Aalto University
School of Science

On cluster structures of NHL players

Bachelor's thesis

Elmeri Lähevirta

Espoo 30.01.2018

Advisor: D.Sc. Lauri Viitasaari
Supervisor: Asst.Prof. Pauliina Ilmonen

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of bachelor's thesis

Author	Elmeri Lähevirta
Title of thesis	On cluster structures of NHL players
Degree programme	Engineering physics and mathematics
Major	Mathematics and operations research
Code of major	SCI3029
Supervisor	Asst.Prof. Pauliina Ilmonen
Thesis advisor(s)	D.Sc. Lauri Viitasaari
Date	15.01.2018
Number of pages	19
Language	English

Abstract

The statistical analysis of NHL players has had a revolution in the 21st century after hockey teams have realized the possibilities of analysing and forecasting the performance of players based on their statistics. NHL teams have hired analysts, even academic statisticians, to try to find a way to harness the hidden information in statistics to give an advantage for their team. The movie called "Moneyball" (2011), which tells the story of a statistical analyst helping his baseball team to success, has been thought to be the starting point of the statistical revolution.

In this thesis, clustering analysis is performed on the skater and forward statistics of the regular NHL season of 2016-17. The goal of this thesis is to investigate the capability of a clustering algorithm to separate the defenders from the skaters into their own clusters and if any of the resulting clusters can be labelled with identifiable player types. Different clustering methods and distance measures are presented, and the *k*-means++ algorithm with the Euclidean distance measure is chosen for the clustering analysis.

The *k*-means++ algorithm was able to separate the forwards from defenders well. In addition, different player types, such as "enforcers" and "high-scoring/offensive-oriented players", were also identified from the resulting clusters. For future research, different clustering algorithms and their performance should be investigated. In addition, a cluster analysis of goaltenders would be an interesting topic to investigate.

Keywords NHL, ice-hockey, statistical analysis, cluster analysis, k-means

Aalto-yliopisto
Perustieteiden
korkeakoulu

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi

Tekniikan kandidaatintyön tiivistelmä

Tekijä	Elmeri Lähevirta
Työn nimi	On cluster structures of NHL players
Koulutusohjelma	Teknillinen fysiikka ja matematiikka
Pääaine	Matematiikka ja systeemitieteet
Pääaineen koodi	SCI3029
Vastuuopettaja	Apulaisprofessori Pauliina Ilmonen
Työn ohjaaja(t)	TkT. Lauri Viitasaari
Päivämäärä	15.01.2018
Sivumäärä	19
Kieli	Englanti

Tiivistelmä

NHL-pelaajien tilastollinen analyysi on kokenut vallankumouksen 2010-luvulla. Jääkiekkojoukkueet ovat ymmärtäneet tilastoihin perustuvan pelaajien suorituskyvyn analysoimisen ja ennustamisen mahdollisuudet. NHL-joukkueet ovat palkanneet analyytikkoja, jopa akateemisia tilastotieteilijöitä, yrittämään löytää keinot valjastaa tilastoissa piilevän informaation. Elokuva Moneyball (2011) kertoo tarinan, kuinka tilastoanalyytikko auttaa pienen budjetin baseball-joukkueen menestymään. Sen ajatellaan olevan urheilun tilastollisen vallankumouksen aloituspiste.

Tässä kandidaatintyössä tutkittiin klusterointianalyysin menetelmin NHL-hyökkääjien ja -puolustajien kauden 2016-17 runkosarjan tilastoja. Työn tavoitteena oli tutkia, kuinka hyvin klusterointialgoritmit pystyvät erottelemaan hyökkääjät ja puolustajat toisistaan ja pystytäänkö tuloksena syntyviä ryhmiä luokittelemaan tunnettujen pelaajatyyppien mukaan. Työssä esitellään erilaisia klusterointimetodeja ja etäisyysmittoja, joista päädyttiin käyttämään *k*-means++-algoritmia euklidisella etäisyysmitalla. Klusterointi suoritettiin neljällä eri klusterointiryhmien määrällä.

Työn tuloksien perusteella *k*-means++-algoritmi pystyi erottamaan hyökkääjät puolustajista erittäin hyvin ja erilaisia pelaajatyyppejä, kuten ”voimahyökkääjät” ja ”pisteiden tekijät/hyökkäysorientoituneet pelaajat”, pystyttiin tunnistamaan tuloksena syntyneistä klustereista. Tulevaisuuden jatkotutkimuksen aiheita on tutkia eri klusterointimenetelmien ja etäisyysmittojen tehokkuutta pelaajien ryhmittelyssä sekä suorittaa klusterointianalyysi myös maalivahdeille.

Avainsanat NHL, jääkiekko, tilastollinen analyysi, klusterointianalyysi, *k*-means

Contents

1 Introduction

The National Hockey League (NHL) is a men's professional ice-hockey league in North America widely recognized as the world's premier hockey league. The NHL was established in 1917 and only six franchises at most played in the league until 1967, when six new teams joined. After that the league has greatly expanded and the NHL currently comprises 31 franchises of which seven are in Canada and 24 in the United States. The teams in the NHL compete for the prestigious Stanley Cup trophy, which is awarded to the winner of the playoffs each season. [1]

The statistical analysis of NHL players has had a revolution in the 21st century: hockey teams have realized the possibilities of analyzing and forecasting the performance of players based on their statistics. NHL teams have hired analysts, even academic statisticians, to try to find a way to harness the hidden information in statistics to give an advantage for their team [2]. The movie called "Moneyball" (2011) [3] has been thought to be the starting point of the statistical revolution. The movie is based on a book called "Moneyball: The Art of Winning an Unfair Game" [4] which tells the true story of a statistical analyst helping a small budget baseball team to success. Scientific statistical research of sports is not a new thing (see for example articles [5, 6, 7]), but now the sport teams have made it part of their processes and organizations.

Clustering, one tool of statistical analysis, is the method of grouping objects into subgroups such that some similarity is maximized within subgroups and minimized between subgroups. The interest of clustering is the resulting clusters: the analysis of clustering is the task of trying to identify and label them. Clustering is usually divided into two subgroups: hierarchical and partitioning algorithms. For more about clustering and clustering algorithms, see [8, 9, 10, 11].

In this thesis, a clustering analysis is performed for forward and defender statistics of the NHL regular season 2016-17 with k-means clustering algorithm. The goal of this thesis is to investigate the capability of the k-means clustering algorithm to separate the defenders from the skaters into their own clusters and if any of the resulting clusters can be labeled with identifiable player types. The clustering analysis is performed with cluster amounts k of 2, 3, 4 and 5. Goaltenders are not a part of the clustering analysis in this thesis.

The statistics used for the clustering analysis are presented and analyzed by their location and scatter measures in Chapter 2. Chapter 3 reviews clustering, including different distance measures and clustering methods. The standardization of the data, clustering analysis and the results of the clustering are presented in Chapter 4.

2 Statistics of the NHL players

The statistics used in this thesis are those of the NHL players of the regular season 2016-17 and they contain the statistics of all forwards and defenders. The statistics are originally from the homepage of NHL obtained by hockey analytic Robert Vollman from the Hockey Abstract [12].

The statistics contained originally data of 589 forwards and 299 defenders. To avoid outliers, players with 9 or less games played were left out. After excluding these players, 486 forwards and 250 defenders remained.

2.1 Forwards

There are three different forward positions in the hockey lineup: left winger (LW), center (C) and right winger (RW). Centers take primarily always the faceoffs and are often more defense-oriented as wingers.

The variables chosen for forwards are:

1. Goals Average (GA). The number of goals made by the player on average per 60 minutes of on-ice time.
2. Assists Average (AA). The number of goal assists given by the player on average per 60 minutes of on-ice time.
3. Shots Average (SA). The number of shots on goal by the player on average per 60 minutes of on-ice time.
4. Hits Average (HA). The number of hits given by the player on average per 60 minutes of on-ice time.
5. Blocks Average (BLKA). The number of shots blocked by the player on average per 60 minutes of on-ice-time.
6. Time On Ice per game (TOI/Game). The average time in minutes that the player has been on the ice per played game.
7. Corsi rating average per 60 minutes of on-ice time (Corsi). Corsi rating is the difference of the team's shot attempts for minus the shot attempts against when the player was on the ice. Only the even-strength plays (5 vs. 5) are calculated in the corsi rating.

Table 1: Location and scatter measures of the forwards' statistics of the regular season 2016-17.

	Min	Max	Mean	Median	SD	Skewness	Kurtosis
GA	0.00	1.78	0.69	0.66	0.37	0.40	0.09
AA	0.00	2.94	0.92	0.87	0.49	0.44	0.15
TOI/GP	5.53	21.46	14.22	14.02	3.34	−0.06	−0.81
Corsi	−37.28	30.92	1.37	2.75	14.33	−0.29	−0.68
SA	2.30	12.47	6.86	6.77	1.95	0.36	−0.08
HA	0.25	24.67	5.53	4.27	4.24	1.36	1.78
BLKA	0.00	5.79	1.91	1.77	0.93	0.95	1.44

The location and scatter measures of the chosen variables of forwards' statistics of regular season 2016-17 are presented in Table 1. From the table, we can see that the goals average GA has values between 0 and 1.78 and assist average AA between 0 and 2.94. The AA has a lot higher maximum value than GA, but the means and medians are closer with the median of GA being 0.66 and the median of AA being 0.87. It makes sense that the assist average is higher than the goal average as there is only one goal credited for one goal made, but maximum of two assists can be credited per goal: primary and secondary assists. The average total on-ice time per game played gets values between 5.53 to 21.46 min per game with the median being 14.02 min and average 14.22 min. That means that on average a forward is one fourth of a regular game time on ice and a little over one third at maximum, so the other variables which are calculated as an average for 60 minutes on ice time should not be confused anyway with a per game played average.

From Table 1, we can see that the corsi rating average for forwards is between -37.28 and 30.92 with the mean being 1.37 and median 2.75. This means that when the player with the best corsi rating average is on the ice, his team creates 30.92 more shot attempts for than is created against them per 60 minutes of on-ice time of that player. The mean and median of the corsi rating are both a little bit above zero, because the players with less than 10 games played were excluded and their total corsi rating average was negative. The corsi rating mean would otherwise be exactly zero if none of the players were excluded. The standard deviation of the corsi rating is quite high 14.33 which means that the corsi rating averages are spread out from the mean. The minimum shot average is 2.30, meaning that every forward with 10 or more games played has been able to create shots on the opposing team's goal, although this minimum amount is very little: with the average on-ice time it is about 1 shot for every two games played. The maximum shot average has been 12.47, so the player with the best shot average has created a shot for every 4.8 minutes of his on-ice time. The shot average mean is 6.86 and median 6.77, meaning that the average player creates a little over one shot for every 10 minutes of his on-ice time.

The hit average of forwards varies between 0.25 to 24.67 so the range of the hit average is big. The mean of the hit average is 5.53 and median 4.27 with the standard deviation being 4.24, meaning that the maximum value of 24.67 seems to be quite an extreme data point and that most of the points are closer to the mean. The block average has values between 0 and 5.79 with the mean being 1.91 and the median 1.77, so forwards do not block shots much as the best blocker blocks a little under one shot per 10 minutes of on-ice time and the average forward blocks only one shot per 30 minutes of on-ice time.

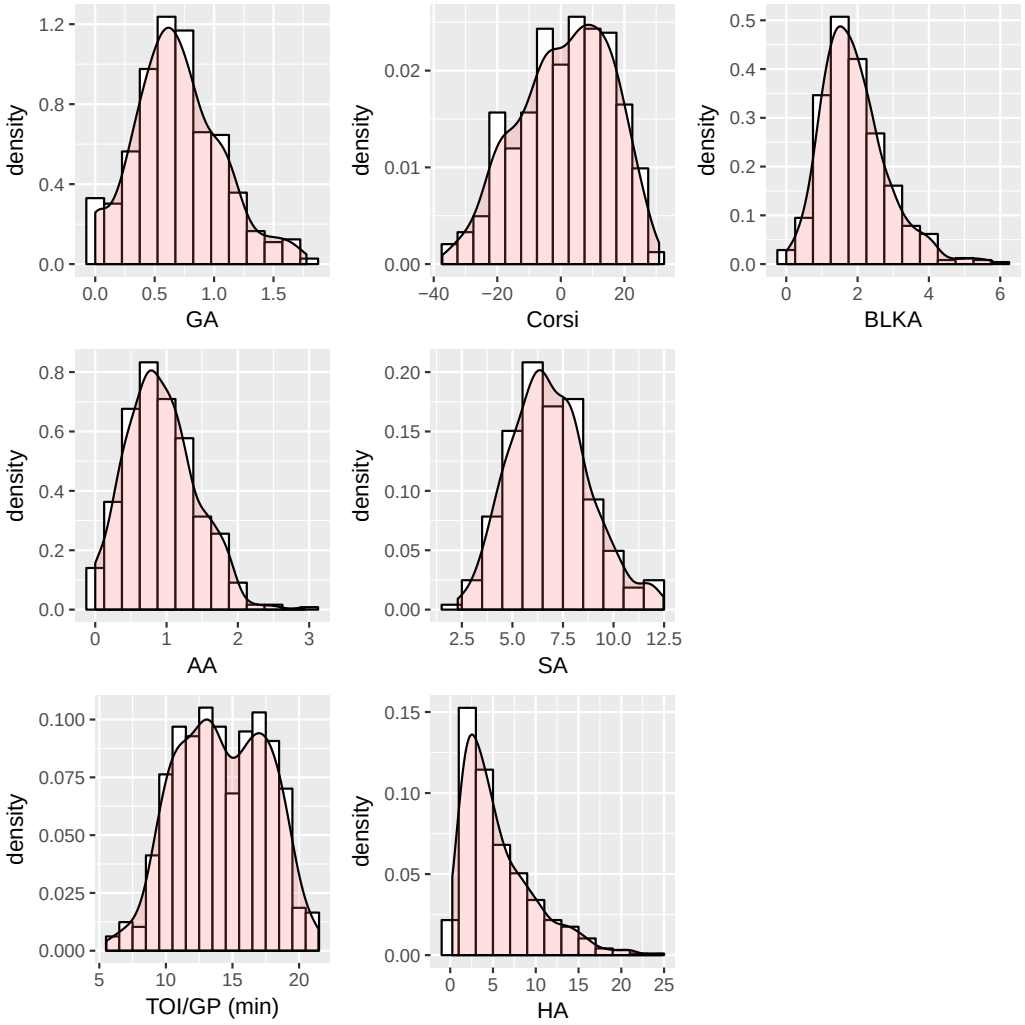

Figure 1: Histograms with kernel density curves of the forward statistics.

The histograms with kernel density curves for all the different variables of the forward statistic data are presented in Figure 1. We can see from these histograms that the distributions of HA and BLKA are clearly skewed to the right. This is also backed up by the skewness and kurtosis values presented in Table 1: the skewness of HA is 1.36 and for BLKA it is 0.95. From the histograms we can also see that the distributions of GA, AA and SA are

slightly skewed to the right, which is backed up by the positive skewness values of 0.40, 0.44 and 0.36 respectively. HA and BLKA have the biggest kurtosis values 1.78 and 1.44, and their histograms have also very strong and sharp peaks. The histograms of Corsi and TOI/GP have clearly the lowest and widest peaks and they also have the lowest kurtosis values of -0.68 and -0.81. The other variables have their kurtosis values near zero.

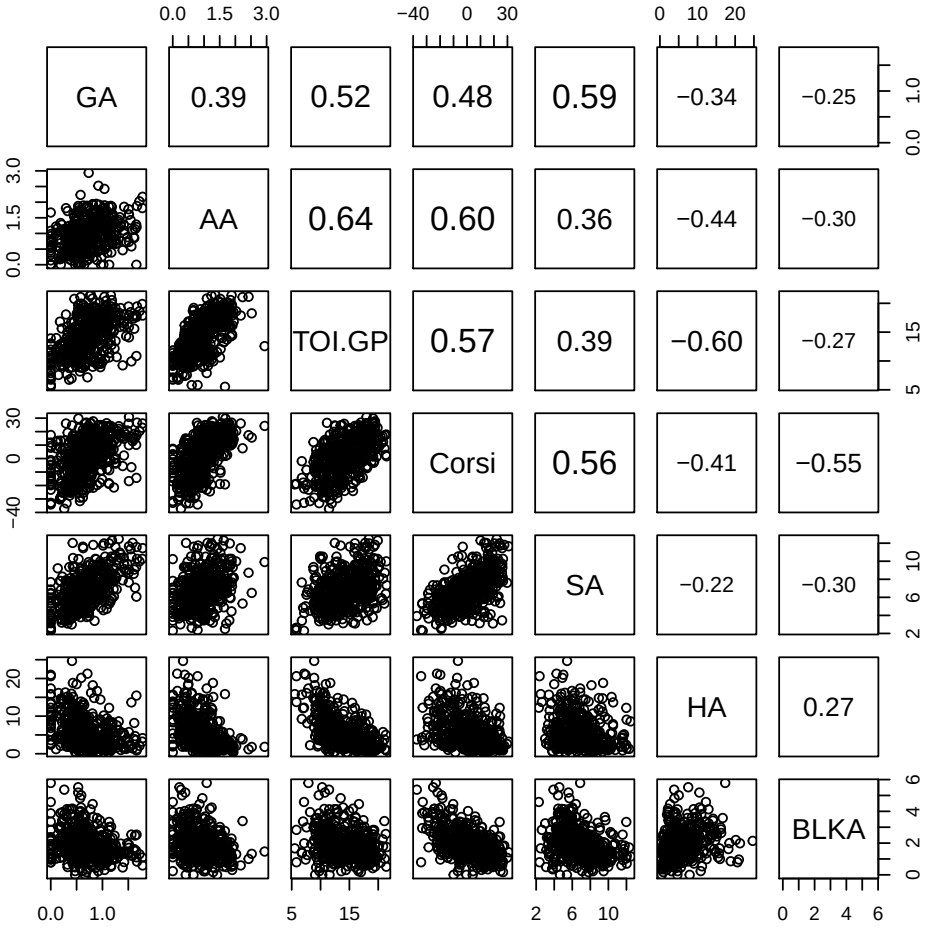

Figure 2: Scatter plots and correlation coefficients of the forward statistics.

The scatter plots and correlation coefficients of the forward statistic data is presented in Figure 2. We can see from the scatter plots and the values of the correlation coefficients, that the variables have quite a lot of correlation between each other. Corsi rating average correlates the most overall with all the other variables and it correlates the most with the assist average with correlation coefficient of 0.60. This could be explained by the fact that players with higher assist average are also able to make the play more in the opponent's end and thus generate also more shot attempts than are created against. Other good reason may just be that the players with greater assist

average play on a better line with good line mates and that better line is able to create the play more in the opponent's end and create more goals which results also to better assist averages. HA and BLKA correlates negatively with all the other variables except with each other. The BLKA correlates the most with the corsi rating average with the correlation coefficient of -0.55 and one reason for it is that blocked shots are also calculated as shot attempts against in the corsi rating average. HA correlates the most with TOI/GP with correlation coefficient -0.60 and that is probably because the players who hit a lot play often in the third or fourth line in roles that do not get that much on-ice time. The biggest single correlation is between the variables AA and TOI/GP with the correlation coefficient of 0.64. SA and GA correlate with correlation coefficient of 0.59, so the more a player shoots on average, the more he also makes goals on average.

2.2 Defenders

In hockey, a line has two defenders: a left defender (LD) and a right defender (RD). Defenders play, as their name suggests, more defense-oriented game. That is why it is expected that the goals, assists and shots averages are lower for defenders than for forwards. The defenders play mostly in front of their own goal and try to protect it anyway possible in the defense zone, so the blocks average is probably higher for defenders than for forwards. The variables chosen for defenders are the same as selected for forwards in the previous section so that defender and forward statistics are easier to compare with each other.

Table 2: Location and scatter measures of the defenders' statistics of the regular season 2016-17.

	Min	**Max**	**Mean**	**Median**	**SD**	**Skewness**	**Kurtosis**
GA	0.00	0.85	0.17	0.19	0.14	0.91	1.33
AA	0.00	1.74	0.61	0.68	0.35	0.60	0.17
TOI/GP	9.63	27.45	18.60	18.68	3.69	−0.02	−0.58
Corsi	−30.63	29.61	−2.73	−3.49	10.63	0.32	−0.14
SA	1.93	9.45	4.20	4.31	1.36	0.89	1.05
HA	0.00	22.27	3.88	4.60	2.97	1.56	4.75
BLKA	1.03	9.06	4.31	4.41	1.35	0.55	0.73

The location and scatter measures of the defenders' statistics are presented in the Table 2. When comparing these to the same estimates of the forwards statistics presented in Table 1, we can see that defenders have lower mean and max values of GA, AA, Corsi, SA and HA. On the other hand, the mean and max values of TOI/GP and BLKA are higher for defenders than for forwards. As was expected, the GA, AA and SA are all lower for defenders as they play more defense-oriented game and the BLKA mean is higher for

defenders for the same reason. The HA mean of forwards 5.53 is higher than the mean of defenders 3.88, but the median of defenders 4.60 is higher than the one of forwards 4.27 suggesting that the hits are a little bit more evenly spread out for defenders. Usually the playing roster of a NHL team consists of 12 forwards and 6 defenders and this explains why the TOI/GP mean is higher for defenders.

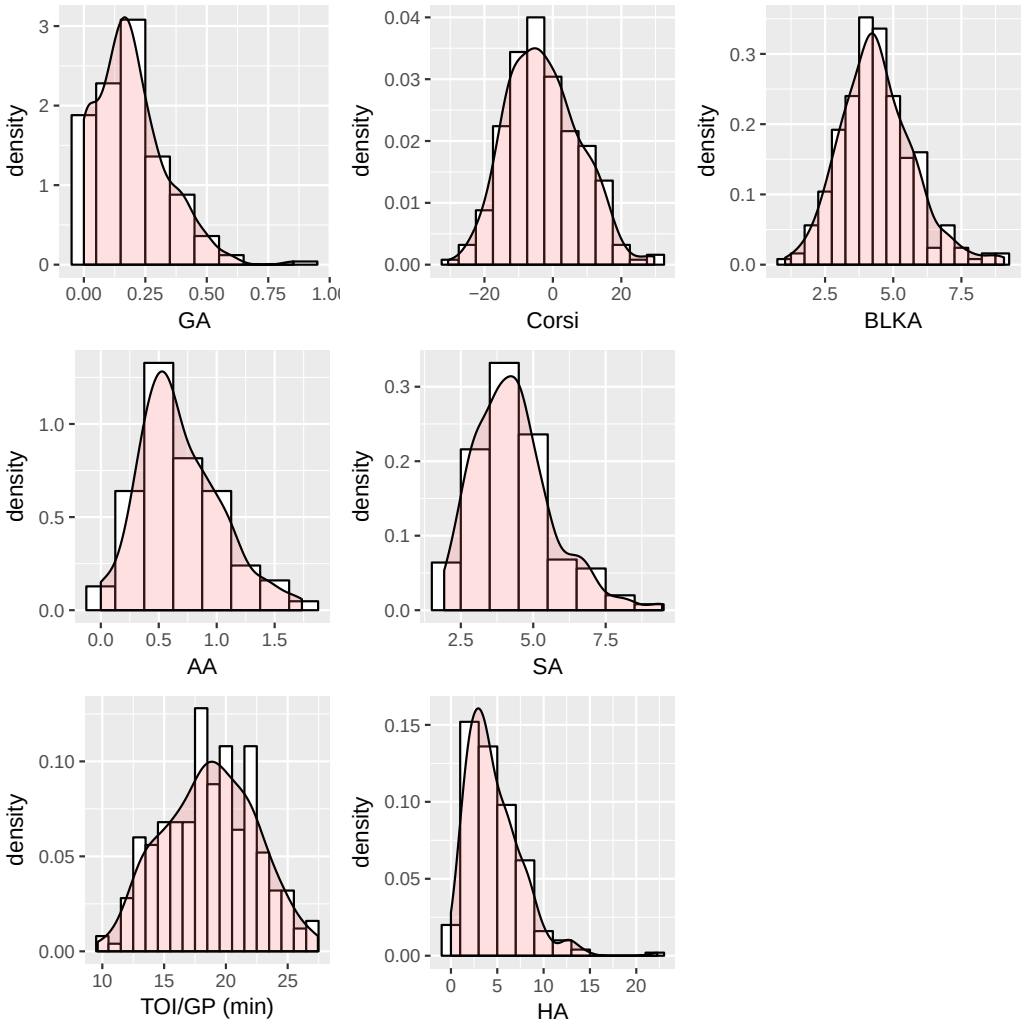

Figure 3: Histograms with kernel density curves of the defender statistics.

The histograms and kernel density curves of the variables of the defenders' statistics are plotted in Figure 3. When comparing these with the histograms of the forward statistics presented in Figure 1 and taking also into account the location and scatter measures of the forward and defender statistics in Tables 1 and 2, we can see that the overall shape of all histograms of the defender statistics are very similar with the ones of the forward statistics. Based on the histograms, the distributions of GA, AA, SA, HA and BLKA are all skewed to the right also for defenders. The skewness values are all also greater for these variables for defender statistics, except for the BLKA, for

which it is smaller for defenders than for forwards. The histogram of Corsi has a right tail for defenders with skewness value of 0.32, when for forwards it has left tail with skewness value of -0.29, so the Corsi histograms are leaning to different directions for forwards and defenders. HA has a very high kurtosis value of 4.75 for defender statistics, so the peak of the HA histogram is a lot sharper when compared to the one of the forward statistics, which had a kurtosis value of 1.78. Based on the histograms of TOI/GP, the one of the defenders' statistics seem to have a clearer peak that the one of the forwards' statistics, but based on the kurtosis values the difference is small (-0.58 for defenders and -0.81 for forwards).

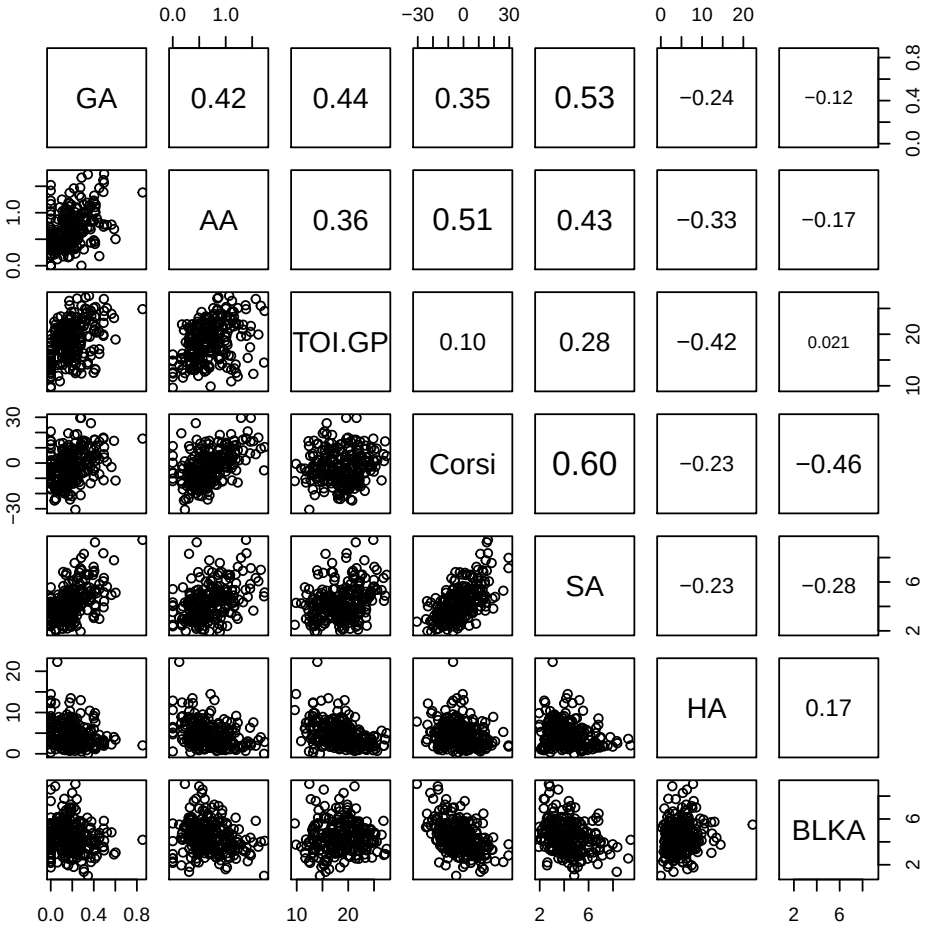

Figure 4: Scatter plots and correlation coefficients of the defender statistics.

The scatter plots and correlation coefficients of all the variables of the defender statistics are plotted in Figure 4. When comparing the correlation coefficients to those of the forward statistics presented in Figure 2, we can see that all correlations are smaller for the defender statistics (the absolute value of the correlation coefficient is smaller) except for correlation coeffi-

cients between the pairs GA-AA, SA-AA, SA-Corsi and SA-HA for which the correlation coefficients are a bit greater for the defender statistics. One large difference between the correlation coefficients of the defenders and forwards data is the correlation between Corsi and TOI/GP which is 0.57 for forwards and only 0.10 for defenders. This means that for defenders greater corsi rating average is not as strongly related to greater total on-ice time per game played. Moreover, the TOI/GP is not correlating as strongly with any of the variables for defenders than it is for forwards.

3 Clustering

Clustering is the task of grouping a set of objects into subgroups (called clusters) in a way that the similarity is maximized within the subgroups and minimized between the subgroups. The goal in clustering is to divide a heterogeneous sample into homogeneous groups and to find suitable labels for the groups. [13, Chapter 14]

Clustering is an unsupervised learning method of which the opposite is a supervised learning method. In supervised learning, one is usually provided with a labeled (pre-classified) objects and the problem is to label a new, yet unlabeled, object. In unsupervised learning, the problem is to group a given group of objects into meaningful subgroups, clusters. Labels are in a sense associated with clusters also, but these category labels are data driven meaning that they are obtained solely from the data [14]. Cluster analysis is usually an iterative process of trials and repetitions and there are no universal and effective criteria to help to select the features and clustering schemes [15].

Clustering has been widely used in a variety of fields ranging from economics (marketing, business), social sciences (sociology, psychology), earth sciences (geography, geology), life and medical sciences (genetics, biology) to computer sciences (web mining, image segmentation) and engineering (machine learning, artificial intelligence) [15].

In order to do the clustering, some sort of measure of similarity is required. There are two main types of approaches: one can apply a distance measure or one can apply a similarity measure.

3.1 Distance measures

There are many different measures that can be applied in measuring the distance between objects and clusters. Let us denote the distance between objects x_i and x_j of a set S as $d(x_i, x_j)$. All valid distance measures must satisfy the following four properties for all $x_i, x_j, x_k \in S$:

1. Non-negativity: $d(x_i, x_j) \geq 0$

2. Symmetry: $d(x_i, x_j) = d(x_j, x_i)$
3. Identity of indiscernibles: $d(x_i, x_j) = 0 \Leftrightarrow x_i = x_j$
4. Triangle inequality: $d(x_i, x_k) \leq d(x_i, x_j) + d(x_j, x_k)$

A distance measure satisfying these properties is known as metric [16, Chapter 2].

The most popular distance measure for numeric data is the Euclidean distance [16, Chapter 2]. Let x_i and x_j be p-dimensional instances, $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$ and $x_j = (x_{j1}, x_{j2}, ..., x_{jp})$. The Euclidean distance between objects x_i and x_j is defined as

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ip} - x_{jp})^2} \ .$$

There are many other distance measures including, for example, Manhattan and Chebychev metrics and there are also different distance measures for binary, nominal and ordinal attributes [16, Chapter 2]. Distance measures are not the only measures to measure the similarity between two objects. For example, the cosine measure is another similarity measure which measures the angle between two data vectors by calculating the normalized inner product between the two. The Euclidean metric is used in this thesis as it is the most common and the attributes of the data used in the analysis are numeric.

The distance between clusters can be defined in many different ways. For example, the average distance measures how far the cluster means are from each other. Other commonly used distances between clusters are the maximum distance and the minimum distance.

Many evaluation criteria measures exist, which try to measure the quality of the clustering. In this thesis, the quality of the clustering results is evaluated by the reasonableness of the resulting clusters. One of the most important factor being, how well the clustering algorithm was able to separate the forwards from defenders.

3.2 Different clustering methods

There exist many different clustering methods that can be divided into two main groups: hierarchical and partitioning methods [17, 18]. Hierarchical methods construct the clusters by iteratively partitioning the data object in a bottom-up or top-down fashion.

Hierarchical clustering methods can be subdivided into agglomerative clustering and divisive clustering methods. In agglomerative clustering, each object is a cluster of its own. Clusters are then successively merged until the designed cluster structure is achieved. In divisive clustering, all data objects are in one cluster at the beginning and then the cluster is divided into

sub-clusters, which are successively divided into their own sub-clusters. This iterative process is repeated until the desired cluster structure is obtained. The merging or division is made according to some similarity measure, for example, the minimum, maximum or average distance between the clusters.

The result of hierarchical clustering methods can be visually presented by a dendrogram. Dendrogram represents the nested grouping of data objects and the similarity levels at which the clusters change. The final clustering of the data objects is obtained by cutting the dendrogram at the desired similarity level.

The other main group of clustering methods is the partitioning methods. Partitioning methods work by relocating the data objects from one cluster to another, starting from some initial partitioning. These methods usually require the number of clusters to be predefined by the user. An exhaustive enumeration process of all possible partitions is required to achieve the global optimality in partitioning clustering and as this is not feasible in practice, different heuristics are used in the form of iterative optimization.

Error minimization algorithms are the most frequently used and intuitive partitioning methods. These algorithms try to find a clustering structure that minimizes a certain error criterion which applies the used distance measure. Sum of squared error (SSE) is the most well-known error criterion and it measures the total squared Euclidean distances of data objects to their representative values. SSE can be globally optimized by exhaustively enumerating all partitions which is very time-consuming. However, the most common alternative is to find an approximate solution by using heuristics, even though it may not give the global optimum [13, Chapter 14].

One of the simplest and commonly used error minimization algorithm, which employs the squared error criterion is the k-means algorithm [16, Chapter 10]. This algorithm partitions the data into k clusters $(C_1, C_2, ..., C_k)$, which are represented by their centers or means. The number of clusters k is predefined by the user. The center of each cluster is calculated as the mean of all the data objects belonging to that cluster. The algorithm starts with an initial set of k cluster centers, chosen uniformly at random or with some heuristic procedure. Then in each iteration, every data object is assigned to the nearest cluster center according to the Euclidean distance and after each iteration, the cluster centers are re-calculated.

K-means algorithm is very sensitive for the initial choice of the cluster centers and it can make the difference between local and global optimum. For this reason, an algorithm named k-means++ has been developed for choosing the initial cluster centers [19]. In k-means algorithm the initial starting centers are uniformly chosen at random from the set of all data objects S. In k-means++ algorithm, the first cluster center is chosen the same way, uniformly at random from S. The difference is that the second and each subsequent cluster center is chosen from the remaining data objects with

probability proportional to its squared distance from the object's nearest existing cluster center. This means that those data objects are selected as new cluster centers with higher probability, which are further away from the existing cluster centers. There are also many other initialization methods for the k-means algorithm which are not discussed in this thesis. For other initialization methods, see [20].

Other distance measures can also be used with the k-means algorithm than the Euclidean distance. For example, the Minkowski distance or Manhattan distance are two other common metrics [21]. There are also other variants of the k-means algorithm, for example, the k-medoids algorithm, in which the cluster centers are always the most centric object of the cluster, not the mean, which may not be a part of the cluster. In this thesis, the k-means algorithm with the Euclidean distance is used because of its popularity. In addition, the k-means++ initialization of the cluster centers is used to ensure stability of the algorithm.

4 Analysis of the statistics

The measurement units used can affect the results of the clustering analysis by giving larger weight for variables with larger variability. The standardization of the data attempts to give an equal weight for all variables in the cluster analysis [16, Chapter 2]. For this reason, the data used in this thesis is standardized before the clustering analysis to avoid the dependence of the choice of the measurement units.

The data is standardized to have sample mean 0 and standard deviation 1 with the following formula:

$$x^* = \frac{x - \mu(x)}{\sigma(x)} \,,$$

in which $\mu(x)$ is the mean of variable x and $\sigma(x)$ is the standard deviation of x.

4.1 Forwards and defenders

The k-means clustering was performed for combined forward and defender statistics with k having values of 2, 3, 4 and 5 to see how well the k-means can differentiate the defenders from forwards with different cluster amounts.

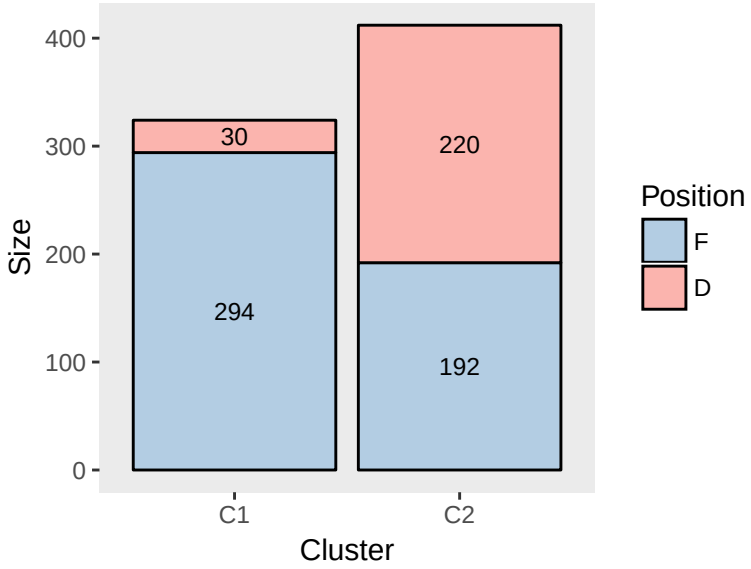

Figure 5: Result of the k-means clustering with $k = 2$.

The result of k-means clustering with $k = 2$ is plotted in Figure 5. We can see from the figure that the clustering algorithm has been able to produce a cluster (C1) which contains almost only forwards. On the other hand, in cluster C2 there are almost as many forwards as defenders. The forwards in cluster C1 are very offensive-oriented with high shot, goal and assist averages. The defenders in cluster C1 are also more offensive-oriented: defenders that shoot more, get more power-play time and create more goals and assists, including defenders like Erik Karlsson, Victor Hedman, Kris Letang and Roman Josi. The forwards in cluster C2 are more defense-oriented: players that do not create that much goals and assists and which play more penalty kill and in the lower 3rd or 4th lines, including forwards like Leo Komarov, Nick Bonino, Ryan Reaves and Antoine Vermette. Cluster C2 contains all the defenders except the most offensive-oriented ones, which are in cluster C1. Cluster C1 can be labeled as ”offensive-oriented players” and cluster C2 can be labeled as ”defense-oriented players”.

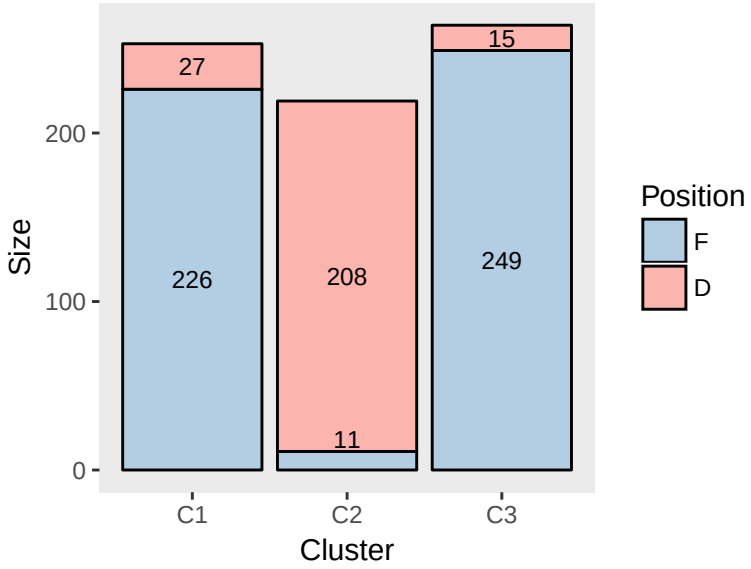

Figure 6: Result of the k-means clustering with $k = 3$.

The result of k-means clustering with $k = 3$ is plotted in Figure 6. We can see from the figure that the clustering algorithm has performed well in separating the forwards from the defenders. Clusters C1 and C3 contain mostly forwards and cluster C2 contains mostly defenders. The difference between clusters C1 and C3 is that the C3 is clearly more offensive-oriented containing players which have high shot, goal, assist and corsi averages and cluster C1 contains more defense-oriented forwards. Cluster C3 is similar with the cluster C1 of the clusters with $k = 2$, containing mainly forwards with high goal, assist, shot and corsi averages. The defenders in cluster C3 are ones with high goal, assist, shot and corsi averages and they have quite low hit and block averages including defenders like Brent Burns, Shayne Gostisbehere and Kevin Shattenkirk. Clustering algorithm has been stricter about the high shot average, high corsi average, low hit average, and low block average. For example, in comparison with cluster C1 with $k = 2$, defenders with high goal and assist averages have been excluded from cluster C3 if they have low shot and corsi averages or high hit and block averages. Defenders like these are, for example, Erik Karlsson, Victor Hedman and Dustin Bufyglien, which are all now in cluster C2. Cluster C2 comprises mostly defenders containing 208 defenders and only 11 forwards. The forwards in cluster C2 are the ones with higher than mean block average and lower than mean hit average, including players like Nick Bonino, Marcus Kruger and Antoine Vermette. Cluster C1 can be labeled as ”enforcers/power forwards”, cluster C2 can be labeled as ”defenders” and cluster C3 can be labeled as ”high-scoring/offensive forwards”.

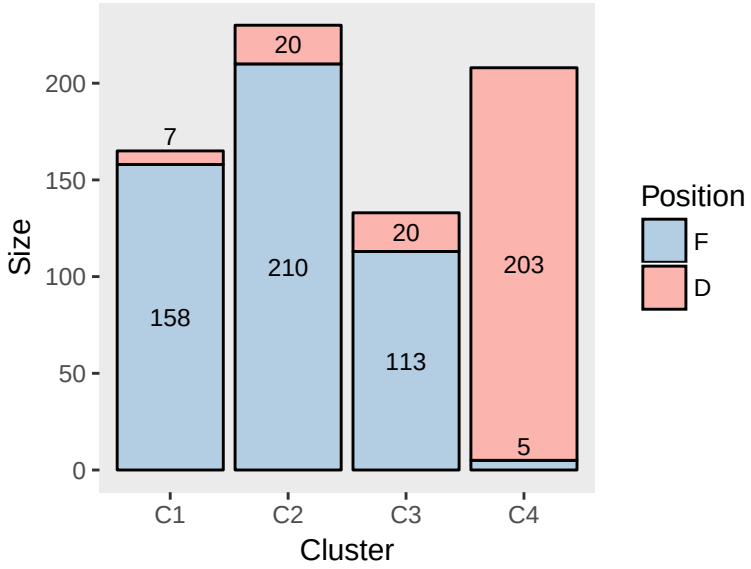

Figure 7: Result of the k-means clustering with $k = 4$.

The result of k-means clustering with $k = 4$ has been plotted in Figure 7. We can see from the figure that the clustering algorithm has again differentiated very well the defenders from the forwards. Clusters C1, C2 and C3 contain mostly forwards and cluster C4 contains mostly defenders. The five forwards in cluster C4 are all also in cluster C2 with $k = 3$. All of them has negative corsi rating averages and all except one have their shot and hit averages below the forward means. On the other hand, all of them have block averages above the forward mean, meaning they are in many ways defender-like what comes to their statistics. Cluster C4 can be labeled as the ”defenders” cluster. The players in cluster C3 are ones with low goal and assist averages and they also have lower than mean total on-ice time per game played. Most of them have also very low corsi rating average and their shot averages are also lower than the forward mean. Players in cluster C3 tend to have very high hit averages and above average block averages as well. So the players in cluster C3 seem to be players that play less than average, more defensive-oriented game and they play with high hit averages. Cluster C3 can be labeled as ”enforcers”. Cluster C1 is the ”high-scoring/offensive-oriented player” cluster with players that have high assist, goal, shot and corsi rating averages and it contains almost the same players as cluster C3 with $k = 3$. Finally, players in cluster C2 could be labeled as ”average forwards”. They do not have high enough goal, assist or corsi rating averages to be included in cluster C1 or high enough block and hit averages to be included in cluster C3. The cluster means of all variables in cluster C2 are also very close to the means calculated from the whole forward data.

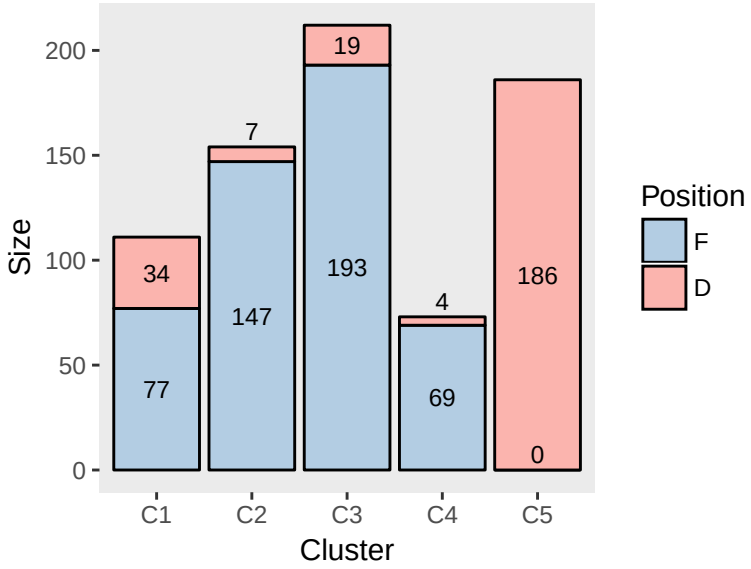

Figure 8: Result of the k-means clustering with $k = 5$.

The result of k-means clustering with $k = 5$ is presented in Figure 8. We can see from the figure that the clustering algorithm has performed well in separating the forwards from defenders for every cluster except cluster C1, which contains 34 defenders and 77 forwards.

Cluster C2 contains the ”high-scoring/offensive-oriented” players and cluster C3 is the ”average forward” cluster. Cluster C4 is the ”enforcers” cluster; it contains the players that do not get much on-ice time (cluster average 10.11 minutes) but have very high hit averages (cluster average 13.62). The cluster is more clearly a group of players with high hit averages when compared to the enforcers cluster with $k = 4$, as the lowest hit average in cluster C4 is now 8.11 when it was only 1.38 with $k = 4$. Also, the cluster hit average was lower (10.56) for the enforcers cluster with $k = 4$. Cluster C5 is a pure ”defenders” cluster as it contains now only defenders.

Cluster C1 is the trickiest to give any label. Players in this cluster have very low corsi ratings (cluster average -14.25), their shot cluster average 4.60 is below the forward mean 6.86 and it is only a little over the defender mean 4.20. Players in cluster C1 have their total on-ice time average below both the forward and defender means with the cluster average value of 12.63 minutes. The cluster average of the hit average 6.21 is above both the forward and defender means (5.53 and 3.88 respectively) and the cluster average of the block average is 3.20, which is higher than the forward mean 1.9, but lower than the defender mean 4.31. The cluster average of goal average is 0.31 and of assist average it is 0.49, which both are low when comparing to the forward mean and about the same size as the defender means. The cluster average of shot average is 4.60, which is lower than the forward mean 6.86 and a little bit higher than the defender mean 4.20. Based on this, the forwards in this cluster are clearly more defense-oriented players with goal,

assist and shot averages being lower than the forward mean and the block average being higher. The on-ice time cluster average is lower and the corsi rating cluster average is lower than the forward mean indicating that the forwards in this cluster play in lower 3rd or 4th lines with a defensive role. The defenders in this cluster have lower on-ice average and block average than the defender average and their hit average is, on the other hand, higher than the defender mean and that is probably the reason why they are not in the "defenders" cluster C5. Furthermore, their hit averages are not high enough to be included in the "enforcers" cluster (C4) and their goal and assist averages are not high enough that they would be included in the high-scoring cluster (C2).

5 Conclusions

In this thesis, a clustering analysis was performed on NHL players' statistics of the regular season 2016-17 to investigate how well a clustering algorithm is able to separate the forwards from the defenders and to assess if other player types can be identified. Different clustering methods and distance measures were presented. The k-means++ clustering algorithm with the Euclidean distance measure was chosen for the analysis based on its popularity and suitability for large datasets. Players with less than 10 games played were excluded to avoid outliers and the data was standardized before the clustering. The k-means clustering algorithm requires the number of cluster k to be fixed before running the algorithm. In this thesis, $k = 2$, 3, 4 and 5 were considered.

Based on the results of the clustering analysis, it can be concluded that the k-means clustering algorithm with the Euclidean distance measure works well for the NHL players' statistics. For $k = 3$, 4 and 5, the clustering algorithm was able to separate the forwards from the defenders into their own clusters. In addition, the algorithm was able to separate different identifiable player types such as "enforcers" and "high-scoring/offensive-oriented players". Moreover, for $k = 2$, the clustering algorithm separated the players into two reasonable clusters of which the first contained almost only forwards. Overall, the clustering algorithm performed surprisingly well and significantly better than was expected.

In this thesis, only the cluster structures of skaters (forwards and defenders) were analyzed. The clustering analysis of goaltenders would also be an interesting research topic. In addition, the k-means clustering analysis for forwards and defenders could be performed with larger values of k to investigate if any other identifiable player types can be found. It would also be interesting to study how well other types of clustering algorithms or distance measures would perform on NHL players' statistics.

References

[1] Marsh J., "National Hockey League", *The Canadian Encyclopedia* (electronic), (2012) [Retrieved 18.1.2018]. Available at: http://www.thecanadianencyclopedia.ca/en/article/national-hockey-league/

[2] Swartz T.B., Hockey analytics, *Wiley StatsRef: Statistics reference online, John Wiley & Sons, Ltd*, (2014).

[3] Dir. Miller B., Perf. Hill J., Pitt B., Moneyball (film), *Columbia Pictures*, (2011).

[4] Lewis M., Moneyball: The art of winning an unfair game, *New York: W.W. Norton*, (2003).

[5] Goldie P.A., McKay G.D., Oakes B.W., Payne W.R., Ankle injuries in basketball: Injury rate and risk factors, *Br J Sports Med. 35(2), p. 103-108*, (2001).

[6] Albright S.C., A statistical analysis of hitting streaks in baseball, *Journal of the American Statistical Association 88(424), p. 1175-1183*, (1993).

[7] Feltz D.L., Lirgg C.D., Perceived team and player efficacy in hockey, *J Appl Psychol. 83(4), p. 557-564*, (1998).

[8] Kaufman L., Rousseeuw P.J., Finding groups in data: Introduction to cluster analysis, *Wiley*, (1990).

[9] Duin R., Jain A.K., Mao J., Statistical pattern recognition: A review, *IEEE Trans. Pattern Anal. Mach. Intell. 22, p. 4-37*, (2000).

[10] McQueen J., Some methods for classification and analysis of multivariate observations, *Proc. 5th Berkeley Symp. Math. Statist.Prob. 1, p. 281-297*, (1967).

[11] Larose D.T., Discovering knowledge in data: An introduction to data mining, *Hoboken: Wiley*, (2005).

[12] Vollman R., NHL 2016-17 Player Data, *Hockey Abstract* (electronic) [Retrieved 6.12.2017]. Available at: http://www.hockeyabstract.com/testimonials/nhl2016-17playerdata

[13] Maimon O., Rokach L., Data Mining and Knowledge Discovery Handbook. 2nd ed., *New York; London: Springer*, (2010).

[14] Flynn P.J., Jain A.K., Murty M.N., Data clustering: a review, *ACM Comput. Surv. 31(3), p. 264-323*, (1999).

[15] Wunsch D., Xu R., Survey of clustering algorithms, *IEEE Transactions on Neural Networks 16(3), p. 645-678*, (2005).

[16] Han J., Kamber M., Pei J., Data mining (third edition), *Boston: Morgan Kaufmann*, (2012).

[17] Fraley C., How many clusters? which clustering method? answers via model-based cluster analysis, *The Computer Journal. 41(8), p. 586-588*, (1998).

[18] Dubes R.C., Jain A.K., Algorithms for clustering data, *Prentice Hall*, (1988).

[19] Arthur D., Gabow H., K-means++: The advantages of careful seeding, *SODA 2007, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, p. 1027-1035*, (2007).

[20] Celebi M.E., Kingravi H.A., Vela P.A., A comparative study of efficient initialization methods for the k-means clustering algorithm, *Expert Systems with Applications 40, p. 200-210*, (2013).

[21] Rana A., Singh A., Yadav A., K-means with three different distance metrics, *International Journal of Computer Applications 67(10), p. 13-17*, (2013).