Appendix A).

3.4.1 Uncontrolled

We begin with a model of all shooting events and regulation wins without controlling for situation, strength, or home/away. We will measure the fit of each of the three shot metrics as linear predictors for win percentage and goal percentage. However, first we will demonstrate multicollinearity between the shot metrics. We define a naive linear regression model: let y_i be the regulation win percentage for the i^{th} team, and x_{1i}, x_{2i}, x_{3i} be the team's shot, Fenwick, and Corsi percentages, respectively.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \quad \varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

The variance inflation factors (VIF) for the predictors are respectively 25.33, 55.75, and 18.59, well above the rule of thumb value of 10. Since all shot metrics

include at least shots on goal in their calculation, it is not surprising that they cause multicollinearity issues. From here, we will model the metrics separately to assess their explanatory value, resulting in three separate models.

We proceed with a multivariate regression model. For the i^{th} observation, let y_{i1} be the regulation win percentage, y_{i2} the regulation goal percentage, and x_{i1} the shot metric. We fit:

$$y_{i1} = \beta_{01} + \beta_{11}x_{i1} + \varepsilon_{i1} \quad \varepsilon_{i1} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_{11}^2)$$

$$y_{i2} = \beta_{02} + \beta_{12}x_{i1} + \varepsilon_{i2} \quad \varepsilon_{i2} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_{22}^2)$$

Letting X be the usual matrix of observed values, we write the multivariate model $Y = XB + \epsilon$ such that:

$$Y = \begin{bmatrix} y_{11} & y_{12} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{bmatrix} = \begin{bmatrix} Y_{(1)} & Y_{(2)} \end{bmatrix}, B = \begin{bmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \end{bmatrix} = \begin{bmatrix} \beta_{(1)} & \beta_{(2)} \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \vdots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} \end{bmatrix} = \begin{bmatrix} \epsilon_{(1)} & \epsilon_{(2)} \end{bmatrix}$$

Under this model, $Cov(\epsilon_{(j)}, \epsilon_{(k)}) = \sigma_{jk}\mathbf{I}$ and $E(\epsilon_{(j)}) = \mathbf{0}$. The shot metrics are evaluated in three separate models: (1) for Shot%, (2) for Fenwick%, and (3) for Corsi%.

Table 3.4: Regression results: Regulation win percentage and regulation goal percentage on shot metrics

Model	Metric	Test stat	F value	DF(num; den)	Pr(>F)
(1)	Shot%	0.417	41.847	2; 117	1.9531e-14
(2)	Fenwick%	0.410	40.606	2; 117	4.044e-14
(3)	Corsi%	0.397	38.574	2; 117	1.3585e-13
<i>Note:</i> (1), (2), and (3) are model IDs for each metric					

For each shot metric, we test the hypothesis that the shot metric has no effect, using Pillai's Trace as our test statistic. Looking at the probabilities in Table 3.4, we can confidently reject the hypothesis for each of the shot metrics. We go on to examine the univariate model fit for each metric to see which one has more explanatory value.

Table 3.5: Regression results: Regulation win percentage on regulation shot metrics

	<i>Dependent variable:</i>		
	Win%		
	(1)	(2)	(3)
Shot%	1.727***		
(Std. Err.)	(0.249)		
Fenwick%		1.623***	
(Std. Err.)		(0.239)	
Corsi%			1.581***
(Std. Err.)			(0.242)
Constant	-0.486***	-0.434***	-0.413***
(Std. Err.)	(0.125)	(0.119)	(0.121)
Observations	120	120	120
R ²	0.290	0.282	0.265
Adjusted R ²	0.284	0.276	0.259
Residual Std. Error (df = 118)	0.070	0.070	0.071
F Statistic (df = 1; 118)	48.118***	46.297***	42.586***

Note:

*p<0.1; **p<0.05; ***p<0.01

(1), (2), and (3) are model IDs for each metric

First we look at the model for win percentage. Looking at the adjusted R^2 values, shot percentage explains the most variation, followed by Fenwick and Corsi, in that order. Shots have the most well-behaved residuals, whereas Q-Q plots for the Fenwick and Corsi residuals suggest that they stray from the usual i.i.d. $\mathcal{N}(0, \sigma^2)$ assumption. (See Figures 5, 6, 7 in Appendix A)

Table 3.6: Regression results: Regulation goal percentage on regulation shot metrics

	<i>Dependent variable:</i>		
	Goal%		
	(1)	(2)	(3)
Shot%	1.009***		
(Std. Err.)	(0.112)		
Fenwick%		0.952***	
(Std. Err.)		(0.108)	
Corsi%			0.938***
(Std. Err.)			(0.109)
Constant	−0.004	0.024	0.031
(Std. Err.)	(0.056)	(0.054)	(0.055)
Observations	120	120	120
R ²	0.407	0.399	0.384
Adjusted R ²	0.402	0.394	0.379
Residual Std. Error (df = 118)	0.031	0.032	0.032
F Statistic (df = 1; 118)	81.044***	78.372***	73.600***

Note: *p<0.1; **p<0.05; ***p<0.01
(1), (2), and (3) are model IDs for each metric

Now we examine the models for goal percentage. Model (1), which uses shot percentage, has the highest adjusted R^2 , narrowly beating Fenwick percentage.

3.4.2 Controlling for situation and strength

Next we consider the same regression models but limit the data to five-on-five play while the game is tied. Our model is the same as in the previous subsection.

We begin by testing the hypothesis that the shot metric has no predictive value.

Table 3.7: Regression results: Regulation win percentage and regulation goal percentage on shot metrics, full-strength tied

Model	Metric	Test stat	F value	DF(num; den)	Pr(>F)
(1)	Shot%	0.384	36.432	2; 117	5.0113e-13
(2)	Fenwick%	0.367	33.977	2; 117	2.3202e-12
(3)	Corsi%	0.382	36.092	2; 117	6.1808e-13
<i>Note:</i> (1), (2), and (3) are model IDs for each metric					

We reject the null hypothesis, and examine the three univariate models for each of their fits with the response variables.

Table 3.8: Regression results: Regulation win percentage on full-strength tied regulation shot metrics

	<i>Dependent variable:</i>		
	Win%		
	(1)	(2)	(3)
Shot%	1.347***		
(Std. Err.)	(0.193)		
Fenwick%		1.304***	
(Std. Err.)		(0.192)	
Corsi%			1.342***
(Std. Err.)			(0.187)
Constant	-0.296***	-0.274***	-0.293***
(Std. Err.)	(0.097)	(0.096)	(0.094)
Observations	120	120	120
R ²	0.291	0.282	0.303
Adjusted R ²	0.285	0.276	0.297
Residual Std. Error (df = 118)	0.070	0.070	0.069
F Statistic (df = 1; 118)	48.533***	46.293***	51.372***

Note: *p<0.1; **p<0.05; ***p<0.01
(1), (2), and (3) are model IDs for each metric

Comparing adjusted R^2 values, Corsi percentage – Model (3) – is now the strongest predictor of win percentage, with a value higher than that of shot percentage in the uncontrolled data. Its residual diagnostics are improved and fall within the assumptions of linear regression models (see Figure 8 in Appendix A). Figure 3.6 illustrates the relationship between win percentage and Corsi percentage.

Table 3.9: Regression results: Regulation goal percentage on full-strength tied regulation shot metrics

	<i>Dependent variable:</i>		
	Goal%		
	(1)	(2)	(3)
Shot%	0.756***		
(Std. Err.)	(0.089)		
Fenwick%		0.761***	
(Std. Err.)		(0.087)	
Corsi%			0.779***
(Std. Err.)			(0.084)
Constant	0.123***	0.120***	0.111***
(Std. Err.)	(0.045)	(0.043)	(0.042)
Observations	120	120	120
R ²	0.378	0.395	0.420
Adjusted R ²	0.372	0.390	0.415
Residual Std. Error (df = 118)	0.032	0.032	0.031
F Statistic (df = 1; 118)	71.561***	77.051***	85.511***

Note: *p<0.1; **p<0.05; ***p<0.01
(1), (2), and (3) are model IDs for each metric

Regressing on percentage of all goals – regardless of situation or strength – shows full-strength tied Corsi percentage has a higher adjusted R^2 than the previous best model. Figure 3.7 illustrates this relationship.

Figure 3.6: Regression: Regulation win percentage vs Full-strength tied Corsi percentage (2012-13 excluded)

Figure 3.7: Regression: Regulation goal percentage vs Full-strength tied Corsi percentage (2012-13 excluded)

3.4.3 Controlling for strength, situation, and venue

The last experiment splits the five-on-five tied data into home and away games. Regulation win percentage refers to the regulation win percentage at home for home games, and the away percentage for away games. The goal percentages are also split into home and away percentages. As before, let y_{i1} be the team's win percentage and y_{i2} be the goal percentage. Our predictors are the shot metric, x_{i1} , and an indicator variable for home status, x_{i2} . Our models are:

$$y_{i1} = \beta_{01} + \beta_{11}x_{i1} + \beta_{21}x_{i2} + \beta_{31}x_{i1}x_{i2} + \varepsilon_{i1} \quad \varepsilon_{i1} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_{11}^2)$$

$$y_{i2} = \beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2} + \beta_{32}x_{i1}x_{i2} + \varepsilon_{i2} \quad \varepsilon_{i2} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_{22}^2)$$

Letting X be the usual matrix of observed values, we write the multivariate model $Y = XB + \epsilon$ such that:

$$Y = \begin{bmatrix} y_{11} & y_{12} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{bmatrix} = \begin{bmatrix} Y_{(1)} & Y_{(2)} \end{bmatrix}, B = \begin{bmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \end{bmatrix} = \begin{bmatrix} \beta_{(1)} & \beta_{(2)} \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \vdots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} \end{bmatrix} = \begin{bmatrix} \epsilon_{(1)} & \epsilon_{(2)} \end{bmatrix}$$

Under this model, $Cov(\epsilon_{(j)}, \epsilon_{(k)}) = \sigma_{jk}\mathbf{I}$ and $E(\epsilon_{(j)}) = \mathbf{0}$.

Table 3.10: Regression results: Regulation win percentage and regulation goal percentage on shot metrics and home indicator, with interaction

Model	Variable	Test stat	F value	DF(num; den)	Pr(>F)
(1)	Shot%	0.267	42.724	2; 235	2.2e-16
(1)	Home	0.147	20.209	2; 235	7.971e-09
(1)	Home*Shot%	0.002	0.206	2; 235	0.8137
(2)	Fenwick%	0.296	49.427	2; 235	2.2e-16
(2)	Home	0.127	17.159	2; 235	1.107e-07
(2)	Home*Fenwick%	0.003	0.337	2; 235	0.714
(3)	Corsi%	0.313	53.616	2; 235	2.2e-16
(3)	Home	0.130	17.619	2; 235	7.419e-08
(3)	Home*Corsi%	0.003	0.384	2; 235	0.6817

Note: (1), (2), and (3) are model IDs for each metric

Models (1), (2), and (3) in Table 3.10 use shot percentage, Fenwick percentage, and Corsi percentage, respectively. The interaction term was not significant in any of the three models, so we reduce them to:

$$y_{i1} = \beta_{01} + \beta_{11}x_{i1} + \beta_{21}x_{i2} + \varepsilon_{i1} \quad \varepsilon_{i1} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_{11}^2)$$

$$y_{i2} = \beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2} + \varepsilon_{i2} \quad \varepsilon_{i2} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_{22}^2)$$

Accordingly, we reduce our B matrix to:

$$B = \begin{bmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix}$$

We proceed with the interaction-free models.

Table 3.11: Regression results: Regulation win percentage and regulation goal percentage on shot metrics and home indicator, no interaction

Model	Variable	Test stat	F value	DF(num; den)	Pr(>F)
(1)	Shot%	0.266	42.854	2; 236	2.2e-16
(1)	Home	0.331	58.354	2; 236	2.2e-16
(2)	Fenwick%	0.296	49.630	2; 236	2.2e-16
(2)	Home	0.341	61.032	2; 236	2.2e-16
(3)	Corsi%	0.313	53.842	2; 236	2.2e-16
(3)	Home	0.346	62.522	2; 236	2.2e-16

Note: (1), (2), and (3) are model IDs for each metric

This time all variables are significant in each model. We now examine the fits of the univariate models.

Table 3.12: Regression results: Regulation win percentage on venue and full-strength tied shot metrics

	<i>Dependent variable:</i>		
	Win%		
	(1)	(2)	(3)
Home (Std. Err.)	0.054*** (0.012)	0.051*** (0.013)	0.050*** (0.012)
Shot% (Std. Err.)	1.176*** (0.157)		
Fenwick% (Std. Err.)		1.185*** (0.157)	
Corsi% (Std. Err.)			1.241*** (0.157)
Constant (Std. Err.)	-0.238*** (0.076)	-0.240*** (0.076)	-0.268*** (0.076)
Observations	240	240	240
R ²	0.334	0.337	0.349
Adjusted R ²	0.329	0.331	0.344
Residual Std. Error (df = 237)	0.089	0.089	0.088
F Statistic (df = 2; 237)	59.561***	60.211***	63.610***

Note: *p<0.1; **p<0.05; ***p<0.01
(1), (2), and (3) are model IDs for each metric

Corsi percentage is again the strongest of the three predictors and adding the home ice indicator improved the model's adjusted R^2 by approximately 0.047.

Table 3.13: Regression results: Regulation goal percentage on venue and regulation full-strength tied shot metrics

	<i>Dependent variable:</i>		
	Goal%		
	(1)	(2)	(3)
Home (Std. Err.)	0.034*** (0.006)	0.031*** (0.005)	0.031*** (0.005)
Shot% (Std. Err.)	0.644*** (0.070)		
Fenwick% (Std. Err.)		0.675*** (0.068)	
Corsi% (Std. Err.)			0.702*** (0.068)
Constant (Std. Err.)	0.161*** (0.034)	0.147*** (0.033)	0.134*** (0.033)
Observations	240	240	240
R ²	0.457	0.476	0.489
Adjusted R ²	0.452	0.472	0.485
Residual Std. Error (df = 237)	0.039	0.039	0.038
F Statistic (df = 2; 237)	99.649***	107.738***	113.598***

Note: *p<0.1; **p<0.05; ***p<0.01
(1), (2), and (3) are model IDs for each metric

Full-strength tied Corsi percentage is again the strongest predictor for goal percentage, and the adjusted R^2 rises from 0.415 to 0.485. However, since each team provided two observations per season, there is correlation between the residuals for home and away games. Future models could remedy this with repeated measures designs.

3.4.4 Conclusions on shot metrics

Without controlling for situation or strength, raw shot percentage appears to be the best predictor for win and goal percentages. However, as the previous section illustrated, not all shot events are generated equally; as the game changes, so does each team's strategy. Evidence points to even-strength tied Corsi percentage as the strongest predictor of win and goal percentage, suggesting that it better reflects a team's abilities.

This is not an altogether shocking idea. While a Corsi event may not reach the net, it is an attack – one team has taken action, and the other is forced to react. This is not reflected in the shot record, as missed and blocked shots are not traditionally considered threats. The importance of a Corsi event is not that it is a direct threat, but rather that for that instant one team has control over the other, compelling them into a defensive response.

One significant benefit to using the Corsi is its sheer availability compared to other metrics. Looking at even-strength tied data, from 2009-10 to 2013-14, home teams saw approximately 94,000 Corsi-for events and 88,000 Corsi-against. In comparison, they took about 50,000 shots and faced about 47,000. This considerable difference almost doubles the number of events that we may use to evaluate teams and players.

3.5 Special teams

Special teams take the ice whenever there is a change in strength. Powerplay units are measured by their ability to score goals, and penalty kill units by their ability to limit them. We will measure powerplay effectiveness by goals scored per two minutes of powerplay time (higher is better), and penalty kill effectiveness by goals allowed per two minutes of shorthanded time (lower is better). We will use them as predictors of regulation win percentage in linear regression. We test them in separate models, as well as combined models with and without interactions. This will allow us to measure their independent and combined predictive values. Letting y_i be the regulation win percentage, x_{1i} the team's goals scored per 2 minutes of powerplay time, and x_{2i} the team's goals allowed per 2 minutes of shorthanded time, our models are:

$$(1) y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \quad \varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$$(2) y_i = \beta_0 + \beta_1 x_{2i} + \varepsilon_i \quad \varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$$(3) y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \quad \varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$$(4) y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i \quad \varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

Table 3.14: Regression results: Regulation win percentage on special teams

	<i>Dependent variable:</i>			
	Win%			
	(1)	(2)	(3)	(4)
PP goals for (Std. Err.)	0.908*** (0.193)		0.885*** (0.177)	1.978* (1.012)
PK goals against (Std. Err.)		-0.807*** (0.176)	-0.785*** (0.161)	0.232 (0.941)
PP/PK interaction (Std. Err.)				-4.897 (4.463)
Constant (Std. Err.)	0.181*** (0.042)	0.554*** (0.039)	0.358*** (0.053)	0.130 (0.214)
Observations	120	120	120	120
R ²	0.157	0.151	0.300	0.307
Adjusted R ²	0.150	0.144	0.288	0.289
Residual Std. Error	0.076	0.076	0.070	0.070
Residual Std. Error DF	118	118	117	116
F Statistic	22.025***	20.956***	25.084***	17.153***
F Statistic DF	(1;118)	(1;118)	(2;117)	(3; 116)

Note: *p<0.1; **p<0.05; ***p<0.01
 (1), (2), (3), and (4) are model IDs for special teams and interactions

Independently each special team has an adjusted R^2 value of approximately 0.15. When the two are combined into one model, this rises to 0.288. The two models' explanatory values are effectively additive, and adding an interaction term does not improve them, suggesting that the strengths of the powerplay and penalty kill units are independent of each other. Since the estimated slopes for the predictors are similar and the adjusted R^2 values are close, it is not clear if the powerplay or penalty kill contributes more to a team's success.

3.6 Shot metrics and special teams

Combining Corsi percentage with special teams is also informative. We will examine models that account for venue and test interactions. In the models accounting for venue, the regulation win percentage is the home regulation win percentage for home games, and the away percentage for away games. Let y_i be the regulation win percentage, x_{1i} the Corsi percentage, x_{2i} the goals scored per 2 minutes of powerplay time, x_{3i} the goals allowed per 2 minutes of penalty kill time, and x_{4i} the indicator for a home game. Our models are:

$$(1a) \ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \quad \varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$$(1b) \ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{1i} x_{2i} + \beta_5 x_{1i} x_{3i} + \varepsilon_i \quad \varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$$(2a) \ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i \quad \varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$$(2b) \ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{2i} x_{4i} + \beta_6 x_{3i} x_{4i} + \varepsilon_i \quad \varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

Table 3.15: Regression results: Regulation win percentage on full-strength tied Corsi percentage and special teams

	<i>Dependent variable:</i>	
	Win%	
	(1a)	(1b)
Corsi% (5v5, Tied)	1.043***	-0.114
(Std. Err.)	(0.181)	(1.416)
PP goals for	0.775***	1.208
(Std. Err.)	(0.158)	(2.376)
PK goals against	-0.459***	-3.566
(Std. Err.)	(0.153)	(2.249)
Corsi*PP		-0.871
(Std. Err.)		(4.716)
Corsi*SH		6.186
(Std. Err.)		(4.468)
Constant	-0.211*	0.374
(Std. Err.)	(0.109)	(0.716)
Observations	120	120
R ²	0.456	0.465
Adjusted R ²	0.442	0.441
Residual Std. Error	0.062 (df = 116)	0.062 (df = 114)
F Statistic	32.375*** (df = 3; 116)	19.806*** (df = 5; 114)
<i>Note:</i>		
*p<0.1; **p<0.05; ***p<0.01		
(1a) and (1b) are model IDs		

The VIFs for the coefficients in Model (1a) are 1.18, 1.02, and 1.16 for Corsi, powerplay effectiveness, and penalty kill effectiveness, respectively. They do not suggest problems with multicollinearity. Adding interaction terms between Corsi% and special team effectiveness does not strengthen the model, as the adjusted R^2 in Model (1b) shows.

Table 3.16: Regression results: Regulation win percentage on venue, full-strength tied Corsi percentage, and special teams

	<i>Dependent variable:</i>	
	Win%	
	(2a)	(2b)
Home	0.040***	-0.178
(Std. Err.)	(0.012)	(0.170)
Corsi%	1.035***	0.919***
(Std. Err.)	(0.149)	(0.206)
PP goals for	0.616***	0.423***
(Std. Err.)	(0.111)	(0.149)
PK goals against	-0.436***	-0.385***
(Std. Err.)	(0.110)	(0.145)
Home*PP		0.524**
(Std. Err.)		(0.226)
Home*PK		-0.211
(Std. Err.)		(0.223)
Home*Corsi		0.296
(Std. Err.)		(0.296)
Constant	-0.198**	-0.113
(Std. Err.)	(0.084)	(0.112)
Observations	240	240
R ²	0.457	0.475
Adjusted R ²	0.447	0.459
Residual Std. Error	0.081 (df = 235)	0.080 (df = 232)
F Statistic	49.383*** (df = 4; 235)	29.965*** (df = 7; 232)
<i>Note:</i>		
*p<0.1; **p<0.05; ***p<0.01		
(2a) and (2b) are model IDs		

The VIFs for the coefficients in Model (2a) are 1.22, 1.04, 1.09, and 1.29 for the home indicator, powerplay effectiveness, penalty kill effectiveness, and Corsi, respectively, suggesting no issues with multicollinearity. Model (2b), which contains interaction terms, offers little improvement in adjusted R^2 over Model (2a). There is a possible interaction between home games and powerplay effectiveness. However, since teams provided two observations per season, there was correlation between their residuals. Future examination could make use of repeated measures methodology.

Comparing Models (1a) and (2a), the adjusted R^2 is effectively the same regardless of the indicator for home games. Model (1a) has considerable explanatory value with an adjusted R^2 of 0.442 compared to models with Corsi percentage and home ice (0.344) and special teams (0.288). Within this model, the estimated slopes suggest that full-strength Corsi percentage contributes the most to victory, while a stronger powerplay adds more than what a weaker penalty kill would take away.

3.7 Save percentage

Save percentage is the ratio of saves made over shots on a non-empty net. For this study we consider the save percentage of a team rather than a goaltender. Regardless of situation, scatterplots (see Figure 3.8) and linear regression suggest no evidence of a link between save percentage and shot metrics. We test one model with three datasets. Letting y_i be the the save percentage and x_i be the Corsi percentage, our model is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

Table 3.17: Regression results: Regulation save percentage on regulation Corsi percentage

	<i>Dependent variable:</i>		
	Save%		
	(1)	(2)	(3)
Corsi%	0.049*	0.034	0.039
(Std. Err.)	(0.026)	(0.026)	(0.027)
Constant	0.900***	0.901***	0.890***
(Std. Err.)	(0.013)	(0.013)	(0.014)
Observations	150	150	150
R ²	0.025	0.011	0.014
Adjusted R ²	0.018	0.004	0.007
Residual Std. Error (df = 148)	0.011	0.009	0.009
F Statistic (df = 1; 148)	3.736*	1.646	2.086

Note:

*p<0.1; **p<0.05; ***p<0.01

(1), (2), and (3), are model IDs for five-on-five, five-on-five tied, and all data, respectively

Models (1), (2), and (3) use tied five-on-five data, all five-on-five data, and all data, respectively. All three failed to demonstrate a relationship between the variables. There was no statistically significant link between Corsi percentage

Figure 3.8: Scatterplot of regulation Corsi percentage and regulation save percentage

and save percentage.

3.8 Data visualization

Visualization is a valuable tool when working with large quantities of data, particularly for identifying relationships or biases. Figures 3.9 – 3.16 are team per-minute Corsi rates plotted against their Corsi percentages. The data points are separated by conference and home/away and are augmented with auxiliary information about a team’s playoff finish (indicated by colour) and whether they won the championship (indicated by shape). An x-intercept is provided to identify teams with Corsi percentages above or below 0.500, and a y-intercept is set to 1 to provide a reference point for Corsi events per minute.

Several things become clear after looking at the images. All five of the most recent champions had Corsi percentages well past 0.500 at home. Teams with percentages below 0.500, away or at home, tended to be eliminated early or to not qualify for the playoffs. Some teams that qualified for the playoffs despite poor shot metrics, such as the 2012-13 Toronto Maple Leafs, stand out as outliers.

Teams such as Los Angeles and Chicago have impressive shot metrics relative to the rest of the league and it becomes clear why they have been the most dominant teams over the last five seasons. Amongst the champions, the Boston Bruins stand out as less dominant when looking at Corsi alone. However, this underscores that the game cannot be predicted entirely with one or two metrics. Factors such as save percentage, special teams, and how well teams defend a lead or recover from a deficit are also important. The Bruins in particular were noted for their strong goaltending and defense, featuring elite players like Zdeno Chara and Tim Thomas.

Characteristics of data collection are also evident in the graphics. Home statistics look noisier than away statistics, which is not surprising as the NHL does not have consistent record-keeping practices between arenas. Thus, these

differences tend to average out in a team's away games, but may become amplified at home; the New Jersey Devils are particularly noteworthy for forming their own cluster among home statistics. Schuckers and Macdonald[20] addressed the issue of rink bias in 2014 and specifically noted the Devils' unusually low counts of blocked and missed shots.

Figure 3.9: Corsi against per minute vs Corsi percentage, Eastern Conference, away games

Figure 3.10: Corsi against per minute vs Corsi percentage, Western Conference, away games

Figure 3.13: Corsi for per minute vs Corsi percentage, Eastern Conference, away games

Figure 3.14: Corsi for per minute vs Corsi percentage, Western Conference, away games

3.9 Player analysis via association rule learning

Analysis of player performance is complicated by factors that are often out of their control. Each player’s performance is subject to implicit facets of the game: coaches prioritize skill sets to the detriment or benefit of a player’s output; line mates can have drastically different levels of talent, buttressing weaker players or limiting stronger ones; strong teams offer chances to excel, while poor ones may limit them.

These myriad constraints stress the importance of context in examining any performance metric. A player’s role will determine how to consider their effectiveness. Some players are used for a particular aspect of their play. In the case of a centre whose main responsibility is the penalty kill, it would be unwise to judge him solely on even-strength performance; his performance at even-strength weighed against his ability to kill penalties would make a better measure. Further, it is critical to avoid divorcing numbers from reality: this is a sport played by people. There are many aspects that cannot be measured, such as commitment to a system, relationships with coaching, management and other players, and personal factors like injuries or life events.

Nonetheless, we can explore new methods of assessing player performance. In particular, this thesis examines association rule learning.¹

3.9.1 Description of association rule learning

Association rule learning is a data mining technique intended to highlight potentially interesting relationships between variables in large databases. In sales applications it is also known as market basket analysis, where it attempts to find

¹Though I developed this method independently, I discovered shortly before the submission deadline that it had been proposed before and applied in other sports. I would like to acknowledge the work of Hipp and Mazlack in 2011 at the first International Conference on Advances in Information Mining and Management, where they concluded that there now exist enough data in ice hockey to make meaningful contributions to our understanding and enjoyment of the game.

items that are purchased together. By highlighting interesting relationships it can shed light on sets of items that could potentially increase revenues, interest in items, etc.

The problem was formally defined in 1993 by Agrawal, Imielinsky, and Swami[21]. We define a set of binary variables called *items* as $I = \{i_1, \dots, i_m\}$, and a database of transactions $D = \{t_1, \dots, t_n\}$. Transactions are binary vectors where $t_k = 1$ if the k^{th} item was purchased and 0 otherwise, for $k \in [1, m]$. An *association rule* is an implication $X \Rightarrow Y$, where $X, Y \subseteq I$ and $X \cap Y = \emptyset$; that is, we look at transactions containing items in X that appear with items in Y . X is referred to as the *antecedent* or *left-hand-side (LHS)* of the rule, and Y is the *consequent* or *right-hand-side (RHS)*. To highlight specific types of rules we introduce several constraints:

- *Syntactic constraints* limit the items that may appear in either part of a rule.
- The *support* of a set of items X , written $Supp(X)$, is the proportion of transactions in D that contain X .
- The *confidence* of $X \Rightarrow Y$ is the ratio of the support of X and Y together to the support of X alone. We write $Conf(X \Rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X)}$. Note that $X \cup Y$ refers to the union of items, which differs from the statistical convention of writing $P(X \cap Y)$.
- The *lift* of $X \Rightarrow Y$ is $Lift(X \Rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X)Supp(Y)}$.
- The *difference of confidence* of $X \Rightarrow Y$ is $Conf(X \Rightarrow Y) - Conf(\bar{X} \Rightarrow Y)$

There are numerous other constraints and interest measures, but we focus on these. If we treat the itemsets as events, we can interpret our interest measures probabilistically:

- $Supp(X) = P(X)$
- $Conf(X \Rightarrow Y) = P(Y|X)$
- $Lift(X \Rightarrow Y) = \frac{P(X \text{ and } Y)}{P(X)P(Y)}$. Note that “ X and Y ” refers to itemsets occurring within the same transaction – recall that this is written as $X \cup Y$ when referring to our itemsets.
- $DOC(X \Rightarrow Y) = P(Y|X) - P(Y|\bar{X})$

3.9.2 Applying association rules to ice hockey

We apply syntactic constraints so that the antecedent X contains only players and the consequent Y contains only in-game events. Thus our interpretation is:

- $Supp(X \cup Y)$ is the probability that a player combination X and event Y occur together. This is implicitly a function of ice time; players that spend more time on the ice will generally be on for more events.
- $Conf(X \Rightarrow Y)$ represents the probability that an event Y occurs given player combination X is on the ice.
- $Lift(X \Rightarrow Y)$ represents the independence of the appearance of X and Y . A value close to 1 suggests that player set X is independent of event Y ; values greater than 1 suggest a positive relationship between the players and event; values less than 1 suggest the opposite.
- $DOC(X \Rightarrow Y)$ represents the difference in probability of an event Y given the presence of a player set X .

Careful consideration must be given to how the transactional database is created and interpreting its results. The choice of events will define the sample space and impact any assessments made. A transactional database that contains

procedural events such as period endings and TV timeouts can impact more meaningful events since it changes their support levels. The choices for minimum thresholds for support and confidence will also have impacts since important events, such as goals, may be especially rare in a given database.

3.9.3 Example: 2013-14 Toronto Maple Leafs

We consider the 2013-14 Toronto Maple Leafs' regular season. Our binary database consists of all non-goalie players who were on the ice for a Corsi event; our database is limited to five-on-five Corsi events when the game is tied. We treat Corsi events equally and do not consider factors such as the quality of competition or how often a player starts play in the defensive or offensive zone. While ostensibly a simple database, we can make meaningful inferences so long as we are mindful of our assumptions.

Table 3.18: A sample of transaction database entries

DION_PHANEUF	PHIL_KESSEL	CORSI_FOR	CORSI_AGAINST
0	1	0	1
1	1	0	1
0	1	1	0
1	0	1	0
1	0	1	0
0	0	1	0

The rules we generate will be $\{PLAYER_SET\} \Rightarrow \{CORSI_EVENT\}$. Our minimum support and confidence levels are 0.001, but we will only examine the highlights. We consider rules for individuals, pairs of individuals (split by forward and defense), and trios (forwards only). We will first examine a table of events for all individuals. The table has been split into two for space.

Table 3.19: Rules for 2013-14 Toronto Maple Leafs

rules	support	conf	lift	doc
{PHANEUF} $\Rightarrow$ {CORSLAGAINST}	0.209	0.606	1.028	0.025
{KESSEL} $\Rightarrow$ {CORSLAGAINST}	0.200	0.556	0.943	-0.052
{GARDINER} $\Rightarrow$ {CORSLAGAINST}	0.195	0.588	0.997	-0.003
{V_RIEMSDYK} $\Rightarrow$ {CORSLAGAINST}	0.189	0.554	0.939	-0.054
{GUNNARSSON} $\Rightarrow$ {CORSLAGAINST}	0.187	0.592	1.003	0.003
{KADRI} $\Rightarrow$ {CORSLAGAINST}	0.172	0.571	0.969	-0.027
{FRANSON} $\Rightarrow$ {CORSLAGAINST}	0.171	0.575	0.976	-0.021
{LUPUL} $\Rightarrow$ {CORSLAGAINST}	0.163	0.630	1.068	0.054
{RIELLY} $\Rightarrow$ {CORSLAGAINST}	0.162	0.587	0.996	-0.004
{RAYMOND} $\Rightarrow$ {CORSLAGAINST}	0.159	0.569	0.965	-0.029
{KESSEL} $\Rightarrow$ {CORSLFOR}	0.159	0.444	1.081	0.052
{V_RIEMSDYK} $\Rightarrow$ {CORSLFOR}	0.152	0.446	1.087	0.054
{BOZAK} $\Rightarrow$ {CORSLAGAINST}	0.145	0.565	0.958	-0.034
{GARDINER} $\Rightarrow$ {CORSLFOR}	0.137	0.412	1.005	0.003
{PHANEUF} $\Rightarrow$ {CORSLFOR}	0.136	0.394	0.959	-0.025

Note: Player names were truncated to preserve formatting.

The Maple Leafs were clearly a poor team when looking at shot metrics. The first rule containing Corsi-for is the 11th entry. When examining full-strength tied Corsi events, the most common – appearing 20.9% of the time – is Dion Phaneuf on the ice facing a shot attempt from the opposition. The lift on the rule is 1.028, which suggests that this event is more likely to occur when he is present. The next most common, at 20%, is Phil Kessel being present during an opposition shot attempt. His lift, however, is 0.943, strongly suggesting that his presence has a negative relationship with the event. Later in the list, we see that 15.9% of the events are shot attempts in the team’s favour, and they happen when he is on the ice. The lift is 1.081, suggesting that his presence drives the event. The difference of confidence plays a similar role, and its relatively high positive value implies that Corsi attempts in the team’s favour are more likely when Kessel is present.

Individual measurements have a limited value as they do not account for the players alongside those being measured. By expanding the LHS, these tables can

be generated for any combination of players and events that appear together. Presenting them as tables makes them difficult to parse, so we must explore visualization strategies.

Visualization of association rules

Visualization was done in R using the package **ggplot2**. Support is measured by the size of a bar, and the hue reflects the intensity of the selected interest measure. We use lift, although difference of confidence is also applicable. A stronger green indicates a higher relative lift, and a stronger purple the opposite. Colours were selected for their differences from each other; a lift closer to 1 appears grayish. Player-event combinations with the lowest support levels were excluded for space, although they also likely have analytic value.

Figure 3.17: Association rules for individuals, 2013-14 Toronto Maple Leafs

Figure 3.18: Association rules for defensemen, 2013-14 Toronto Maple Leafs

Figure 3.19: Association rules for defense pairs, 2013-14 Toronto Maple Leafs

Figure 3.20: Association rules for forwards, 2013-14 Toronto Maple Leafs

Figure 3.21: Association rules for forward pairs, 2013-14 Toronto Maple Leafs

Figure 3.22: Association rules for forward trios, 2013-14 Toronto Maple Leafs

Analysis of association rules

By reading the six plots we can infer some of the Leafs' strengths, weaknesses, and potential areas for improvement. The plot of all individuals (Figure 3.17) illustrates that the Leafs' offense is heavily driven by Phil Kessel and James van Riemsdyk. While the quantity of shot events they generate is important, the bright green on Corsi (for) – representing a relatively high lift – suggests that when they are on the ice they are adept at generating attacks. Unfortunately for each of them, the attacks they generate are outstripped by attacks by opponents, at least on an individual level. Other players, such as Joffrey Lupul and Jay McClement, appear to fare poorly as shot attempts against them heavily outnumber shot attempts for, with the lift indicating their presences may have had a negative impact.

Next we examine pairs of forwards (Figure 3.21). Again, the first line of Kessel, van Riemsdyk, and Bozak dominates shot metrics. The pairing of Kessel and van Riemsdyk has a higher lift than either of them with Bozak. However, Bozak was injured for part of the year and replaced by Nazem Kadri, who appears in pairings further down the list. This provides a valuable comparison of the first line with different centres – Kadri's lifts with van Riemsdyk and Kessel are each higher than those with Bozak.

Finally, examining the forward trios (Figure 3.22), we see that when Kadri played on the top line, it actually had a positive Corsi differential with a very strong lift, compared to the negative differential and lower lift with Bozak. Combining our inferences from the three graphs, we have compiled convincing evidence that the team would likely benefit from having Nazem Kadri as the first line centre.

Table 3.20: Forward pairs with lift

Player pair	Event	Freq	Lift
Bozak, van Riemsdyk	Corsi (F)	287	1.116
Bozak, van Riemsdyk	Corsi (A)	340	0.919
Bozak, Kessel	Corsi (F)	302	1.112
Bozak, Kessel	Corsi (A)	360	0.922
Kadri, van Riemsdyk	Corsi (F)	92	1.212
Kadri, van Riemsdyk	Corsi (A)	93	0.852
Kessel, Kadri	Corsi (F)	105	1.164
Kessel, Kadri	Corsi (A)	115	0.886
Kessel, van Riemsdyk	Corsi (F)	390	1.140
Kessel, van Riemsdyk	Corsi (A)	444	0.903

Going further, we may examine how the two lines played accounting for ice time. In full-strength tied situations, the Bozak line played 17,743 seconds together, and the Kadri line played 4,745 seconds.² Table 3.21 presents the frequency and per-60-minute rates of goals and Corsi events

Table 3.21: Kadri and Bozak lines, with rates/60 minutes

Event	Freq (Kadri)	Rate (Kadri)	Freq (Bozak)	Rate (Bozak)
Corsi (A)	78	59.178	320	64.927
Corsi (F)	85	64.489	279	56.608
Goal (A)	2	1.517	14	2.841
Goal (F)	4	3.035	19	3.855

With Kadri at centre, the line generated 64.489 Corsi events per 60 minutes and faced 59.178; in contrast, the Bozak line generated 56.608 and faced 64.927 – essentially a reversal. The goals for and against are presented as well, though the sample sizes for the Kadri line are low.

Looking at defensemen individually (Figure 3.18, Table 3.22), captain Dion Phaneuf was on the ice for a combined 34.5% of all full-strength tied Corsi events, although the vast majority were against the team. The lift suggests

²These values were estimated using the method described in the section on data collection.

his presence is associated with the opposition attacking, but one should note that the scale on this graph has a smaller range than the forwards'. His partner Carl Gunnarsson fared similarly in terms of his difference between Corsi-for and -against, but his lift is closer to 1. Jake Gardiner was on the ice for more Corsi-for events than any other defenseman, with a lift suggesting that his presence is positively related to those events. Cody Franson had the highest lift relative to the other defensemen indicating that his presence was more associated with a Leafs attack.

Figure 3.19 shows that the Phaneuf-Gunnarsson pairing has a lift that associates them slightly with Corsi events against the team and slightly more against Corsi events in their favour. Gardiner-Franson perform slightly better, but the lifts for the pairing are very close to 1 meaning the events could be independent of their presence. Gardiner-Rielly, however, came closest to parity and their lift indicates their presence is strongly related to Corsi-for events. This could be an indication of their skill, or some combination of their chemistry, skill, and the quality of opposition that they face.

To investigate further, we look at player ice times listed in Tables 3.22 and 3.23.

Table 3.22: Defensemen, sorted by TOI (seconds)

Player	TOI	CF60	Lift (F)	CA60	Lift (A)
Gardiner	29,827	46.106	1.005	65.779	0.997
Phaneuf	28,311	48.193	0.959	74.261	1.028
Gunnarsson	27,524	47.086	0.995	68.275	1.003
Franson	26,621	47.602	1.035	64.505	0.976
Rielly	22,885	49.867	1.006	70.946	0.996
Ranger	16,608	45.087	1.016	63.078	0.989
Gleason	13,139	49.593	0.998	71.512	1.001

Note: CF60 and CA60 refer to Corsi For/Against per 60 minutes

Table 3.23: Defense pairs, sorted by TOI (seconds)

Players	TOI	CF60	Lift (F)	CA60	Lift (A)
Phaneuf, Gunnarsson	21,678	47.163	0.978	70.412	1.015
Gardiner, Franson	12,847	43.154	1.004	61.649	0.997
Gardiner, Ranger	6,022	44.836	1.016	62.770	0.989
Gardiner, Rielly	5,654	56.668	1.130	65.582	0.910
Franson, Rielly	5,611	51.328	1.016	71.859	0.989
Gleason, Rielly	4,556	45.040	0.897	77.436	1.072
Gleason, Franson	4,212	54.701	1.091	67.521	0.937
Ranger, Rielly	3,634	46.560	1.023	64.392	0.984
Phaneuf, Gleason	2,045	54.572	1.021	75.697	0.985
Gardiner, Phaneuf	1,922	48.699	0.845	91.779	1.108
Gardiner, Gunnarsson	1,707	42.179	0.855	78.032	1.101
Franson, Ranger	1,223	32.379	0.958	50.041	1.029
Gunnarsson, Ranger	1,211	38.646	0.813	77.291	1.130

Note: CF60 and CA60 refer to Corsi For/Against per 60 minutes

The beleaguered Phaneuf-Gunnarsson pairing faced 70.412 Corsi-against per 60 minutes – bad, but not the worst on the list. Looking at individuals, Phaneuf alone faces 74.261 Corsi events against per 60 minutes, while Gunnarsson sees 68.275. Phaneuf’s pairing with Gardiner also fared very poorly, facing an enormous 91.779 shot attempts but generating only 48.699. Phaneuf may have been struggling in his role or within the system coach Randy Carlyle implemented at the time.

The Gardiner-Rielly pairing, earlier highlighted for its very high lift for favourable Corsi events, has the highest rate of Corsi-for events, with 56.668 per 60 minutes. This rate is higher than either player's individual CF60, suggesting – as the high lift indicates – that they have good chemistry together.

3.10 Future applications of association rules

There is significant potential for exploration of events in hockey games beyond those presented here. Many recorded events can be customized or modified to be more informative. Takeaways and giveaways can be categorized by whether they occurred in a defensive or offensive context, and offsides and icings, while not officially credited against a team in the RTSS, can be categorized as advantageous or disadvantageous to a team. Mapping these results to the players on the ice could serve as a measure of which are effective at stymieing offense and which have difficulty breaking through defenses.

Analysis is not limited to full-strength tied play. Given a well-defined database of events, it will be possible to objectively evaluate the performance of unusual or novel tactics, such as powerplays with five forwards or reintroducing the rover as position. Additionally, this analysis separated forwards and defensemen – there is value in examining the contributions of combinations like two forwards and one defenseman.

New sources of data present the greatest potential for application. Amateur efforts, such as those to manually track passes, zone exits, and zone entries, will create tens of thousands of data points to analyze. Location-tracking technology that follows player and puck movements presents an even greater quantity of data. One could define a puck as being in a player's possession if it is within a certain proximity of a player; once it enters the other team's possession, it can be categorized as a turnover. Over a season it would be possible to make inferences about which players are more successful at controlling or driving play given a particular context. By carefully defining the events that we wish to examine together, we can add significantly to our understanding of the game; given the quantity of data that analysts will soon face, it will be necessary to apply data mining techniques.

Chapter 4

Conclusion

4.1 Team metrics

There are several clear conclusions for team metrics. Over a full season, goal percentage remains the strongest predictor of success, for obvious reasons. Home ice advantage was clear from the data examined, with home teams outscoring away teams regularly. However, because of their relative rarity, goals have limited analytical value, which motivates the use of shot metrics.

Shot metrics are subject to score effects, where the score of the game determines the intensity at which a team attacks. On average, teams with a lead play a more passive role, while teams in a deficit are more aggressive. This is especially pronounced in the third period. Tied teams at full-strength have a mean Corsi percentage of roughly 0.5. Using raw data, linear regression shows shot percentage is the superior metric for explaining regulation win and goal percentages; when limiting data to full-strength tied situations, Corsi percentage becomes the strongest variable. Adding an indicator variable for home/away status improves predictive power further. The predictive power of Corsi events

is a boon for analysis as they are relatively abundant compared to shots. Linear regression on special teams alone against regulation win percentage shows each of them explains roughly the same amount of variation. No relationship between save percentage and shot metrics could be found.

4.2 Association rules

Association rule learning has demonstrable value in player evaluation. While full-strength tied Corsi is not a new metric, within a team there are hundreds of potential player combinations, with dozens of those regularly realized over the course of a season. Using current techniques results in hundreds of aggregated data points – a rich but unfortunately labyrinthine collection. Association rule learning allows users to examine many of these combinations at once, filtered through a number of constraints and interest measures. From there one can decide on whichever measures are most relevant for investigation; this is, at its core, the purpose of data mining.

In the example case of the Toronto Maple Leafs, we believe it has identified strengths and weaknesses in individuals and groups alike. Visualizations of interest measures highlight team performance and player chemistry. These highlights can be examined further within their relevant contexts, adding information for players, coaches, and management and informing decisions about strategies and tactics, as well as personnel selection.

There is great potential for further exploration of hockey statistics using association rule learning, particularly given the expected explosion of data from newly introduced player- and puck-tracking technologies. Careful use of statistical methods and data mining techniques will only serve to better inform the hockey world, seeking not to displace traditional evaluation methods, but rather to augment them. Players and tactics that have gone unnoticed in the

past might now be revealed, possibly translating into complementary benefits: for teams, a more competitive performance; and for fans, a more creative, more entertaining game.

Appendix A

Figure 1: Win% on Goal% regression diagnostics (1)

Figure 2: Win% on Goal% regression diagnostics (2)

Figure 3: Win% on Goal% regression diagnostics (3)

Figure 4: Win% on Goal% regression diagnostics (4)

Figure 5: Win% on Shot% regression diagnostics (QQ)

Figure 6: Win% on Fenwick% regression diagnostics (QQ)

Figure 7: Win% on Corsi% regression diagnostics (QQ)

Figure 8: Win% on Corsi% (5v5, tied) regression diagnostics (QQ)

Figure 9: Autocorrelation function for Win% on Corsi% (5v5, tied) to test autocorrelation along seasons. Winnipeg/Atlanta were removed to avoid the possible impact of a team changing cities.

References

- [1] Arik Parnass. *Analytics, not statistics, driving NHL evolution*. Feb. 22, 2015. URL: <http://www.nhl.com/ice/news.htm?id=754099>.
- [2] Eliote Friedman. *In search of Vic Ferrari*. Aug. 6, 2014. URL: <http://www.cbc.ca/sports-content/hockey/opinion/2014/08/in-search-of-vic-ferrari.html>.
- [3] Brian Macdonald. “Adjusted Plus-Minus for NHL Players using Ridge Regression with Goals, Shots, Fenwick, and Corsi”. In: *Journal of Quantitative Analysis in Sports* 8 (3 Oct. 2012). DOI: 10.1515/1559-0410.1447.
- [4] Brian Macdonald. “An Improved Adjusted Plus-Minus Statistic for NHL Players”. In: *Sloan Sports Analytics Conference*. MIT. Mar. 2011. URL: <http://www.sloansportsconference.com/?p=2838>.
- [5] Brian Macdonald. “An Expected Goals Model for Evaluating NHL Teams and Players”. In: *Sloan Sports Analytics Conference*. MIT. Mar. 2012. URL: <http://www.sloansportsconference.com/?p=6135>.
- [6] Robert B. Gramacy, Shane T. Jensen, and Matt Taddy. “Estimating player contribution in hockey with regularized logistic regression”. In: *Journal of Quantitative Analysis in Sports* 9 (1 Mar. 2013), pp. 97–111. DOI: 10.1515/jqas-2012-0001.

- [7] Michael E. Schuckers, Dennis F. Lock, Chris Wells, C. J. Knickerbocker, and Robin H. Lock. *National Hockey League Skater Ratings Based upon All On-Ice Events: An Adjusted Minus/Plus Probability (AMPP) Approach*. 2011. URL: <http://myslu.stlawu.edu/~msch/sports/LockSchuckersProbPlusMinus113010.pdf>.
- [8] Michael E. Schuckers and James Curro. “Total Hockey Rating (THoR): A comprehensive statistical rating of National Hockey League forwards and defensemen based upon all on-ice events”. In: *Sloan Sports Analytics Conference*. MIT. Mar. 2013. URL: <http://www.sloansportsconference.com/?p=10193>.
- [9] Nathan Spagnola. “The Complete Plus-Minus: A Case Study of The Columbus Blue Jackets”. Master’s thesis. University of South Carolina - Columbia, 2013. URL: <http://scholarcommons.sc.edu/etd/2708>.
- [10] Claude B. Vincent and Byron Eastman. “Defining the Style of Play in the NHL: An Application of Cluster Analysis”. In: *Journal of Quantitative Analysis in Sports* 5.10 (1 2009), p. 1133.
- [11] Sean McIndoe. *Farewell, Enforcers: A Eulogy for an NHL Institution*. Nov. 14, 2014. URL: <http://grantland.com/the-triangle/farewell-enforcers-a-eulogy-for-an-nhl-institution/>.
- [12] Samuel Buttrey, Alan Washburn, and Wilson Price. “Estimating NHL Scoring Rates”. In: *Journal of Quantitative Analysis in Sports* 7 (3 July 2011), p. 1334.
- [13] Patrice Marek, Blanka Šedivá, and Tomáš ěoupal. “Modeling and prediction of ice hockey match results”. In: *Journal of Quantitative Analysis in Sports* 10 (3 Sept. 2014), pp. 357–365. DOI: 10.1515/jqas-2013-0129.

- [14] Paramjit S. Gill. “Late-Game Reversals in Professional Basketball, Football, and Hockey”. In: *The American Statistician* 54 (2 May 2000), pp. 94–99. DOI: 10.1080/00031305.2000.10474518.
- [15] Andrew C Thomas. “Inter-arrival Times of Goals in Ice Hockey”. In: *Journal of Quantitative Analysis in Sports* 3 (3 July 2007), p. 1064.
- [16] Eric Tulskey, Geoffrey Detweiler, Robert Spencer, and Corey Sznajder. “Using Zone Entry Data To Separate Offensive, Neutral, And Defensive Zone Performance”. In: *Sloan Sports Analytics Conference*. MIT. Mar. 2013. URL: <http://www.sloansportsconference.com/?p=10561>.
- [17] R Pollard and G Pollard. “Long-term trends in home advantage in professional team sports in North America and England (1876 - 2003)”. In: *Journal of Sports Sciences* 23 (4 Apr. 2005), pp. 337–350. DOI: 10.1080/02640410400021559.
- [18] Vincent L. Liardi and Albert V. Carron. “An analysis of National Hockey League face-offs: Implications for the home advantage”. In: *International Journal of Sport and Exercise Psychology* 9 (2 June 2011), pp. 102–109. DOI: 10.1080/1612197X.2011.567100.
- [19] Tore Purdy. *Shots, Fenwick and Corsi*. Feb. 16, 2011. URL: <http://objectivenhl.blogspot.ca/2011/02/shots-fenwick-and-corsi.html>.
- [20] Michael Schuckers and Brian Macdonald. *Accounting for Rink Effects in the National Hockey League’s Real Time Scoring System*. Dec. 2, 2014. URL: http://statsportsconsulting.com/main/wp-content/uploads/Schuckers_Macdonald_RinkEffects_Final.pdf.

- [21] Rakesh Agrawal, Tomasz Imielenski, and Arun Swami. “Mining association rules between sets of items in large databases”. In: *ACM SIGMOD Record* 22 (2 June 1993), pp. 207–216. DOI: 10.1145/170036.170072.
- [22] Joanne M. Doyle and Benjamin Leard. “Variations in Home Advantage: Evidence from the National Hockey League”. In: *Journal of Quantitative Analysis in Sports* 8 (2 June 2012). DOI: 10.1515/1559-0410.1446.
- [23] Andrew C Thomas. “The Impact of Puck Possession and Location on Ice Hockey Strategy”. In: *Journal of Quantitative Analysis in Sports* 2 (1 Jan. 2006), p. 1007.
- [24] Rob Vollman, Tom Awad, and Iain Fyffe. *Hockey Abstract 2014*. Jan. 2015. ISBN: 9781500717711.

Bibliography

Bovas Abraham and Johannes Ledolter. *Introduction to Regression Modeling*. Thomson Brooks/Cole, 2006. ISBN: 0-534-42075-3.

Jason Abrevaya and Robert McCulloch. “Reversal of fortune: a statistical analysis of penalty calls in the National Hockey League”. In: *Journal of Quantitative Analysis in Sports* 10 (2 June 2014), pp. 207–224. DOI: 10.1515/jqas-2013-0067.

Jim Albert. “Looking at spacings to assess streakiness”. In: *Journal of Quantitative Analysis in Sports* 9 (2 June 2013), pp. 151–163. DOI: 10.1515/jqas-2012-0015.

James Cochran and Rob Blackstock. “Pythagoras and the National Hockey League”. In: *Journal of Quantitative Analysis in Sports* 5 (2 May 2009), p. 1181.

Adam Hipp and Lawrence Mazlack. “Mining Ice Hockey: Continuous Data Flow Analysis”. In: IMMM 2011, The First International Conference on Advances in Information Mining and Management. IARIA, Oct. 23, 2011, pp. 31–36. ISBN: 978-1-61208-162-5. URL: http://www.thinkmind.org/index.php?view=article&articleid=immm_2011_2_10_20124.

Ali Jarvandi, Shahram Sarkani, and Thomas Mazzuchi. “Modeling team compatibility factors using a semi-Markov decision process: a data-driven

approach to player selection in soccer”. In: *Journal of Quantitative Analysis in Sports* 9 (4 Dec. 2013). DOI: 10.1515/jqas-2012-0054.

Marshall Jones. “Responses to Scoring or Conceding the First Goal in the NHL”. In: *Journal of Quantitative Analysis in Sports* 7 (3 July 2011), pp. 1324–.

Benjamin Leard and Joanne Doyle. “The Effect of Home Advantage, Momentum, and Fighting on Winning in the National Hockey League”. In: *Journal of Sports Economics* 12 (5 Oct. 2011), pp. 538–560. DOI: 10.1177/1527002510389869.

Xinrong Lei and Brad R. Humphreys. “Game importance as a dimension of uncertainty of outcome”. In: *Journal of Quantitative Analysis in Sports* 9 (1 Mar. 2013), pp. 25–36. DOI: 10.1515/jqas-2012-0019.

Daniel Nevo and Ya’acov Ritov. “Around the goal: examining the effect of the first goal on the second goal in soccer using survival analysis methods”. In: *Journal of Quantitative Analysis in Sports* 9 (2 June 2013), pp. 165–177. DOI: 10.1515/jqas-2012-0004.

Garritt L. Page, Bradley J. Barney, and Aaron T. McGuire. “Effect of position, usage rate, and per game minutes played on NBA player production curves”. In: *Journal of Quantitative Analysis in Sports* 9 (4 Dec. 2013), pp. 337–345. DOI: 10.1515/jqas-2012-0023.

H. Pileggi, C. D. Stolper, J. M. Boyle, and J. T. Stasko. “SnapShot: Visualization to Propel Ice Hockey Analytics”. In: *IEEE Transactions on Visualization and Computer Graphics* 18 (12 Dec. 2012), pp. 2819–2828. DOI: 10.1109/TVCG.2012.263.

Manuel Ruiz, Jose A. Martinez, Fernando A. López-Hernández, and Almudena Castellano. “The relationship between concentration of scoring

and offensive efficiency in the NBA”. In: *Journal of Quantitative Analysis in Sports* 10 (1 Jan. 2014), pp. 27–36. DOI: 10.1515/jqas-2013-0060.

Michael E. Schuckers. “DIGR: A Defense Independent Rating of NHL Goaltenders using Spatially Smoothed Save Percentage Maps”. In: *Sloan Sports Analytics Conference*. MIT. Mar. 2011. URL: <http://www.sloansportsconference.com/?p=648>.

Robert Vollman. *Howe and Why. Ten Ways to Measure Defensive Contributions*. Mar. 4, 2010. URL: <http://www.hockeyprospectus.com/puck/article.php?articleid=480>.

Software

Python Software Foundation. *Python*. Version 2.7. 2015. URL: <https://www.python.org/>.

Michael Hahsler, Christian Buchta, Bettina Gruen, and Kurt Hornik. *arules: Mining Association Rules and Frequent Itemsets*. R package version 1.1-6. 2014. URL: <http://CRAN.R-project.org/package=arules>.

Michael Hahsler, Bettina Gruen, and Kurt Hornik. “arules – A Computational Environment for Mining Association Rules and Frequent Item Sets”. In: *Journal of Statistical Software* 14.15 (Oct. 2005), pp. 1–25. ISSN: 1548-7660. URL: <http://www.jstatsoft.org/v14/i15/>.

Marek Hlavac. *stargazer: LaTeX/HTML code and ASCII text for well-formatted regression and summary statistics tables*. R package version 5.1. Harvard University. Cambridge, USA, 2014. URL: <http://CRAN.R-project.org/package=stargazer>.

Beautiful Soup Team. *Beautiful Soup*. Version 4. 2015. URL: <http://www.crummy.com/software/BeautifulSoup/>.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2014. URL: <http://www.R-project.org>.

Hadley Wickham. “Reshaping Data with the reshape Package”. In: *Journal of Statistical Software* 21.12 (2007), pp. 1–20. URL: <http://www.jstatsoft.org/v21/i12/>.

Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN: 9780387981406. URL: <http://had.co.nz/ggplot2/book>.

Hadley Wickham. “The Split-Apply-Combine Strategy for Data Analysis”. In: *Journal of Statistical Software* 40.1 (2011), pp. 1–29. URL: <http://www.jstatsoft.org/v40/i01/>.

Glossary

- **Corsi/Corsi event:** An attempted shot; either a shot on goal, a missed shot, or a blocked shot. Attempts by one's team are **Corsi For** and attempts against one's team are **Corsi Against**. It can be expressed as the percentage of all Corsi events.
- **CF60/CA60:** Corsi events for/against per 60 minutes of play.
- **Fenwick/Fenwick event:** An attempted shot, excluding blocked shots. Only missed shots and shots on goal are counted.
- **NHL:** National Hockey League
- **PBP:** Play-By-Play. A list of all events – such as goals, hits, shots, etc. – during a game.
- **PK:** Penalty kill. A team is on the penalty kill when they have fewer players on the ice as a result of a penalty. Also referred to as playing **shorthanded**.
- **PP:** Powerplay. A team is on the powerplay when they outnumber their opponent's players on the ice because of a penalty.
- **RTSS:** Real Time Scoring System. The NHL's system containing the Play-By-Play and Time-On-Ice.

- **SH:** Shorthanded. See **PK**.
- **TOI:** Time-On-Ice. A table listing the times when players are on the ice.
- **Win percentage (regulation):** The proportion of games in which a team is leading at the end of the third period. Conventionally written as a decimal.