Identifying Play Styles of Football Players Based on Match Event Data

Mark Riezebos
450121

Master Thesis Business Analytics and Quantitative Marketing*
Erasmus School of Economics
Erasmus University Rotterdam

Supervisor & second assessor
Dr. Michel van de Velden
Prof. dr. S. Ilker Birbil

Supervisors ORTEC Sports BV
Ruud van der Knaap
Dr. Bertus Talsma

ERASMUS UNIVERSITY ROTTERDAM

ORTEC SPORTS

March 12, 2021

*The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Abstract

In this thesis, we identify play styles of football players based on match event data. In order to do so, we first use the average locations of players in football matches to determine to which position group(s) they belong, which are goalkeepers, centre-backs, wing-backs, central midfielders, wingers and centre-forwards. Second, for each position group, we have statistics based on match event data for each player in that position group. We apply a dimension reduction technique to the data of player statistics, and then cluster football players based on the reduced player statistics to identify different play styles. For this we consider two approaches, of which the first is subsequently applying Principal Component Analysis (PCA) and hierarchical clustering. The second approach is the joint dimension reduction and clustering method Reduced k-means. For some football players (but not all) we know beforehand what their play style is (based on expert opinion). Using these players as representatives for the play styles they belong to, we show for central midfielders and centre-forwards that in general, Reduced k-means performs better than PCA and hierarchical clustering for identifying play styles. In terms of accuracy, PCA and hierarchical clustering achieves 52.1% and 69.0% for central midfielders and centre-forwards, respectively, whereas the accuracy for Reduced k-means is 58.3% and 75.9% for central midfielders and centre-forwards, respectively. We also calculate similarities of representing players to the cluster that most closely resembles the play style they represent. PCA and hierarchical clustering achieves an average similarity of 87.8% and 77.4% for central midfielders and centre-forwards, respectively, while Reduced k-means achieves an average similarity of 87.8% and 80.6% for central midfielders and centre-forwards, respectively.

Contents

1 Introduction

Professional football clubs aspire to accomplish success, which is achieved by winning as many matches as possible. Each club tries to allocate its resources in such a way that the performance of the team is maximised every single match. The use of data to analyse performance and achieve better results has become common amongst professional football clubs and other professional sports organizations (McLaughlin, 2018; Nalton, 2020; Volpicelli, 2020). One example of data analysis being used in sports and leading to better results is the story of the Oakland Athletics, an American professional baseball team competing in the Major League Baseball (MLB) (Goldstein, 2017). Ahead of the 2002 season, the general manager of the Oakland Athletics, Billy Beane, had to work with a relatively small budget. He decided to sign some players who were undervalued according to his and Paul DePodesta's data analysis. That season, the Oakland Athletics would go on to win the American League West division and win 20 games in a row, which was a record at the time. This story was later used to write the book *Moneyball: The Art of Winning un Unfair Game* (Lewis, 2003), of which a film has been made as well.

When recruiting an employee for a specific job, companies distinguish between job applicants based on skills that are needed for that job (Branine, 2008). This distinction between job applicants can be regarded as grouping potential employees into clusters, and the cluster a potential employee belongs to is one of the deciding factors for the company when making the decision on whether to hire that person (Zide et al., 2014). The recruitment of football players by professional football clubs can be regarded as a similar process. Each football player can be described by certain characteristics, such as his technical and tactical skills, his physical strength, and his mental strength. A football club is interested in a player when his characteristics match the characteristics that the club is looking for in a player. The "play style" of a football player is one such characteristic. Since a player's play style is a rather subjective concept, we have to define what we consider a play style to be. We use a similar definition for a football player's play style as Decroos and Davis (2019):

Play style (definition): A football player's play style is defined by the area(s) on the football field that he occupies most, and by the choices the player makes in these areas. In addition, the play styles of multiple football players can be compared by the choices they make in similar situations.

Furthermore, in our definition of a football player's play style, we make the following assumption:

Assumption: A football player's play style arises from the interplay between his skills and the tactics employed by the team. Football clubs tend to employ specific tactics that match the philosophy of the club of how football should be played in order to win matches. Tactics are also influenced by managers and even though managers do not always stay at a football club for multiple years, the majority of clubs do not change their manager during a season. As a result, we assume that a football player's play style does not change in a short period of time. That is, in a sequence of matches in a season, each player exhibits the same play style.

When recruiting players, it is important for football clubs to be able to identify play styles of individual football players, since the success of a player at a club is partly determined by how well his play style fits the tactics employed by the team. The identification of play styles of players can be done by scouts. This can however be regarded as subjective. In addition, scouts are not able to analyse all football players around the world. In this thesis, the goal is to identify play styles of football players based on statistics that are calculated from match event data, such that more football players can be analysed in a more objective manner (i.e., all players for which match event data is available). The match event data are collected by ORTEC Sports, and based on these data, ORTEC Sports calculates statistics for each player per match. These player match statistics, together with the playing positions reported by the media, are provided by ORTEC Sports for all matches of the 2018/2019 season of the England Premier League, Spanish Primera Division, Italian Serie A, German Bundesliga and France Ligue 1.

To describe a football player's play style, we distinguish between four aspects: a player's technical and tactical skills, a player's physique, and a player's mentality. Match event data captures two of these aspects to a certain extent, namely a player's technical and tactical skills. Since we are working with statistics based on match event data in this thesis, we are describing play styles based on those two aspects.

Scouting departments of football clubs can benefit from the identification of different play styles based on match event data, since it makes it easier to compare many players in an objective way based on their play style. Research that has been done thus far on analysing football players is mainly concerned with evaluating the performance of players. Research on identifying different play styles of individual football players however appears to be rather limited. The identification of play styles based on match event data can also be useful for other purposes than player recruitment. For example, it can be useful for player development, since the identification of a player's play style can contribute to assessing his strengths and/or weaknesses. This can help players and staff to determine the areas a player needs to work on. The identification of different play styles of players can also be useful for team development. Certain team tactics require specific players who fit the tactics of that team. Being able to objectively identify play styles of players can thus help the manager in deciding which players to line up based on the way he wants to play, or determining the team tactics based on the players he has to his disposal. In addition, the identification of different play styles can be helpful for game preparation and match analysis. Being able to recognize how the players of the opponent play and what their strengths and weaknesses are can be useful for adjusting the team tactics, and for preparing players for the type of opponents they are going to play against.

Having given our definition of a play style, we need some way to identify a football player's play style. First, as can be seen from the definition, a football player's play style is partly defined by the area(s) on the football field he occupies most, or in other words, his playing position. Hence, when identifying different play styles of football players, it is important to be aware of the fact that each play style can only belong to players who play in the same group of similar positions. For example, it would not make sense to compare the play style of a centre-forward with the play style of a centre-back, as the main task of a centre-forward is to score goals, whereas the main task of a centre-back is to prevent the opposing team from scoring goals.

Consequently, play styles of football players are position dependent, and should be identified separately for groups of similar playing positions. To do so, we introduce position groups. Each position group consists of multiple similar positions, with any two positions being regarded as similar if they are close to each other in terms of position on the field, and if they involve similar tasks during a football match. When deciding which positions are regarded as similar, we should also take into account to which team formation a playing position belongs. The groups of similar playing positions are defined and discussed in Section 3.1.1. After having defined which playing positions are regarded as similar, we have to determine for each player what his playing position is in every single match, such that we can derive for each player to which position group he belongs for every match. We then also have to decide how we determine to which position group(s) each player belongs over multiple matches, given the position group he belongs to for every single match.

Second, the considered definition of a play style implies that a football player's play style is partly defined by the choices he makes compared to other players in similar situations. These choices are reflected by the actions that a player performs in certain situations, as captured by match event data. We use the statistics that are calculated by ORTEC Sports based on the match event data for each player per match to determine a football player's play style. In particular, we do this by assessing how the statistics of a player differ compared to other players. For example, if a relative high number of the actions of a centre-forward are goal attempts and a relative low number of his actions are passes, this could indicate that he is a centre-forward who tries to score in every situation, and that he is more focused on scoring goals himself than creating opportunities for teammates to score goals. To be more precise in how we determine football players' play styles, we use the player match statistics to cluster players, such that each resulting cluster represents a certain play style. In this way, we compare players with different play styles between clusters, and players with similar play styles are compared within a cluster. In the following subsection we give a formal definition of the problem that we investigate in this thesis.

1.1 Problem definition

The main goal of this thesis is to identify different play styles of individual football players based on match event data. As a result, the main research question is:

"Can we identify play styles of individual football players based on match event data?"

A football player's play style partly depends on his playing position. Hence, to answer the main research question, we first have to discuss which playing positions and team formations are used by ORTEC Sports to register media lineups, and how these positions within formations can be categorized into different groups of similar playing positions. We should then assess for each football player per match in which of the playing positions within a team formation he plays, from which we can derive to which position group he belongs for every single match. This enables us to determine for each player to which position group(s) he belongs over multiple matches, given the position groups he belongs to in single matches. In this way we can create groups of

players that play in similar positions. Subsequently, for each position group we need to decide which player statistics we regard as relevant for describing play styles. Thereafter, we can start to identify different play styles within the created groups of players based on relevant player statistics, which is the main goal of this thesis. In this way we create distinct clusters of football players with similar play styles. As a result, in order to answer the main research question, we need to answer the following sub-questions:

* *"Can we assess to which position group(s) a football player belongs over multiple matches?"*
* *"Given for each football player to which position group(s) he belongs over multiple matches and thus to which group(s) of players with similar playing positions he belongs, can we identify play styles (by clustering players) based on relevant player statistics?"*

1.2 Outline

This thesis is structured as follows. An overview of related literature is given in Section 2. Section 3 gives a description of the data used in this thesis. We also discuss the playing positions and team formations that are used by ORTEC Sports to register media lineups, and we define how these playing positions within formations can be categorized into different groups of similar playing positions. The methods used to identify play styles of football players are discussed in Section 4. In Section 5 we present the results. We draw our final conclusion and discuss the limitations of the work done in this thesis in Section 6.

2 Literature review

In the field of football statistics, the prediction of football match outcomes is a much studied topic. This is reflected by the size of the sports betting industry, which in 2013 was estimated to be worth between $700 billion and $1 trillion per year, according to Darren Small, the director of integrity at the sports betting company Sportradar (Keogh and Rose, 2013).

Next to predicting football match outcomes, a lot of research has been done with respect to football analytics, which consists of analysing football matches, teams and players. A part of football analytics consists of research that is concerned with the identification of play styles of football players or teams. The play style of a football team differs from the play style of a football player, in the sense that the play style of a football team is defined by the collective choices that the players of a team make. In general, the problem of having to identify play styles of football players or teams can be treated as supervised or unsupervised, both having their advantages and disadvantages. The difference between the two is that with supervised learning, some benchmark players or teams are labeled with a certain play style, and other players or teams are classified to one of the play styles of those benchmark players or teams. On the other hand, with unsupervised learning, no players or teams are labeled with a play style, but play styles are identified by clustering players or teams based on statistics. Treating the problem as supervised means that some knowledge from a football expert is needed on which play styles belong to certain players, or which play styles are employed by certain teams. The advantage of this approach is the utilization of expert knowledge. The disadvantage however, is that it makes

the assessment less objective as one has to make a decision on which teams or players are used as a benchmark. Also, the gathering of expert knowledge can be expensive and/or time-consuming. When treating the problem as unsupervised, the advantages are that the resulting play styles are more flexible in the sense that they do not have to adhere to the predefined play styles based on expert knowledge, and that the process of gathering expert knowledge is not needed. However, the disadvantage of this approach is that it may be harder to recognize certain play styles without the use of expert knowledge.

Wensveen (2016) tried to identify play styles of football teams based on match event data in an unsupervised way. She showed that treating the problem as unsupervised does not lead to satisfying results, which is why the decision was made to treat the problem as supervised by making use of expert knowledge. Four benchmark team play styles were defined, and for each of those play styles a team was chosen that best represented that play style. Subsequently, the k-nearest neighbours algorithm was used to assign a team in a specific match to one of the predefined play styles. In a second, more flexible approach than k-nearest neighbours, Principal Component Analysis (PCA) was used together with expert knowledge to determine characteristic variables for describing play styles of football teams. Wensveen (2016) showed that both of the approaches that make use of expert knowledge lead to satisfying results.

Research has also been done with respect to identifying play styles of football players. For example, Taylor et al. (2004) discriminate between play styles of players in similar positions by using Chi-square tests to test for significant differences between statistics of different players. They make a distinction between centre-backs, wing-backs, midfielders and forwards, and the identification of play styles is based on performance indicators suggested by professional football coaches and expert match analysts. Peña and Navarro (2015) use affinity propagation clustering to find play styles of players based on their passing motifs, and Van de Ven (2018) uses player attributes obtained from the FIFA video games by EA Sports to cluster football players. The three previously mentioned papers however all use other data than the data used in this thesis, which is based on match event data.

Using match event data, the play style of a football player can be captured in a "player vector" that can be interpreted by human experts as well as machine learning systems (Geerts et al., 2018; Decroos and Davis, 2019). Geerts et al. (2018) and Decroos and Davis (2019) compare certain players based on those player vectors, but they do not describe any play styles that occur in general for different positions. PCA and hierarchical clustering are subsequently applied to match event data by Kalenderoğlu (2019) to find three different play styles for defenders, four play styles for midfielders and five play styles for forwards. Aalbers and Van Haaren (2018) propose 21 different play styles for six groups of positions. These groups of similar positions are goalkeepers, central defenders, wing-backs, central midfielders, wingers and centre-forwards. For example, for centre-forwards the authors propose the play styles Second Striker, Target Man, Poacher, and Mobile Striker. Football players are classified to one of the proposed play styles of the group of positions they belong to, using statistics that are calculated based on match event data. For each player, this is done by performing a probabilistic binary classification for each play style, such that the probability of a player belonging to a certain play style is obtained for each player and each play style. For each centre-forward this results in the probabilities of that

player belonging to the play styles Second Striker, Target Man, Poacher, and Mobile Striker. Each centre-forward is then assigned to the play style which has the highest probability of the four play styles.

The research performed in this thesis differs from the mentioned papers in some aspects. We aim to describe play styles that occur in general for different groups of similar playing positions, which are goalkeepers, central defenders, wing-backs, central midfielders, wingers and centre-forwards. The positions of players in single matches are determined based on their average locations, instead of using the lineups that are reported by the media. This can help with describing specific player roles. For identifying play styles of football players for each group of similar positions, we first consider the approach of subsequently applying PCA and hierarchical clustering to player statistics. As a second approach we use the joint dimension reduction and clustering method called Reduced k-means, instead of treating dimensionality reduction and clustering as separate methods. In this way, the variables that are found by the dimension reduction technique are chosen in such a way that we obtain an optimal clustering of football players.

3 Data description

In this section we describe the data used in this thesis, which consists of two components: formation data and player statistics based on match event data. The formation data contain each player's playing position as reported by the media for each match. The player statistics that are based on match event data consist of 180 statistics for each player per match. ORTEC Sports has provided us with these data for all matches of the 2018/2019 season of the England Premier League, Spanish Primera Division, Italian Serie A, German Bundesliga and France Ligue 1.

In Section 3.1 we describe the formation data, together with the average x- and y-locations of players in matches (the average x- and y-locations are 2 of the 180 player statistics for each match). The other player match statistics are described in Section 3.2.

3.1 Formation data

ORTEC Sports has provided us with formation data that contain the starting lineups reported by the media for each match. In football, each team is represented by 11 players, each of who has his own position on the football pitch. These 11 players can be positioned on the football pitch in multiple ways. To capture these differences in the collective positioning of a football team, formations are used. A formation is usually described by three or four numbers, which denote how many players are in each row of the formation, from the most defensive row to the most forward row. For example, the often used formation 4-3-3 consists of 4 defenders, 3 midfielders and 3 forwards. Since each team always plays with a goalkeeper irregardless of the formation they play in, the goalkeeper is not reported in formations, which is why the numbers of a formation always add up to 10. As a result of the fact that the 11 players of a team can be positioned on the football pitch in multiple ways, it holds that more than 11 playing positions exist in football. To be more precise, ORTEC Sports makes use of 22 unique playing positions, which are listed in Table 1.

Position label	Abbreviation	Position label	Abbreviation
Goalkeeper	GK	Centre-midfield	CM
Left-back	LB	Right-centre-midfield	RCM
Left-centre-back	LCB	Right-midfield	RM
Centre-back	CB	Attacking-midfield-left	AML
Right-centre-back	RCB	Attacking-midfield	AM
Right-back	RB	Attacking-midfield-right	AMR
Defensive-midfield-centre-left	DMCL	Left-winger	LW
Defensive-midfield-centre	DMC	Centre-forward-left	CFL
Defensive-midfield-centre-right	DMCR	Centre-forward	CF
Left-midfield	LM	Centre-forward-right	CFR
Left-centre-midfield	LCM	Right-winger	RW

Table 1: The playing positions used by ORTEC Sports with their abbreviations.

To register media lineups of teams in football matches, ORTEC Sports uses 12 different formations, each of which are formed by selecting 11 of the 22 playing positions. We define $\mathcal{F}$ as the set of formations that are used by ORTEC Sports, that is, $\mathcal{F} = \{$3-4-1-2, 3-4-3, 3-5-2, 4-1-4-1, 4-2-3-1, 4-3-1-2, 4-3-3, 4-4-1-1, 4-4-2, 4-5-1, 5-3-2, 5-4-1$\}$. These 12 different formations are listed in Table 2, with for each formation the number of times it is present in the data and the abbreviations of the playing positions that make up that formation. Note that we have analysed the correctness of standard position labels given by ORTEC Sports to certain formations. Based on this analysis, for some formations we have made some small adjustments to the playing positions that make up the formation (see Appendix A).

Formation (frequency)	Player 1	Player 2	Player 3	Player 4	Player 5	Player 6	Player 7	Player 8	Player 9	Player 10	Player 11
4-3-3 (890)	GK	LB	LCB	RCB	RB	LCM	CM	RCM	LW	CF	RW
4-2-3-1 (760)	GK	LB	LCB	RCB	RB	DMCL	DMCR	AML	AM	AMR	CF
4-4-2 (580)	GK	LB	LCB	RCB	RB	LM	LCM	RCM	RM	CFL	CFR
3-4-3 (412)	GK	LCB	CB	RCB	LM	LCM	RCM	RM	LW	CF	RW
3-5-2 (407)	GK	LCB	CB	RCB	LM	LCM	CM	RCM	RM	CFL	CFR
4-1-4-1 (150)	GK	LB	LCB	RCB	RB	DMC	LM	LCM	RCM	RM	CF
4-3-1-2 (117)	GK	LB	LCB	RCB	RB	LCM	CM	RCM	AM	CFL	CFR
4-4-1-1 (101)	GK	LB	LCB	RCB	RB	LM	LCM	RCM	RM	AM	CF
5-3-2 (68)	GK	LB	LCB	CB	RCB	RB	LCM	CM	RCM	CFL	CFR
5-4-1 (60)	GK	LB	LCB	CB	RCB	RB	LM	LCM	RCM	RM	CF
3-4-1-2 (49)	GK	LCB	CB	RCB	LM	LCM	RCM	RM	AM	CFL	CFR
4-5-1 (30)	GK	LB	LCB	RCB	RB	LM	LCM	CM	RCM	RM	CF

Table 2: The formations that are contained in $\mathcal{F}$, with for each formation the number of times it is present in the data and the abbreviations of the playing positions that make up that formation.

A player's playing position as reported by the media is not always his actual playing position. In order to make a comparison between a player's reported playing position and his actual playing position, we use the average x- and y-locations of players in a match. For example, the media lineup of Liverpool in round 28 of the 2018/2019 England Premier League is a 4-3-3 formation. If we however look at the average locations of the Liverpool players in this match (see Figure 1), we observe that Origi is labeled as a centre-forward (CF) and Mané as a left-winger (LW), whereas their average locations suggest that Origi played as a LW and Mané as a CF. This is just

one example of incorrect position labels. When comparing the average locations of players with the corresponding media lineups for more matches, we however observe that incorrect position labels occur regularly in the data set. In order to correct for these incorrect position labels, we determine the playing positions of players in matches based on their average x- and y-locations. How we do this exactly is explained in Section 4.1.

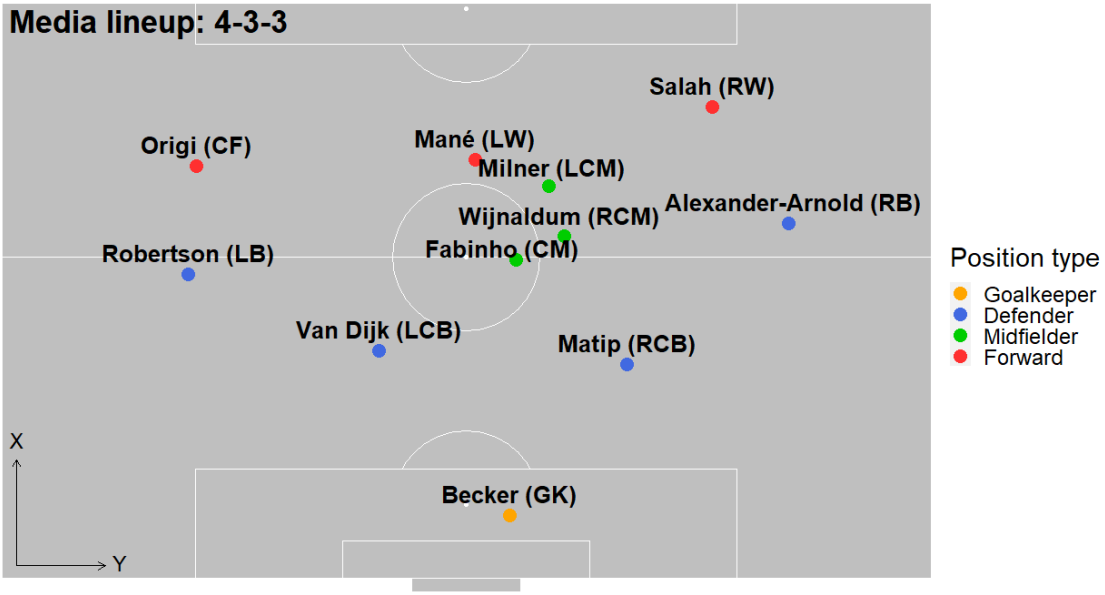

Figure 1: The average locations of Liverpool players in round 28 of the 2018/2019 England Premier League, with for each player the labeled position that corresponds to the media lineup.

Despite the fact that we observe incorrect position labels, we can see from Figure 2 (in which the weighted average location of each playing position is shown) that on average, the playing positions reported by the media are correct for each position label. The weighted average location of a playing position is calculated by taking the weighted average of all average locations of players in a match who played in that reported position (according to the media). The weights are the number of ball-related match events a player was involved in in a match (which is one of the 180 player match statistics). In this way, players with a high number of ball-related match events get a higher weight than players with a low number of ball-related match events.

3.1.1 Groups of similar playing positions

Having discussed which playing positions and team formations are used by ORTEC Sports to register lineups that are reported by the media, here we define which playing positions within formations are regarded as similar in this thesis. The groups of similar playing positions can be found in Table 3. Note that some position groups contain formation-specific positions. For example, a player who plays in LM is regarded as a wing-back if he plays in a 3-4-1-2, 3-4-3 or 3-5-2 formation, and he is regarded as a winger if he plays in a 4-1-4-1, 4-4-1-1, 4-4-2, 4-5-1 or 5-4-1 formation.

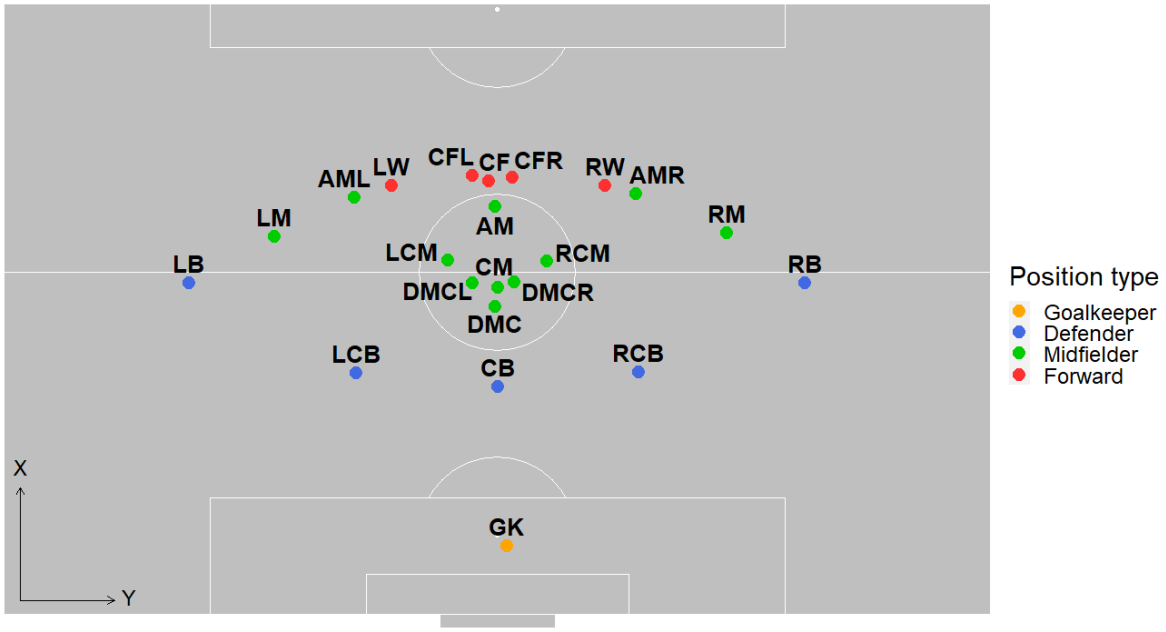

Figure 2: The weighted average locations of all positions.

Position group	Position(s)
Goalkeepers	GK
Centre-backs	LCB, CB, RCB
Wing-backs	LB, RB LM in 3-4-1-2, 3-4-3 or 3-5-2 RM in 3-4-1-2, 3-4-3 or 3-5-2
Central midfielders	DMCL, DMC, DMCR LCM, CM, RCM AM
Wingers	LM in 4-1-4-1, 4-4-1-1, 4-4-2, 4-5-1 or 5-4-1 RM in 4-1-4-1, 4-4-1-1, 4-4-2, 4-5-1 or 5-4-1 AML, AMR LW, RW
Centre-forwards	CFL, CF, CFR

Table 3: Overview of the groups of positions that are regarded as similar.

3.2 Player statistics based on match event data

The player statistics that we work with in this thesis are calculated based on match event data. In general there are two different types of data for football, i.e., match event data and tracking data. The difference between the two is that match event data consists of all events in a match in which the ball is involved (and some non-ball events), whereas tracking data is obtained by tracking the ball and all 22 players on the field for every single moment of the game. With match event data, for every moment of the game we thus do not have information on the 21 or 22 players (depending on whether an action is performed on the ball) who are not performing an action on the ball, whereas we do have this information with tracking data.

The collection of match event data yields approximately 1500 to 2000 observations on match events per match, from which 180 statistics are calculated for each player per match. These statistics range from simple count statistics like number of passes and number of goal attempts,

to more sophisticated statistics such as number of key actions and the grade a player gets for his performance in a match (this grading is explained in Appendix B.2). According to which aspect of football each statistic belongs to, the 180 statistics can be divided into 9 different categories. These categories are goalkeeper statistics, passing statistics, duel statistics, goal attempt statistics, possession statistics, goal type and goal attempt type statistics, set play statistics, defensive statistics and other statistics. An overview of the statistics that belong to each category can be found in Tables 29 up to 37 in Appendix B.3, with for each statistic a short description. In this thesis we are particularly interested in player statistics that can be used to describe football players' play styles and distinguish between different play styles. Which of the player statistics we use for this is discussed in Section 4.3.

4 Methodology

In this section we want to identify play styles of football players. To do this, we first determine the playing positions of football players in single matches based on their average locations. From the playing positions in single matches we can derive for each player to which position group he belongs for every single match. Then we need to determine to which position group(s) each player belongs over multiple matches, given their position groups in single matches. How we determine the playing positions of players in single matches and the position groups of players over multiple matches is both explained in Section 4.1. After determining to which position group(s) each football player belongs over multiple matches, our goal is to create distinct clusters of football players with similar play styles for each position group. In order to do this, we first need to decide for each position group which statistics are relevant for describing play styles of players in that position group. This selection of player statistics is discussed in Section 4.2. Then, for each position group we cluster football players based on the selected statistics for that position group. Two approaches are considered for clustering football players. In the first approach we subsequently apply Principal Component Analysis (PCA) and hierarchical clustering to the selected player statistics. In the second approach we use use the joint dimension reduction and clustering technique Reduced k-means to cluster football players in terms of their play styles. Both of these approaches are explained in Section 4.3. In Section 4.4 we explain how we evaluate the resulting clusters of play styles.

4.1 Determining football players' playing positions and position groups

In order to determine for each player to which position group(s) he belongs over multiple matches, we first determine the playing positions of players in single matches, which is explained in Section 4.1.1. From the playing positions in single matches we can derive for each player to which position group he belongs for every single match. In Section 4.1.2 we explain how we determine for each player to which position group(s) he belongs over multiple matches, given for every single match to which position group he belongs.

4.1.1 Determining football players' playing positions in single matches

The problem we are faced with, is having to determine the playing positions of football players in single matches. We do this by considering the 11 players in the starting lineup for every team in every match, and determining the formation of those 11 players based on the average locations of the 11 players. From the resulting formation for every team in every single match, we can derive the playing position of every player in every single match. Below we explain in detail how we determine the formation of one team in a particular match, which is applied to every team in every single match.

Since we are working with data from one season (2018/2019) for five different competitions (England Premier League, Spanish Primera Division, Italian Serie A, German Bundesliga and France Ligue 1), each team in a particular match can uniquely be identified by the team and the playing round of the match. As a result, let us denote the average locations of the 11 players in the starting lineup for team t in playing round d by $\left\{\bar{x}_i^{(t,d)}, \bar{y}_i^{(t,d)} : i = 1, \ldots, 11\right\}$. For each match this gives us the average locations of 11 players, which we call a match profile.

Recall that $\mathcal{F}$ is the set of formations that we use in this thesis (see Table 2). For each formation $f \in \mathcal{F}$, we define P_f as the set that consists of all positions that are in formation f (it thus hold that $|P_f| = 11$ for each formation $f \in \mathcal{F}$). For example, for the formation 4-3-3 we have $P_{\text{4-3-3}} = \{\text{GK, LB, LCB, RCB, RB, LCM, CM, RCM, LW, CF, RW}\}$. For each formation $f \in \mathcal{F}$, we calculate weighted averages of the 11 positions that are in P_f. Each weighted average of a position p in a formation f is calculated over all teams t, playing rounds d and players i. However, we only include the average location $\left\{\bar{x}_i^{(t,d)}, \bar{y}_i^{(t,d)}\right\}$, if in playing round d, team t plays in formation f and player i of team t ($i = 1, \ldots, 11$) plays in position p (according to the reported media positions). Hence, for each formation $f \in \mathcal{F}$ and each position $p \in P_f$, the weighted average location of position p is calculated as

$$\left\{\bar{x}_p^{(f)}, \bar{y}_p^{(f)}\right\} = \left\{\frac{\sum_{t,d}\sum_{i=1}^{11} I_{fpi}^{(t,d)} w_i^{(t,d)} \bar{x}_i^{(t,d)}}{\sum_{t,d}\sum_{i=1}^{11} I_{fpi}^{(t,d)} w_i^{(t,d)}}, \frac{\sum_{t,d}\sum_{i=1}^{11} I_{fpi}^{(t,d)} w_i^{(t,d)} \bar{y}_i^{(t,d)}}{\sum_{t,d}\sum_{i=1}^{11} I_{fpi}^{(t,d)} w_i^{(t,d)}}\right\}, \tag{1}$$

with the weights $w_i^{(t,d)}$ the number of ball-related match events that player i of team t was involved in in playing round d. In this way, the average locations of players with a high number of ball-related match events get a higher weight than the average locations of players with a low number of ball-related match events. Furthermore, $I_{fpi}^{(t,d)}$ is an indicator variable which is defined as

$$I_{fpi}^{(t,d)} = \begin{cases} 1 & \text{if in playing round } d, \text{ team } t \text{ plays in formation } f, \text{ and player } i \text{ of team} \\ & t \text{ plays in position } p \text{ according to the lineup reported by the media} \\ 0 & \text{otherwise.} \end{cases}$$

For each formation $f \in \mathcal{F}$, the 11 weighted average positions $\left\{\bar{x}_p^{(f)}, \bar{y}_p^{(f)} : p \in P_f\right\}$ make up a formation profile. Hence, we get 12 formation profiles (since $|\mathcal{F}| = 12$).

We determine the formation of team t in playing round d by assigning the corresponding match profile to one of the 12 formation profiles. We do this by assessing how similar the match

profile is to each of the 12 formation profiles, and then assigning the match profile to the formation profile that has the highest similarity. Before we assess how similar a match profile is to a formation profile, we standardize each match profile and each formation profile. Each match/-formation profile is standardized by calculating the average (x, y) coordinate of that profile, and then subtracting the average (x, y) coordinate from each of the average locations/weighted average positions of that profile. In this way we ensure that how offensive a team plays, does not influence the formation profile that that team gets assigned in a match. The reason that we take this into account is that different teams can play in the same formation, but at the same time one team can play more offensively than another team (which is usually the result of differences in how good teams are). For example, Manchester City tends to play in an offensive 4-3-3 formation, whereas Brighton and Hove Albion tends to play in an defensive 4-3-3 formation.

We have now explained how we obtain the match profile $\left\{\bar{x}_i^{(t,d)}, \bar{y}_i^{(t,d)} : i = 1, \ldots, 11\right\}$ for team t in playing round d, and how we obtain the 12 formation profiles $\left\{\bar{x}_p^{(f)}, \bar{y}_p^{(f)} : p \in P_f\right\}$. What we want to do, is calculate how similar each match profile is to each of the formation profiles. This similarity measure can be calculated by assigning each of the 11 players in the match profile to one of the 11 positions in the formation profile in such a way that the total distance between the players and the positions the players are assigned to is minimized. To this end, for team t in playing round d and the formation profile that corresponds to the formation f, we define $\mathbf{Z}_f^{(t,d)}$ as an (11×11) matrix, with the element in row i and the column corresponding to position $p \in P_f$ being given by

$$z_{f,ip}^{(t,d)} = \begin{cases} 1 & \text{if in playing round } d, \text{ player } i \text{ of team } t \\ & \text{is assigned to position } p \in P_f \\ 0 & \text{otherwise.} \end{cases}$$

The problem at hand can then mathematically be stated as

$$\min_{\mathbf{Z}_f^{(t,d)}} \quad \sum_{i=1}^{11} \sum_{p \in P_f} z_{f,ip}^{(t,d)} * d\Big(\Big\{\bar{x}_i^{(t,d)}, \bar{y}_i^{(t,d)}\Big\}, \Big\{\bar{x}_p^{(f)}, \bar{y}_p^{(f)}\Big\}\Big) \tag{2}$$

$$\text{s.t.} \quad \sum_{p \in P_f} z_{f,ip}^{(t,d)} = 1 \text{ for } i = 1, \ldots, 11 \tag{3}$$

$$\sum_{i=1}^{11} z_{f,ip}^{(t,d)} = 1 \text{ for } p \in P_f, \tag{4}$$

$$z_{f,ip}^{(t,d)} \in \{0, 1\} \text{ for } i = 1, \ldots, 11 \text{ and } p \in P_f. \tag{5}$$

Here, $d(\cdot)$ is a function to calculate the Euclidean distance between $\left\{\bar{x}_i^{(t,d)}, \bar{y}_i^{(t,d)}\right\}$ and $\left\{\bar{x}_p^{(f)}, \bar{y}_p^{(f)}\right\}$, which is defined as

$$d\Big(\{x_i, y_i\}, \{x_j, y_j\}\Big) = \sqrt{\Big(x_i - x_j\Big)^2 + \Big(0.7 * \Big(y_i - y_j\Big)\Big)^2}. \tag{6}$$

Note that the y-coordinates are multiplied by the factor 0.7. This is because the x- and y-coordinates are both measured on a [0,100] scale, while a football field is longer in the x-direction

than in the y-direction. Hence, ORTEC Sports always uses the factor 0.7 to adjust for this, since the length of a football pitch in the y-direction is on average 0.7 times the length of a football pitch in the x-direction. The constraints (3), (4) and (5) ensure that each player is assigned to exactly one position, and that each position is assigned to exactly one player. This minimization problem is a combinatorial optimization problem which is known as the assignment problem (Bertsekas, 1998). Generally stated, the assignment problem consists of q agents and r tasks. Each agent can be assigned to any of the r tasks, and for each agent-task combination it is known what the cost is of letting that agent i perform task j. This cost is denoted as $c(i, j)$. What we then want to find is an allocation of agents to tasks, such that as many tasks as possible are performed while minimizing the total cost. Additional constraints are that each agent can be assigned to at most one task, and each task can be performed by at most one agent. This is the same problem as the one we have mathematically described in Equations (2), (3), (4) and (5), with the difference being that in our case the agents are players belonging to a match profile which are represented by their average locations, and the tasks are positions belonging to a formation profile which are represented by their weighted average locations. In our case, the cost of assigning player i of the match profile corresponding to team t in playing round d to position p of the formation profile that corresponds to the formation f, is thus given by $d\Big(\Big\{\bar{x}_i^{(t,d)}, \bar{y}_i^{(t,d)}\Big\}, \Big\{\bar{x}_p^{(f)}, \bar{y}_p^{(f)}\Big\}\Big)$, as we want to minimize the total distance of the players to their assigned positions. Furthermore, we have $q = r = 11$, since each team consists of 11 players and each formation consists of 11 positions. The fact that we have $q = r$ in our case means that we are dealing with a special case of the assignment problem, which is called the balanced assignment problem. Solving the balanced assignment problem by iterating over all possible assignments and storing the assignment with the lowest total cost becomes inefficient as q becomes large. This is because in the balanced assignment problem, there are $q!$ different assignments possible, which in our case is $11! = 39,916,800$ (which is just for calculating the distance between one match profile and one formation profile). Kuhn (1955) recognized the inefficiency of this brute-force solution as well, and came up with a heuristic to solve the balanced assignment problem more efficiently. This heuristic is called the Hungarian method, and it is implemented by the function `solve_LSAP` from the package "clue" in R (Hornik, 2019), which we use to solve the minimization problem given by Equations (2), (3), (4) and (5). The minimization problem is solved for each team t in each playing round d and for each formation $f \in \mathcal{F}$. Then, for each team t in each playing round d, we take the formation f^* for which the objective function has the lowest value, and we assign that formation f^* to team t in playing round d. From the resulting assigned formation for every team in every single match, we derive the playing position of every player in every single match.

4.1.2 Determining football players' position groups over multiple matches

Given the playing positions of football players in single matches, we want to determine to which position group(s) each player belongs over multiple matches. First we can derive from Table 3 for each player per match to which position group he belongs based on his assigned position. Then, for each player we know to which position group he belongs for every match he played, and we also know for each of those matches how many minutes the player has played (this is one of the

statistics that can be found in Table 37). The way we determine the position groups of players over multiple matches, is by calculating for each player how many minutes he played in total in each position group, and if a player has played at least 450 minutes in a position group (which is equal to playing 5 full matches), we regard that player as belonging to that position group for the corresponding matches. We then aggregate his statistics over those matches, and these aggregated statistics are used to cluster football players in terms of their play styles. How we aggregate statistics over multiple matches is explained in Section 4.2. An advantage of the used approach is that it allows for players to belong to multiple position groups, as there are football players who are able to play in multiple positions that do not belong to the same position group.

4.2 Selection of player statistics to use for clustering

Having performed the methods described in Section 4.1, for each position group we are provided with players who have played at least 450 minutes in that position group. For each of these players we have the match statistics for the matches in which they played in the corresponding position group. What we now want to do, is select the player statistics that are relevant for describing play styles of football players, and aggregate these statistics over multiple matches. The resulting aggregated statistics for each player in each position group can then be used to cluster football players. For the results in Section 5, we only focus on the position groups central midfielders and centre-forwards, since we think these are the two position groups for which the identification of play styles is most relevant and interesting, and because the evaluation of results would otherwise become rather extensive. Consequently, we also only discuss the selected player statistics for central midfielders and centre-forwards. For the other position groups, the selected player statistics and the implementation of the other methods from this section can be found in the scripts that are written in R (R Core Team, 2020) for this thesis.

We aggregate each statistic for each player in each position group over the multiple matches by taking the sum of that statistic over the multiple matches. For example, if a player has played 3 matches as a central midfielder and in those 3 matches he has 66, 84 and 51 passes, respectively, then his aggregated number of passes becomes $66 + 84 + 51 = 201$. Then, for each position group we select the aggregated statistics that we consider to be relevant for describing play styles of players in that position group, and each aggregated statistic is taken relative to or as a percentage of another aggregated statistic. For example, for central midfielders as well as for centre-forwards we have included the statistic `passesPercentageForward`, which is defined as the total number of forward passes divided by the total number of passes of a player. We take relative statistics because we think that these better reflect the choices that a player makes than absolute statistics (e.g., the absolute number of forward passes of a player).

Because we are calculating relative statistics, it can occur that we get missing values because we are dividing by a statistic that has a value of zero. One possible way to deal with these missing values is by replacing them with zero's. However, this can destroy the multivariate structure of the data, which can lead to biased results when analysing the data. For example, when calculating the statistic `shotsPercentageWithHead` (which is defined as the number of headed shots divided by the number of shots), it can occur that a player has produced zero shots. This however does not mean that if that player would produce a shot, that this will never

be a headed shot.

Consequently, we use k-nearest neighbor imputation with $k = 10$ (this is the default) to replace the missing values (Troyanskaya et al., 2001). For each player that has any missing values for the statistics, k-nearest neighbor imputation computes the k players that are closest to that player in terms of Euclidean distance, based on the statistics that are not missing for the player in question. These k players are called the k nearest neighbors. Each candidate neighbor might be missing some of the statistics used to calculate the distance. In this case, the distance between the player in question and the candidate neighbor is calculated based on the statistics that are not missing for both players, and this distance is then multiplied by the factor $\frac{n_n+n_m}{n_n}$. Here, n_n is the number of non-missing values for both the player in question and the candidate neighbor, and n_m is the number of statistics for which the value is missing for the player in question, the candidate neighbor, or both. The statistics for which the value is missing for the player in question are then replaced by the means of those statistics for the k nearest neighbors. The resulting selected player statistics for central midfielders and centre-forwards can be found in Appendix C.

4.3 Clustering football players based on player statistics

After determining to which position group(s) football players belong over multiple matches and which statistics are relevant for describing play styles, our goal is to create distinct clusters of football players with similar play styles. First we use the same approach as Kalenderoğlu (2019) to cluster football players based on their style of play. This approach consists of first applying Principal Component Analysis (PCA) to the selected player statistics, and then applying hierarchical clustering to the principal components that are found by PCA. An explanation of these two methods is given in Section 4.3.1. In the second approach for clustering football players, we use a joint dimension reduction and clustering method instead of treating dimensionality reduction and clustering as two separate parts. This joint dimension reduction and clustering technique is called Reduced k-means, and it is explained in Section 4.3.2.

Before we explain the methods used to cluster football players in terms of their play styles, let us first introduce some general notation. Based on the position group(s) found for each player in Section 4.1 and the selected player statistics from Section 4.2, for each position group we obtain an $(N \times q)$ data matrix $\mathbf{X}$, with N the number of players that are in the position group in question, and q the number of selected statistics for that position group. We can write $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)'$, with $\mathbf{x}_i$ a vector of length q containing the statistics of the i-th player. Note that this is just general notation for one position group, and that the values of N and q can differ per position group. The methods explained in Sections 4.3.1 and 4.3.2 to cluster football players are applied to each of the six position groups.

4.3.1 PCA and hierarchical clustering

In the first approach to cluster football players in terms of their play styles, we first apply PCA to reduce the dimension q of the $(N \times q)$ data matrix $\mathbf{X}$. How this works is explained in Section 4.3.1.1. Let us denote the matrix that results from PCA by $\mathbf{X}^*$, which has dimension $(N \times p)$

with $p < q$. We then apply hierarchical clustering to $\mathbf{X}^*$ to cluster the N football players, which is explained in Section 4.3.1.2.

4.3.1.1 Principal Component Analysis (PCA) We are given the $(N \times q)$ data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_q)$, of which we want to reduce the number of variables/columns. We do this by applying Principal Component Analysis (Mardia et al., 1979). However, before we perform PCA we should standardize the columns of $\mathbf{X}$. This prevents variables with high standard deviations to have a greater influence on the results than other variables. Hence, each observation x_{ij} (for $i = 1, \ldots, N$ and $j = 1, \ldots, q$) is first subtracted by the mean of the corresponding variable ($\bar{x}_j$), and then divided by the standard deviation of the corresponding variable (s_j). It holds that $\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$ and $s_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{ij} - \mu_j)^2}$ for $j = 1, \ldots, q$.

PCA is a method to transform a high-dimensional $(N \times q)$ data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_q)$ to a lower-dimensional $(N \times p)$ matrix $\mathbf{X}^* = (\mathbf{x}_1^*, \ldots, \mathbf{x}_p^*)$, with $p < q$. To be precise, PCA actually finds a matrix that still has q variables (which are called principal components), but the dimension of the matrix is reduced by only retaining the first p principal components. The transformation is taken in such a way that each of the p principal components are a linear combination of the q original variables, which explain a decreasing amount of the variance in the original data matrix. The transformation is also taken in such a way that each of the p principal components are linearly uncorrelated. Hence, the first principal component $\mathbf{x}_1^*$ is the linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_q$ for which $\mathrm{Var}(\mathbf{x}_1^*)$ is maximized, the second principal component $\mathbf{x}_2^*$ is the linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_q$ for which $\mathrm{Var}(\mathbf{x}_2^*)$ is maximized and such that $\mathbf{x}_2^*$ is orthogonal to $\mathbf{x}_1^*$, the third principal component $\mathbf{x}_3^*$ is the linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_q$ for which $\mathrm{Var}(\mathbf{x}_3^*)$ is maximized and such that $\mathbf{x}_3^*$ is orthogonal to $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$, and so on. Hence, the i-th principal component can be written as

$$\mathbf{x}_i^* = \sum_{j=1}^{q} \alpha_{ij}\mathbf{x}_j, \text{ for } i = 1, \ldots, p. \tag{7}$$

Here, the elements $\{\alpha_{ij} : j = 1, \ldots, q\}$ are called the loadings of the i-th principal component, and it holds that $\sum_{j=1}^{q} \alpha_{ij}^2 = 1$ for $i = 1, \ldots, p$. This restriction for the loadings prevents that the variance of a principal component can become arbitrary large. It also holds that $\mathrm{Var}(\mathbf{x}_1^*) > \mathrm{Var}(\mathbf{x}_2^*) > \cdots > \mathrm{Var}(\mathbf{x}_{p-1}^*) > \mathrm{Var}(\mathbf{x}_p^*)$, and $\mathrm{Cov}(\mathbf{x}_i^*, \mathbf{x}_j^*) = 0$ for $i \neq j$.

In order to perform PCA, we should find the loadings $\{\alpha_{ij} : j = 1, \ldots, q\}$ for each of the principal components ($i = 1, \ldots, p$), while satisfying the given constraints. These loadings can be found by performing an eigendecomposition on the $(q \times q)$ correlation matrix Σ of $\mathbf{X}$. However, according to Mardia et al. (1979), for numerical accuracy it is preferred to perform a Singular Value Decomposition (SVD) on $\mathbf{X}$, which can also be used to find the loadings. Therefore, we use the latter approach to find the loadings $\{\alpha_{ij} : j = 1, \ldots, q\}$ for each of the principal components ($i = 1, \ldots, p$). This approach of performing PCA by using a SVD works as follows. If we perform a Singular Value Decomposition on $\mathbf{X}$, we obtain

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}', \tag{8}$$

with $\mathbf{U}$ and $\mathbf{V}$ orthogonal matrices of dimensions $(N \times N)$ and $(q \times q)$, respectively. The columns

of $\mathbf{U}$ and $\mathbf{V}$ are called the left-singular and right-singular vectors of $\mathbf{X}$, respectively. The left-singular vectors of $\mathbf{X}$ consist of the orthonormal eigenvectors of $\mathbf{XX'}$, and the right-singular vectors of $\mathbf{X}$ consist of the orthonormal eigenvectors of $\mathbf{X'X}$. Furthermore, $\mathbf{\Lambda}$ is an $(N \times q)$ matrix for which the element in row i and column i is a non-negative number which we denote as λ_i, and all other elements of $\mathbf{\Lambda}$ are 0. The values $\{\lambda_i : i = 1, \ldots, \min\{N, q\}\}$ are the square roots of the non-negative eigenvalues of both $\mathbf{XX'}$ and $\mathbf{X'X}$, which are called the singular values of $\mathbf{X}$, and we order them such that $\lambda_1 > \lambda_2 > \cdots > \lambda_{\min\{N,q\}-1} > \lambda_{\min\{N,q\}}$. In order to obtain the lower-dimensional matrix $\mathbf{X}^*$, we only retain the first p columns of $\mathbf{V}$, such that we get the $(q \times p)$ matrix $\mathbf{V}^*$. The i-th column of $\mathbf{V}^*$ then contains the loadings $\{\alpha_{ij} : j = 1, \ldots, q\}$ for each of the principal components $(i = 1, \ldots, p)$, and we get that $\mathbf{X}^* = \mathbf{XV}^*$. Note that for this solution it holds that $\sum_{j=1}^{q} \alpha_{ij}^2 = 1$ for $i = 1, \ldots, p$ as $\mathbf{V}$ is an orthogonal matrix. For the proof that it also holds that $\text{Var}(\mathbf{x}_1^*)$, ..., $\text{Var}(\mathbf{x}_p^*)$ are maximized for this solution while satisfying the constraints $\text{Var}(\mathbf{x}_1^*) > \text{Var}(\mathbf{x}_2^*) > \cdots > \text{Var}(\mathbf{x}_{p-1}^*) > \text{Var}(\mathbf{x}_p^*)$, and $\text{Cov}(\mathbf{x}_i^*, \mathbf{x}_j^*) = 0$ for $i \neq j$, we refer to Mardia et al. (1979).

4.3.1.2 Hierarchical clustering After applying PCA we obtain an $(N \times p)$ matrix $\mathbf{X}^*$, which contains the scores on each of the p principal components for each of the N football players. We apply hierarchical clustering to $\mathbf{X}^*$ in order to cluster the football players. We can write $\mathbf{X}^* = (\mathbf{x}_1^*, \ldots, \mathbf{x}_N^*)'$, with $\mathbf{x}_i^*$ a vector of length p containing the principal components of the i-th player.

In general, clustering is a method that groups a set of objects in such a way that objects from the same cluster are more similar than objects from different clusters. In our case, these objects are football players. To cluster the football players, we need to define (dis)similarity between players. We define the dissimilarity between two players to be the distance between those two players. That is, the dissimilarity between players i and j ($i, j \in \{1, \ldots, N\}$ and $i \neq j$) is given by the distance between $\mathbf{x}_i^*$ and $\mathbf{x}_j^*$. Hence, opting for a clustering with a high intra-cluster similarity and a low inter-cluster similarity, requires a clustering in which players from the same cluster are close to each other, and players from different clusters should be far away from each other. Several distance measures have been used in the literature to calculate the distance between two objects, including Euclidean distance and Manhattan distance. Aggarwal et al. (2001) compare the effectiveness of Euclidean distance and Manhattan distance for clustering, and they show that using Manhattan distance is preferred for a dimensionality of 20 or higher. As can be seen in Section 5, the highest dimension that we use for clustering is 5. As a result, we choose to use Euclidean distance for calculating the distance between two players. Hence, the distance between $\mathbf{x}_i^*$ and $\mathbf{x}_j^*$ is given by

$$d(\mathbf{x}_i^*, \mathbf{x}_j^*) = \sqrt{\sum_{m=1}^{p} \left(\mathbf{x}_{im}^* - \mathbf{x}_{jm}^*\right)^2},$$

with $\mathbf{x}_{im}^*$ and $\mathbf{x}_{jm}^*$ the m-th principal component scores of players i and j, respectively.

In this thesis we use agglomerative hierarchical clustering (Jain and Dubes, 1988). This clustering algorithm has a bottom up approach: each football player begins as its own cluster, and in each step of the algorithm two clusters are merged to form one larger cluster. This

process is iteratively repeated until all football players are merged into one cluster. In each iteration of the algorithm we thus have to decide which two clusters are going to be merged. This is done by taking the two clusters that are most similar to each other. The dissimilarity between two clusters is defined in terms of distance, which means that the two clusters that are closest to each other are merged in each iteration. Note that the distance between two clusters is something different than the distance between two players, as the distance between two clusters is the distance between two sets of players. A distance measure for calculating the dissimilarity between two clusters is called a linkage criterion. In this thesis we consider three linkage criteria: complete linkage, average linkage and Ward linkage. Complete linkage computes the distance between two clusters as the maximum distance between any two players of the two clusters. With average linkage, the distance between two clusters is computed by taking the average of the distances between all possible pairs of players from the two clusters. Given two clusters A and B, the similarity measures complete linkage and average linkage are mathematically defined by the expressions given below.

$$\text{Complete linkage: } d(A,B) = \max_{\mathbf{x}_i^* \in A, \mathbf{x}_j^* \in B} \left\{ d\big(\mathbf{x}_i^*, \mathbf{x}_j^*\big) \right\} \tag{9}$$

$$\text{Average linkage: } d(A,B) = \frac{1}{|A||B|} \sum_{\mathbf{x}_i^* \in A, \mathbf{x}_j^* \in B} d\big(\mathbf{x}_i^*, \mathbf{x}_j^*\big) \tag{10}$$

In Equation (10), we use $|.|$ to denote the cardinality of a cluster. If we define the centroid of a cluster A to be $C_A = \frac{1}{|A|} \sum_{\mathbf{x}_i^* \in A} \mathbf{x}_i^*$ (with a similar expression for the centroid of a cluster B), then the Ward linkage is given by the expression below (Ward Jr., 1963; Murtagh and Legendre, 2014).

$$\text{Ward linkage: } d(A,B) = \frac{|A||B|}{|A| + |B|} d(C_A, C_B)^2 \tag{11}$$

After applying the hierarchical clustering algorithm, we obtain a hierarchical clustering of the football players. If we start at the top of the hierarchy, we have one cluster that contains all football players. From this point, we can move down the hierarchy, and in each step we take down the hierarchy, one cluster is divided into two clusters. Hence, say that we would want to obtain k clusters. Then we should take $k-1$ steps down the hierarchy from the top. This would give us a clustering which divides the N football players into k clusters.

4.3.2 Reduced k-means

In the previously considered approach to cluster football players in terms of their play styles, we subsequently applied PCA and hierarchical clustering to the $(N \times q)$ data matrix $\mathbf{X}$ that contains q player statistics for each of the N players. However, the principal components that are found by PCA may not contribute much to finding clusters of football players. This is because PCA selects linear combinations of the variables (which are the columns of $\mathbf{X}$) such that as much variance is retained in as few dimensions as possible. This however provides us

with an $(N \times p)$ matrix $\mathbf{X}^*$ that does not necessarily contain well separated clusters. Hence, in the second approach to cluster football players in terms of their play styles, we use a joint dimension reduction and clustering method, which is called Reduced k-means (De Soete and Carroll, 1994). This method simultaneously performs dimension reduction and clustering to divide the N players into k clusters, which is done in such a way that the variance between the found clusters is maximized.

We are again given the $(N \times q)$ data matrix $\mathbf{X}$. In the same way as we did before applying PCA, we first standardize each column of $\mathbf{X}$. In this way, we prevent variables with high standard deviations to have a greater influence on the results than other variables. When applying Reduced k-means, we want to find a clustering of the N players in $\mathbf{X}$ in a lower-dimensional subspace of the q columns of $\mathbf{X}$. To this end, we define k as the number of clusters, p as the lower dimension (it thus holds that $p < q$), and $\mathbf{Z}$ as an $(N \times k)$ matrix that indicates for each player to which cluster he belongs, with the element in row i and column j of $\mathbf{Z}$ being given by

$$z_{ij} = \begin{cases} 1 & \text{if player } i \text{ belongs to cluster } j \\ 0 & \text{otherwise.} \end{cases}$$

We impose that $\sum_{j=1}^{k} z_{ij} = 1$ for $i = 1, \ldots, N$, such that each player gets assigned to exactly one cluster. Furthermore, we define the $(k \times p)$ matrix $\mathbf{C} = (c_1, \ldots, c_k)'$ as the matrix that contains the cluster centroids of each of the k clusters in the lower dimension p. Lastly, we define $\mathbf{L}$ to be a $(q \times p)$ orthonormal matrix (that is, $\mathbf{L}'\mathbf{L} = I_p$) that contains the loadings to transform the variables in the original q-dimensional space to the reduced p-dimensional space. In Reduced k-means, the dimension reduction and cluster allocation is performed in such a way that we maximize the variance between clusters in the reduced space. This is done by minimizing the objective function

$$f(\mathbf{Z}, \mathbf{C}, \mathbf{L}) = ||\mathbf{X} - \mathbf{Z}\mathbf{C}\mathbf{L}'||_F^2, \tag{12}$$

with $||\cdot||_F$ the Frobenius norm of a matrix, which is defined as $\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}$ for an $(m \times n)$ matrix $\mathbf{A}$. The cluster centroids in the reduced space are given by $\mathbf{C} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\mathbf{L}$. If we obtain k clusters with each their own centroid, then these k centroids that are given by $\mathbf{C}$ always define a $(k-1)$-dimensional subspace. Hence, if we choose p such that $p < k$, we are able to achieve dimension reduction. Substituting $\mathbf{C} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\mathbf{L}$ in the objective function gives

$$\begin{aligned} f(\mathbf{Z}, \mathbf{L}) &= ||\mathbf{X} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\mathbf{L}\mathbf{L}'||_F^2 && (13) \\ &= ||\mathbf{X} - \mathbf{P}\mathbf{X}\mathbf{L}\mathbf{L}'||_F^2, && (14) \end{aligned}$$

with $\mathbf{P} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ (Yamamoto and Hwang, 2014). Using the trace operator, it holds that $||\mathbf{A}||_F^2 = \text{tr}(\mathbf{A}'\mathbf{A})$ for an $(n \times m)$ matrix $\mathbf{A}$, such that we can write

$$\begin{aligned} f(\mathbf{Z}, \mathbf{L}) &= ||\mathbf{X} - \mathbf{P}\mathbf{X}\mathbf{L}\mathbf{L}'||_F^2 \\ &= \text{tr}\big((\mathbf{X} - \mathbf{P}\mathbf{X}\mathbf{L}\mathbf{L}')'(\mathbf{X} - \mathbf{P}\mathbf{X}\mathbf{L}\mathbf{L}')\big) \end{aligned}$$

$$
\begin{aligned}
&= \text{tr}(\mathbf{X}'\mathbf{X}) - \text{tr}(\mathbf{X}'\mathbf{P}\mathbf{X}\mathbf{L}\mathbf{L}') - \text{tr}(\mathbf{L}\mathbf{L}'\mathbf{X}'\mathbf{P}'\mathbf{X}) + \text{tr}(\mathbf{L}\mathbf{L}'\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{X}\mathbf{L}\mathbf{L}') \\
&= \text{tr}(\mathbf{X}'\mathbf{X}) - 2 \cdot \text{tr}(\mathbf{X}'\mathbf{P}\mathbf{X}\mathbf{L}\mathbf{L}') + \text{tr}(\mathbf{L}'\mathbf{L}\mathbf{L}'\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{X}\mathbf{L}) \\
&= \text{tr}(\mathbf{X}'\mathbf{X}) - 2 \cdot \text{tr}(\mathbf{L}'\mathbf{X}'\mathbf{P}\mathbf{X}\mathbf{L}) + \text{tr}(\mathbf{L}'\mathbf{X}'\mathbf{P}\mathbf{X}\mathbf{L}) \\
&= \text{tr}(\mathbf{X}'\mathbf{X}) - \text{tr}(\mathbf{L}'\mathbf{X}'\mathbf{P}\mathbf{X}\mathbf{L}).
\end{aligned}
$$

Here we use that for an $(n \times m)$ matrix $\mathbf{A}$ and an $(m \times n)$ matrix $\mathbf{B}$ it holds that $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$. In addition, we use that $\mathbf{L}'\mathbf{L} = I_p$ and $\mathbf{P}'\mathbf{P} = \mathbf{P}$. As the second term of the objective function (which is $\text{tr}(\mathbf{L}'\mathbf{X}'\mathbf{P}\mathbf{X}\mathbf{L})$) is the variance between clusters in the reduced space, we see that minimizing the objective function $f(\mathbf{Z}, \mathbf{L})$ with respect to $\mathbf{Z}$ and $\mathbf{L}$ corresponds to maximizing the variance between clusters. The objective function can be minimized by performing alternating least-squares (ALS). For a given allocation of players to clusters (which is given by $\mathbf{Z}$), the loadings $\mathbf{L}$ can be found by performing an eigendecomposition on $\mathbf{X}'\mathbf{P}\mathbf{X}$, and taking the orthonormal eigenvectors that correspond to the p largest eigenvalues. For a given loadings matrix $\mathbf{L}$, maximizing $\text{tr}(\mathbf{L}'\mathbf{X}'\mathbf{P}\mathbf{X}\mathbf{L})$ with respect to $\mathbf{Z}$ comes down to performing k-means on the data in the reduced space, i.e., $\mathbf{X}\mathbf{L}$. If we have an initial allocation of players to clusters given by $\mathbf{Z}$, we can iteratively perform the two steps for updating $\mathbf{L}$ and $\mathbf{Z}$ until convergence. Convergence is reached when the allocation of players to clusters ($\mathbf{Z}$) does not change anymore. The ALS algorithm is given in Algorithm 1, which can be found in Appendix D.

In the ALS algorithm, we use the clustering method k-means to update the allocation of players to clusters (Hartigan and Wong, 1979). This k-means algorithm divides objects (players in our case) into k clusters. First, k initial centroids for the clusters are obtained by randomly selecting k objects from the data set without replacement. The algorithm then calculates for each object the distance to each of the k centroids, and each object gets assigned to the cluster that corresponds to the centroid that is closest. This gives us k clusters of objects. The centroids of the clusters are then updated by calculating the mean of the objects in each cluster. This procedure of assigning objects to clusters and updating the cluster centroids is iteratively repeated until convergence, that is, until the allocations of objects to clusters does not change anymore.

4.4 Evaluation of results

After performing the two approaches discussed in Section 4.3 to cluster football players in terms of their play styles, for each position group we obtain two clusterings (one for each approach). We want to compare the two approaches to evaluate which one performs best. In order to compare two clusterings for a position group, we use expert opinion to choose some players from that position group to represent a certain play style. In this way, for a given position group we have several pre-defined play styles, and for each of these play styles we have some players who represent that play style. Important to note here is that for each position group, only part of the players in that position group are chosen to represent one of the pre-defined play styles, which means that for the remaining players we do not know to which of the pre-defined play styles they belong. For each of the resulting clusters of a clustering we can compute the centroid, which represents the average profile of a cluster. The quality of the resulting clusters of a clustering can then be evaluated by

- The accuracy, which is the percentage of the representing players who end up in the cluster that has the highest resemblance to their pre-defined play style;
- The similarity, which reflects how similar a player who represents a certain play style is to the cluster that has the highest resemblance to the corresponding pre-defined play style.

For the second evaluation step we calculate for each player who represents a certain play style how similar he is to the cluster that has the highest resemblance to the corresponding pre-defined play style. This is done by first calculating for each player who represents a certain play style what his distance is to the centroid of each cluster in the p-dimensional subspace of the columns of $\mathbf{X}$ (for this we use Euclidean distance). If we have k clusters, then the distances to the centroids of the k clusters for a player can be denoted by $d_1, \ldots, d_k$. If that player then represents cluster i ($i \in \{1, ..., k\}$), we calculate the similarity of that player to cluster i as

$$S_i = 1 - \frac{d_i}{\sum_{j=1}^{k} d_j}.$$

Compared to the accuracy measure, the advantage of the similarity measure is that it gives an indication of how close a representing player is to the centroid of the cluster that has the highest resemblance to the play style that he represents, relative to how close that player is to the centroids of the other clusters. Hence, the similarity measure gives us additional information for evaluating the results, as opposed to just evaluating whether a representing player ends up in the right cluster. Note that it if we have two players from the same cluster i for which the first player is closer to the centroid of cluster i than the second player, it could still be the case that the second player has a higher similarity to cluster i than the first player. The reason for this is that even though the second player is further away from the centroid than the first player, this could also mean that the second player is further away from the centroids of the other clusters.

In this thesis we do not evaluate the results for all six position groups, as this would become rather extensive. Instead, we focus on the two position groups for which we consider the identification of play styles to be most relevant and interesting, which are central midfielders and centre-forwards. Hence, for central midfielders and centre-forwards, we must define play styles and players to represent those play styles. In order to do this, we use the play styles that are proposed by Aalbers and Van Haaren (2018). These play styles are based on extensive sports-media research and the game Football Manager 2018, which is often acclaimed for its realism, and is used by professional football clubs to recruit players (Sullivan, 2016). For each of these play styles, Aalbers and Van Haaren (2018) and Quint (2020) provide representing players, which we use as well. In addition, we do a sports-media research ourselves to choose more players to represent certain play styles (also, some of the provided players by Aalbers and Van Haaren (2018) and Quint (2020) are not in our data set due to the use of data from different seasons). The resulting pre-defined play styles with their representing players for central midfielders and centre-forwards are given in Sections 4.4.1 and 4.4.2, respectively. The representing players that are provided by Aalbers and Van Haaren (2018) and Quint (2020) are marked with an asterisk, and for the other representing players the references are given.

4.4.1 Central midfielders

For central midfielders we use the five pre-defined play styles with representing football players as given below.

* **Ball-winning midfielder:** A midfielder with this play style mainly focuses on regaining possession of the ball. When the opposing team is in possession of the ball, this player defends actively and aggressively by trying to close down opponents and cut off pass supply lines. This player tries to disturb the build-up of the opposing team, and occasionally makes strategic fouls such that his own team is able to reorganize. When in possession of the ball or after gaining possession, this player mainly plays simple passes.
  Representing players: N'Golo Kanté* (Chelsea), Casemiro* (Real Madrid), Giovani Lo Celso* (Real Betis), Remo Freuler* (Atalanta Bergamo), Konrad Laimer* (RB Leipzig), Mahmoud Dahoud* (Borussia Dortmund), Moussa Sissoko* (Tottenham Hotspur), Fernandinho (Manchester City) and Pierre Højbjerg (Southampton) (Beuvink, 2018; Bodell, 2020).

* **Holding midfielder:** A midfielder with this play style mainly focuses on protecting the defensive line. When the opposing team is in possession of the ball, this player defends passively, tries to keep the defensive line compact, reduces space in front of the defense and tries to shadow the attacking midfielders of the opponent. When the team is in possession of the ball, this player dictates the pace of the game.
  Representing players: Sergio Busquets* (FC Barcelona), Nemanja Matić* (Manchester United), Wilfred Ndidi* (Leicester City), Allan* (Napoli), Luka Milivojević* (Crystal Palace), Javi Martínez* (Bayern München), Rodri (Atletico Madrid), Fabinho (Liverpool) and Thomas Partey (Atletico Madrid) (Wright, 2019; Zavala, 2020; Bate, 2019; Kelly, 2019; Wright, 2020).

* **Deep-lying playmaker:** A midfielder with this play style mainly focuses on dictating the pace of the game, creating chances for teammates to score goals and exploiting the space in front of his team's defense. This player has excellent vision and timing, is technically gifted and has accurate passing skills to potentially cover longer distances with his passes as well. Hence, this player is more focused on the build-up play than on defending.
  Representing players: Jorginho* (Chelsea), Cesc Fàbregas* (AS Monaco), Granit Xhaka* (Arsenal), Thiago Alcántara* (Bayern München), Marco Verratti* (PSG), Santi Cazorla* (Villareal CF), João Moutinho (Wolverhampton), Dani Parejo (Valencia CF) and Miralem Pjanić (Juventus) (Baldi, 2019; Roden, 2019; Achanta, 2020; Mukherjee, 2020).

* **Box-to-box midfielder:** A midfielder with this play style is a dynamic player, whose main focus is on excellent positioning, both defensively and offensively. When the opposing team is in possession of the ball, this player concentrates on breaking up the play and guarding the defensive line. When in possession of the ball, this player often dribbles forward to then pass the ball to teammates higher up the pitch, and he often arrives late in the penalty area of the opposing team to create chances.

Representing players: Georginio Wijnaldum* (Liverpool), Blaise Matuidi* (Juventus), Arturo Vidal* (FC Barcelona), Leon Goretzka* (Bayern München), Franck Kessié* (AC Milan), Emre Can* (Juventus), Maraoune Fellaini (Manchester United), Weston McKennie (Schalke 04) and Sergej Milinković-Savić (Lazio Roma) (Parker, 2018; Shelat, 2020; Cooper, 2020; Pearson, 2020).

* **Advanced playmaker:** A midfielder with this play style is the prime creator of the team, by trying to occupy space between the midfield and defensive line of the opposing team. This player is technically skilled, has good passing skills, can hold up the ball and has excellent vision and timing. A midfielder with this play style tries to bring teammates in good scoring positions by giving perfectly timed through passes.
Representing players: Kevin de Bruyne* (Manchester City), Luis Alberto* (Lazio Roma), Christian Eriksen* (Tottenham Hotspur), Gylfi Sigurdsson* (Everton), Mario Götze* (Borussia Dortmund), James Rodríguez* (Bayern München), Hakan Çalhanoğlu (AC Milan), James Maddison (Leicester City), Nabil Fekir (Olympique Lyon), David Silva (Manchester City), Mesut Özil (Arsenal) and Isco (Real Madrid) (Sengupta, 2019; El-Shaboury, 2019; Agate, 2019; Chambers, 2020c; Kaynak, 2019; Fitzpatrick, 2019).

4.4.2 Centre-forwards

For centre-forwards we use the four pre-defined play styles with representing football players as given below.

* **Shadow striker:** A centre-forward with this play style mainly gives short passes to teammates, which are often into the final third. Compared to other centre-forwards, this player has a high amount of shots from outside the penalty area of the opposing team, and he mainly operates from outside the penalty area of the opposing team. When the opposing team is in possession of the ball, this player rarely presses high up the pitch, such that he has enough energy to be able to contribute offensively when his own team is in possession of the ball.
Representing players: Kai Havertz* (Bayer Leverkusen), Andrej Kramarić* (Hoffenheim), Antoine Griezmann* (Atletico Madrid), Roberto Firmino* (Liverpool), Heung-Min Son* (Tottenham Hotspur), Marco Reus* (Borussia Dortmund), Paulo Dybala (Juventus) and Memphis Depay (Olympique Lyon) (Zavala, 2019; Macdonald, 2020; Kircher, 2020).

* **Target man:** In general, a centre-forward with this play style does not shoot a lot, but the shots he does produce are usually headers or shots from receiving crosses. This player operates less inside the penalty area of the opposing team than other centre-forwards. When the opposing team is in possession of the ball, a player with this play style presses the opposition high up the pitch and tends to have a relative high number of tackles, fouls and duels.
Representing players: Dominic Calvert-Lewin* (Everton), Álvaro Morata* (Chelsea and Atletico Madrid), Mario Mandžukić* (Juventus), Diego Costa (Atletico Madrid), Sébastien Haller (Eintracht Frankfurt), Shane Long (Southampton), Aleksandar Mitrović (Fulham)

and Patrik Schick (AS Roma) (Carlisle, 2018; Sandford, 2019; Jackson, 2020; Bourgeois, 2019; Harris, 2018; Jacob, 2019).

* **Poacher:** A player with this play style mainly operates inside the penalty area of the opposing team, where he often receives the ball to try to shoot. When this player receives a cross from a teammate, he also tries to shoot immediately. This player is not likely to create chances for teammates to score. In addition, when the opposing team is in possession of the ball, a player with this play style tends to have a low contribution in terms of defensive actions, such as pressing the opposition high up the pitch or making tackles and dueling the opposition.
  Representing players: Romelu Lukaku* (Manchester United), Edin Džeko* (AS Roma), Yussuf Poulsen* (RB Leipzig), Mauro Icardi (Inter Milan) and Radamel Falcao (AS Monaco) (Elliott, 2018; Warrier, 2018).

* **Mobile striker:** A player with this play style tends to be involved in the build-up play, which he does by receiving the ball, dribbling with the ball in the final third and shooting from inside the penalty area of the opposing team. This player regularly gets in a position lower on the pitch to receive possession, but he is not likely to be involved in defensive actions.
  Representing players: Robert Lewandowski* (Bayern München), Gabriel Jesus* (Manchester City), Harry Kane* (Tottenham Hotspur), Joshua King* (Bournemouth), Pierre-Emerick Aubameyang* (Arsenal), Wissam Ben Yedder (Sevilla FC), Ciro Immobile (Lazio Roma) and Jamie Vardy (Leicester City) (Bliss, 2019; Chambers, 2020a,b; Smith, 2020).

5 Results

The results obtained by implementing the methods described in Section 4 are presented in this section. All used methods are implemented in R (R Core Team, 2020). The k-nearest neighbor imputation is implemented by the function `impute.knn` from the package "impute" (Hastie et al., 2020). We use the function `prcomp` from the package "stats" (R Core Team, 2020) to implement PCA, and the function `hcut` from the package "factoextra" (Kassambara and Mundt, 2020) is used to implement hierarchical clustering. For the implementation of Reduced k-means, we use the function `cluspca` from the package "clustrd" (Markos et al., 2019). The function `cluspca` uses the algorithm by Hartigan and Wong (1979) for the part where the k-means algorithm is applied. We only discuss the results for central midfielders and centre-forwards in this thesis, as discussing the results for all six position groups would become rather extensive. In Section 5.1 we present the results for central midfielders, and in Section 5.2 we present the results for centre-forwards. In terms of performance, the results of PCA and hierarchical clustering are compared to the results of Reduced k-means in Section 5.3.

5.1 Central midfielders

For the player statistics of central midfielders we have 16 missing values for the statistics `crossPassesPercentageCompleted`, `crossPassesPercentageToGoalAttempt`, `crossPassesPer-`

`centageLate` and `crossPassesPercentageHigh`. Furthermore, we have 7 missing values for the statistics `goalAttemptsPercentageInsidePenaltyBox`, `shotsPercentageInsidePenaltyBox` and `shotsPercentageWithHead`. Since we have 519 central midfielders and 51 statistics, the number of missing values is not too high in order for it to have a significant influence on the results. Each of the missing values is replaced by using k-nearest neighbor imputation with $k = 10$.

The results that follow from applying PCA and hierarchical clustering for central midfielders are presented in Section 5.1.1, and in Section 5.1.2 we present the results that follow from applying Reduced k-means for central midfielders.

5.1.1 PCA and hierarchical clustering

We have applied PCA to the $(N \times q)$ data matrix $\mathbf{X}$ for central midfielders to reduce the dimension q. For central midfielders we have $N = 519$ and $q = 51$, meaning that we have 519 central midfielders and 51 statistics. Having applied PCA, we should decide how many principal components to retain. One way to do this is by using Kaiser's rule (Kaiser, 1960). When using this rule, one only retains the principal components that have an eigenvalue greater than 1. The idea behind this rule is that it would not make sense to retain a principal component that explains less variance than a single original variable. Another way to decide how many principal components to retain, is by using Parallel Analysis (Horn, 1965). With Parallel Analysis, one only retains the principal components for which the eigenvalues are larger than the 95^{th} percentile of a distribution of randomly generated eigenvalues, which are derived from random observations from a standard normal distribution. Raîche et al. (2013) proposed two additional methods to decide how many principal components to retain, which are both based on the scree plot of the principal components. A scree plot is a visualization of the eigenvalues of the principal components, with the eigenvalues ordered from high to low. This scree plot can be used to determine the number of principal components to retain, which is done by finding the "elbow" of the scree plot (Cattell, 1966). The "elbow" of a scree plot is the point at which the slope of the curve changes most drastically. The two methods proposed by Raîche et al. (2013) are both numerical solutions to find the "elbow" of a scree plot, and they are called Optimal Coordinates and the Acceleration Factor. Applying PCA to the statistics of central midfielders yields the scree plot that is displayed in Figure 3, with also the suggested number of principal components to retain for the four discussed methods. In Table 4 we show the percentages of how much of the variance of the original variables is explained by the first 10 principal components. As the most suggested number of principal components to retain is 4 (see Figure 3), we decide to retain the first four principal components. Hence, 53.4% of the variance of the original variables is retained.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Percentage of explained variance	26.5	12.7	9.2	5.0	3.9	3.8	3.3	2.8	2.5	2.4
Cumulative percentage of explained variance	26.5	39.2	48.4	53.4	57.3	61.0	64.3	67.1	69.7	72.1

Table 4: Percentages of explained variance for the first 10 principal components.

Interpreting the retained principal components can be hard, as it can be that the principal

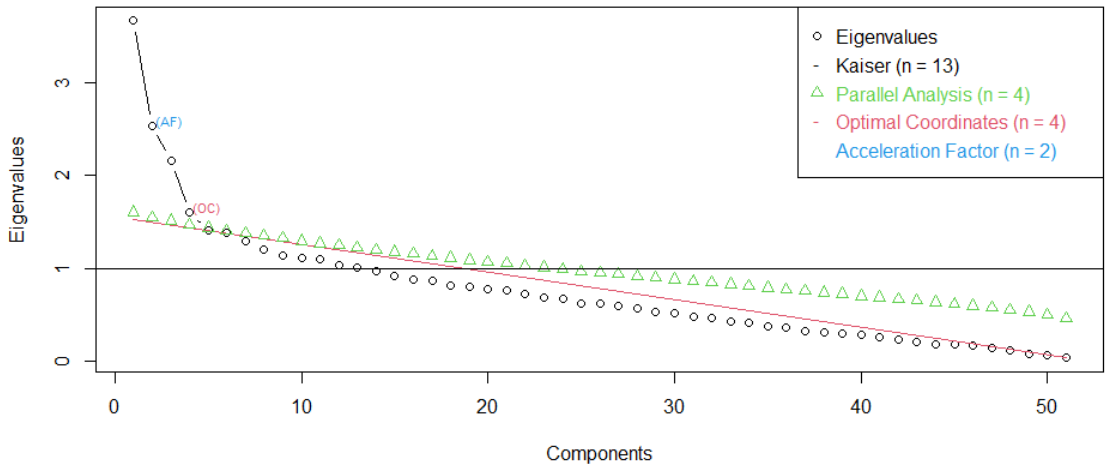

Figure 3: Scree plot of the principal components found for central midfielders.

components have high loadings (in absolute terms) for many variables. Hence, we use varimax rotation to rotate the retained principal components (Kaiser, 1958). Varimax rotation is a method that rotates the axes of the principal components, such that the sum of the variances of the squared loadings are maximized (the variance of the squared loadings is taken for each component). For each component, this should result in few of the original variables having high loadings, and the loadings of the other variables being closer to zero (compared to the non-rotated components), making the components easier to interpret. Important to note is that the rotated components are only used to interpret the principal components, and to interpret the found clusters after applying hierarchical clustering. Hence, for hierarchical clustering and for calculating the similarities of players to clusters (see Section 4.4), we use the non-rotated retained principal components.

The loadings of the four varimax rotated principal components for central midfielders can be found in Table 6, and in Table 5 the variables to which the loadings belong are displayed. In Appendix E.1.1, the 10 highest loadings (in absolute terms) are given for each varimax rotated principal component, which can be used for easier interpretation of the components. The interpretation of the four varimax rotated principal components is given below.

* **PC1** (creating chances and duel weakness): can be interpreted as how likely a central midfielders is to create chances for teammates to score goals, and how weak he is in duels.

* **PC2** (involvement, risky passing and operating outside penalty box): can be interpreted as how involved a central midfielder is in the play of his team when they are in possession of the ball, how risky his passing is, and how much he operates outside of the penalty area of the opposing team. Risky passes are passes that are typically forward and/or long.

* **PC3** (simple passing): can be interpreted as how simple the passes of a player are. Simple passes are passes that are typically short and either wide or backwards.

* **PC4** ((involvement in) goal attempts and duel strength): can be interpreted as how likely

a central midfielder is to make a goal attempt or to be involved in the possession moment leading to a goal attempt, and how strong he is in duels.

Variable	Variable name	Variable	Variable name
V1	shareInBallActionsPercentage	V27	groundDuelsPercentageWon
V2	shareInPassesPercentage	V28	airDuelsPercentageWon
V3	passesPercentageCompleted	V29	defensiveDuelsPercentageWon
V4	passesPercentageForward	V30	defensiveDuelsPercentageOwnHalf
V5	passesForwardPercentage-Completed	V31	attackingDuelsPercentageWon
V6	passesPercentageWide	V32	keyActionsPerBallAction-Percentage
V7	passesPercentageLong	V33	shareInKeyActionsPercentage
V8	keyPassesPerBallAction-Percentage	V34	goalAttemptsPerBallAction-Percentage
V9	shareInKeyPassesPercentage	V35	shareInGoalAttemptsPercentage
V10	passesPercentageOpponentHalf	V36	goalAttemptsPercentageInside-PenaltyBox
V11	passesOwnHalfPercentage-Completed	V37	shotsPerBallActionPercentage
V12	passesOpponentHalfPercentage-Completed	V38	shareInShotsPercentage
V13	passesPercentageFinalThird	V39	shotsPercentageInsidePenaltyBox
V14	passesFinalThirdPercentage-Completed	V40	shotsPercentageWithHead
V15	shareInPassFirstInPossession-Percentage	V41	shareInPossessionWithGoal-AttemptsPercentage
V16	passFirstInPossession-PercentageForward	V42	shareInPossessionWithGoals-Percentage
V17	shareInReceivedFirstPassIn-PossessionPercentage	V43	possessionLossPerBallAction-Percentage
V18	passesPercentageToBox	V44	possessionRegainInPlayPer-BallActionPercentage
V19	passesPercentageCrosses	V45	possessionRegainInPlay-PercentageByInterception
V20	crossPassesPercentageCompleted	V46	possessionRegainInPlay-PercentageOpponentHalf
V21	crossPassesPercentageToGoal-Attempt	V47	foulsPerBallActionPercentage
V22	crossPassesPercentageLate	V48	foulsPercentageOwnHalf
V23	crossPassesPercentageHigh	V49	foulsSufferedPerBallAction-Percentage
V24	duelsPercentageWon	V50	cardsPerFoul
V25	dribblesPerBallAction-Percentage	V51	actionsPercentageInOpponentBox
V26	slidingsPerDuel		

Table 5: Variables for central midfielders.

After performing PCA, we apply hierarchical clustering to the 4 principal components that we found for the 519 central midfielders that we have. We consider three different linkage criteria for hierarchical clustering, which are complete linkage, average linkage and Ward linkage. To decide on the number of clusters that we want to obtain, we can use the number of play styles that are defined by Aalbers and Van Haaren (2018). However, we also look at the cluster dendogram, which can be used to decide on the number of clusters that we want to obtain, by cutting the tree at the height at which the difference in heights is largest. In addition, we consider the average

	PC1	PC2	PC3	PC4		PC1	PC2	PC3	PC4		PC1	PC2	PC3	PC4
V1	0.01	**0.32**	0.03	0.09	V18	0.18	0.06	-0.18	0.07	V35	0.00	0.02	-0.08	**0.37**
V2	0.02	**0.32**	0.05	0.00	V19	0.11	0.02	-0.17	0.06	V36	-0.02	**-0.23**	0.02	0.07
V3	0.02	0.01	**0.35**	-0.04	V20	0.07	0.06	0.01	-0.02	V37	-0.01	-0.11	-0.05	**0.33**
V4	-0.05	**0.23**	**-0.21**	-0.09	V21	0.12	0.04	-0.02	-0.05	V38	-0.02	-0.01	-0.07	**0.37**
V5	0.03	0.03	**0.33**	-0.04	V22	0.03	-0.14	0.03	0.07	V39	-0.02	**-0.23**	0.02	0.07
V6	-0.07	-0.03	**0.24**	0.04	V23	-0.02	0.14	-0.01	0.00	V40	-0.16	-0.17	0.00	0.09
V7	-0.10	**0.24**	-0.19	0.01	V24	**-0.27**	0.13	0.11	0.17	V41	0.14	**0.22**	0.07	**0.26**
V8	**0.28**	0.01	0.00	0.03	V25	**0.20**	0.01	0.10	0.03	V42	0.10	0.10	0.04	**0.21**
V9	**0.28**	0.12	-0.03	0.02	V26	-0.04	0.06	-0.01	-0.08	V43	0.05	-0.10	-0.24	0.07
V10	**0.22**	-0.11	0.02	0.11	V27	**-0.22**	0.12	0.12	0.18	V44	**-0.23**	0.02	-0.08	-0.13
V11	0.04	0.00	**0.33**	-0.02	V28	**-0.22**	0.05	0.00	0.05	V45	0.05	0.09	0.03	-0.02
V12	0.04	0.01	**0.34**	-0.03	V29	**-0.21**	0.06	0.05	**0.25**	V46	0.16	-0.08	0.08	-0.01
V13	**0.21**	-0.12	-0.02	0.12	V30	**-0.21**	0.09	-0.03	0.05	V47	-0.10	-0.15	-0.16	-0.11
V14	0.04	0.00	**0.32**	0.00	V31	-0.17	0.10	0.12	0.19	V48	-0.09	0.08	0.00	-0.04
V15	-0.03	**0.26**	0.02	0.02	V32	**0.29**	0.08	-0.03	0.03	V49	-0.08	-0.06	-0.07	0.17
V16	-0.08	0.17	**-0.20**	-0.15	V33	**0.27**	0.19	-0.04	0.04	V50	-0.03	0.07	0.01	0.01
V17	0.07	**0.29**	0.01	0.11	V34	0.00	-0.08	-0.06	**0.33**	V51	0.05	**-0.21**	-0.04	0.19

Table 6: Loadings of the varimax rotated principal components for central midfielders. The loadings that are greater than or equal to 0.2 (in absolute terms) are shown in bold.

silhouette width (ASW) (Rousseeuw, 1987) and the Caliński-Harabasz (CH) index (Caliński and Harabasz, 1974) for different number of clusters to decide on the number of clusters. The ASW indicates how compact the found clusters are, and whether they are well separated. The ASW takes values between -1 and 1. The CH index is defined as the between-cluster variance divided by the within-cluster variance, and it is corrected by the number of clusters. The CH index takes values between 0 and ∞. For both the ASW and the CH index, it holds that in general high values indicate well separated clusters.

When applying hierarchical clustering with complete linkage or average linkage, we find that there is one cluster that contains approximately 99% of the central midfielders, and all other clusters only contain less than 1% of the central midfielders. On the other hand, applying hierarchical clustering with Ward linkage provides us with a clustering of central midfielders in which the players are more evenly distributed over the clusters. Hence, we decide to use Ward linkage, which gives the cluster dendrogram that is shown in Figure 4. The values of the ASW and the CH index for 2 until 7 clusters are displayed in Table 7.

Number of clusters	ASW	CH index
2	0.318	231
3	0.170	170
4	0.172	147
5	0.166	134
6	0.162	129
7	0.159	123

Table 7: Values of the average silhouette width (ASW) and the Caliński-Harabasz (CH) index for 2 until 7 clusters for central midfielders.

The dendrogram, ASW and CH index each indicate that it would be optimal to set the number of clusters to 2. However, given that Aalbers and Van Haaren (2018) have defined 5 play styles for central midfielders, we have strong reason to believe that there are more than 2 play styles for central midfielders. As the dendrogram, ASW and CH index do not clearly indicate which

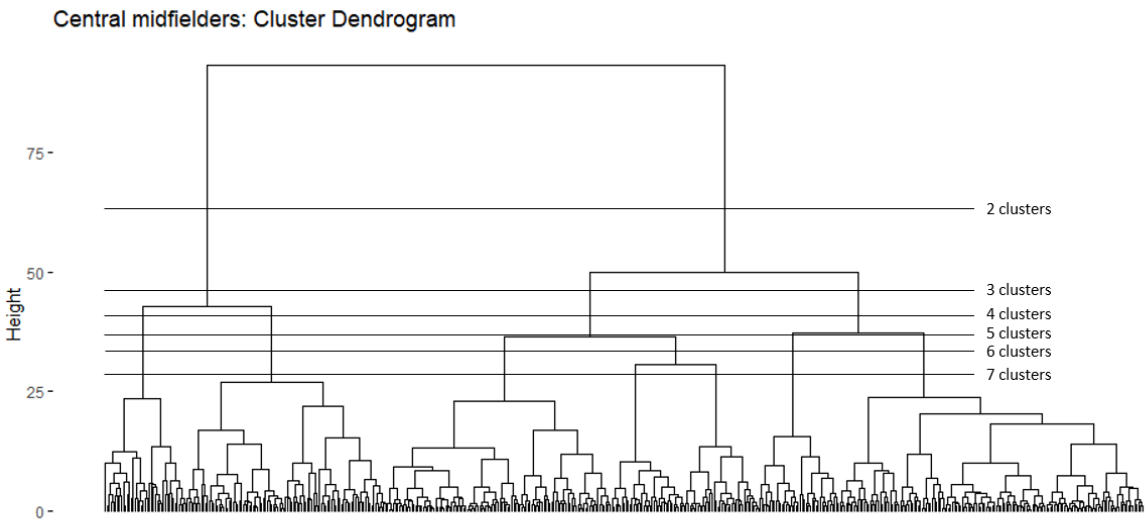

Figure 4: Cluster dendrogram for central midfielders that follows from applying hierarchical clustering with Ward linkage.

number of clusters would be optimal when considering more than 2 clusters, we decide to set the number of clusters to 5. In Figure 5, the resulting clusters are plotted based on the first two varimax rotated principal components, and for the the central midfielders who are chosen as representing players the names are displayed.

To interpret the found clusters, we calculate the average profile of each cluster. An average profile of a cluster is the centroid of that cluster, and it represents the average player of a cluster. The average profiles of the 5 clusters of central midfielders are displayed in Table 8.

From the average profiles we notice that there is one cluster that stands out in particular, which is the fifth cluster. In the dendrogram that is shown in Figure 4, this is the most left cluster. Compared to the other clusters, the fifth cluster scores high on the variables `passesPercentageFinalThird`, `goalAttemptsPercentageInsidePenaltyBox` and `actionsPercentageInOpponentBox`. When we take a look at the players who are in the fifth cluster, we see that it contains players such as Danny Ings (Southampton), Mario Mandžukić (Juventus), Radamel Falcao (AS Monaco), Rodrigo (Valencia CF), Yussuf Poulsen (RB Leipzig), Troy Deeney (Watford), Joshua King (Bournemouth), Aleksandar Mitrović (Fulham), Sébastien Haller (Eintracht Frankfurt) and Roberto Firmino (Liverpool). These are all players who usually tend to play as centre-forwards, which explains why the average profile of the fifth cluster scores high on the variables `passesPercentageFinalThird`, `goalAttemptsPercentageInsidePenaltyBox` and `actionsPercentageInOpponentBox`. This interpretation of the fifth cluster can mean two things: the players from this cluster who usually tend to play as centre-forwards play defensively compared to other centre-forwards, or our approach for determining players' playing positions can be improved. For example, our approach for determining players' playing positions can possibly be improved by using median locations of players instead of average locations (since the median is a more robust statistic than the mean). It could also help to take into account the standard deviations of players' locations in matches, in both the x-direction and the y-direction. Either way, we have found a cluster that represents an additional play style compared to the 5 play

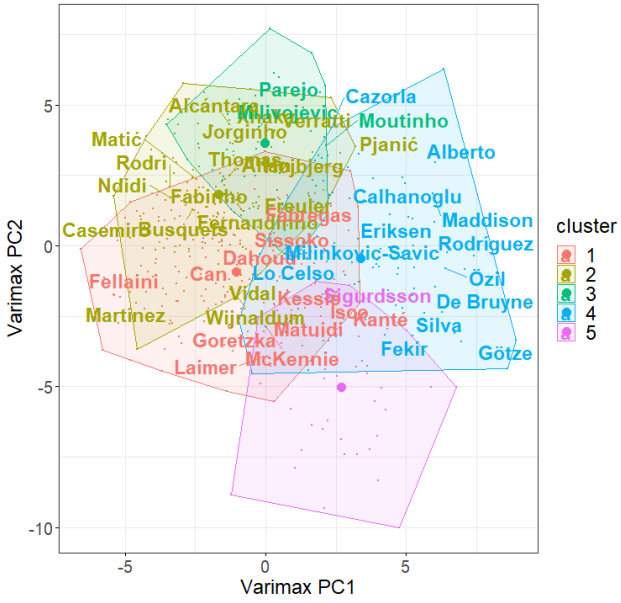

Figure 5: Plot of the found clusters based on the first two varimax rotated principal components for central midfielders, after applying hierarchical clustering with the number of clusters set to 5.

	Cl. 1	Cl. 2	Cl. 3	Cl. 4	Cl. 5		Cl. 1	Cl. 2	Cl. 3	Cl. 4	Cl. 5
V1	9	10	11	9	7	V27	50	57	53	48	45
V2	9	11	12	10	6	V28	47	53	50	36	33
V3	83	88	80	81	75	V29	56	60	58	55	57
V4	52	52	58	51	47	V30	65	66	66	58	57
V5	76	82	73	74	65	V31	41	50	45	40	38
V6	18	20	17	16	16	V32	1	1	2	2	2
V7	13	14	19	12	10	V33	7	8	15	16	9
V8	1	1	1	2	2	V34	1	1	1	2	4
V9	8	8	12	15	10	V35	7	6	9	12	16
V10	52	49	49	64	70	V36	47	33	23	46	70
V11	85	89	82	64	77	V37	1	1	1	2	4
V12	81	87	78	79	73	V38	7	6	7	12	16
V13	21	17	19	34	40	V39	47	33	24	48	71
V14	75	82	72	74	71	V40	15	10	4	7	22
V15	9	12	12	9	8	V41	34	41	47	47	40
V16	53	53	60	50	43	V42	32	37	44	48	46
V17	9	10	12	10	8	V43	10	7	10	12	18
V18	4	3	6	8	7	V44	9	8	8	6	5
V19	2	1	2	3	3	V45	61	63	65	61	60
V20	21	26	26	26	24	V46	32	32	31	37	39
V21	9	10	12	14	12	V47	3	2	2	2	3
V22	58	47	43	62	74	V48	50	51	53	41	35
V23	60	66	69	56	46	V49	2	2	2	2	3
V24	50	56	53	47	43	V50	0	0	0	0	0
V25	4	4	4	6	4	V51	2	1	1	4	8
V26	4	5	6	4	2						

Table 8: Average profiles for the 5 clusters of central midfielders that are found by subsequently applying PCA and hierarchical clustering.

styles defined by Aalbers and Van Haaren (2018). Hence, we take one step down in the hierarchical clustering that we applied to the 4 principal components that are retained for central midfielders, such that we get 6 clusters instead of 5. We again plot the resulting clusters based on the first two varimax rotated principal components (see Figure 6), and for the the central midfielders who are chosen as representing players the names are displayed. The average profiles of the 6 clusters are displayed in Table 9.

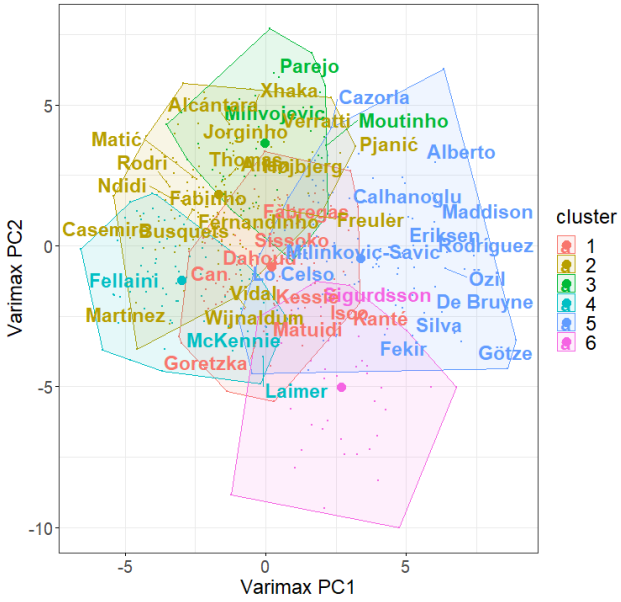

Figure 6: Plot of the found clusters based on the first two varimax rotated principal components for central midfielders, after applying hierarchical clustering with the number of clusters set to 6.

The cluster consisting of players who usually tend to play as centre-forwards has stayed exactly the same, and it is now the sixth cluster. If we interpret each of the other clusters as well and relate them to the play styles defined in Section 4.4.1, we get the interpretation of the clusters as given below. Note that we have ordered the clusters such that they are in the same order as the play styles defined in Section 4.4.1.

* **Cluster 1: 112 central midfielders (22%)**
  The average profile of this cluster scores low on `passesPercentageForward`, `passesPercentageWide` and `passesPercentageLong`, and above average on `possessionRegainInPlayPerBallActionPercentage` and `possessionRegainInPlayPercentageOpponentHalf`. Hence, players from this cluster mainly play simple passes when they are in possession of the ball, and when not in possession of the ball, they actively try to regain possession. This cluster therefore most closely resembles ball-winning midfielders.

* **Cluster 2: 152 central midfielders (29%)**
  The average profile of this cluster scores low on `passesPercentageFinalThird` and `actionsPercentageInOpponentBox`, and high on `shareInPassFirstInPossessionPercentage`. This indicates that players from this cluster do not play offensively, and that they have an important role in determining what the team is going to do when they gain possession of the ball. As a result, this cluster is most closely related to holding midfielders.

	Cl. 1	Cl. 2	Cl. 3	Cl. 4	Cl. 5	Cl. 6		Cl. 1	Cl. 2	Cl. 3	Cl. 4	Cl. 5	Cl. 6
V1	9	10	11	8	9	7	V27	49	57	53	52	48	45
V2	10	11	12	9	10	6	V28	44	53	50	53	36	33
V3	85	88	80	80	81	75	V29	55	60	58	58	55	57
V4	51	52	58	53	51	47	V30	63	66	66	69	58	57
V5	78	82	73	72	74	65	V31	40	50	45	43	40	38
V6	18	20	17	18	16	16	V32	1	1	2	1	2	2
V7	13	14	19	14	12	10	V33	9	8	15	5	16	9
V8	1	1	1	1	2	2	V34	1	1	1	1	2	4
V9	9	8	12	5	15	10	V35	7	6	9	6	12	16
V10	55	49	49	47	64	70	V36	44	33	23	51	46	70
V11	87	89	82	82	84	77	V37	1	1	1	1	2	4
V12	83	87	78	77	79	73	V38	7	6	7	6	12	16
V13	23	17	19	18	34	40	V39	45	33	24	51	48	71
V14	78	82	72	71	74	71	V40	11	10	4	21	7	22
V15	10	12	12	9	9	8	V41	37	41	47	30	47	40
V16	51	53	60	57	50	43	V42	34	37	44	28	48	46
V17	9	10	12	8	10	8	V43	9	7	10	10	12	18
V18	5	3	6	4	8	7	V44	8	8	8	10	6	5
V19	2	1	2	2	3	3	V45	61	63	65	60	61	60
V20	22	26	26	20	26	24	V46	34	32	31	28	37	39
V21	10	10	12	7	14	12	V47	2	2	2	3	2	3
V22	59	47	43	55	62	74	V48	49	51	53	51	41	35
V23	60	66	69	60	56	46	V49	2	2	2	2	2	3
V24	49	56	53	53	47	43	V50	0	0	0	0	0	0
V25	4	4	4	2	6	4	V51	2	1	1	2	4	8
V26	4	5	6	5	4	2							

Table 9: Average profiles for the 6 clusters of central midfielders that are found by subsequently applying PCA and hierarchical clustering.

* **Cluster 3: 40 central midfielders (8%)**
  The average profile of this cluster scores high on `shareInBallActionsPercentage`, `shareInPassesPercentage`, `passesPercentageForward`, `shareInKeyPassesPercentage` and `keyActionsPerBallActionPercentage`, indicating that players from this cluster are likely to be involved in the build-up play. In addition, they try to create chances for teammates to score goals. This cluster thus most most closely resembles deep-lying playmakers.

* **Cluster 4: 74 central midfielders (14%)**
  The average profile of this cluster scores low on `shareInKeyActionsPercentage` and `shareInPossessionWithGoalAttemptsPercentage`, which indicates that a player from this cluster is not very involved in the play of his team when they are in possession of the ball. Furthermore, he scores high on `passesPercentageForward` and `goalAttemptsPercentageInsidePenaltyBox`, meaning that he tries to pass the ball to teammates higher up the pitch and that he has a tendency to arrive late in the penalty area of the opposing team to make a goal attempt. This cluster can thus be related to box-to-box midfielders.

* **Cluster 5: 100 central midfielders (19%)**
  The average profile of this cluster scores high on `keyPassesPerBallActionPercentage`, `shareInKeyPassesPercentage`, `passesPercentageFinalThird`, `passesPercentageToBox`, `goalAttemptsPerBallActionPercentage`, `shareInGoalAttemptsPercentage`, `shareInPossessionWithGoalAttemptsPercentage` and `actionsPercentageInOpponentBox`. This

shows that players from this cluster tend to play offensively, and that they try to create chances for teammates to score goals. Hence, this cluster is most closely related to advanced playmakers.

* **Cluster 6: 41 central midfielders (8%)**
  This is the cluster of players who usually tend to play as centre-forwards.

Table 10 is a confusion table which shows how many of the representing players of each play style are assigned to each cluster. It also shows the accuracy and average similarity for each play style and in total. We notice that there are four representing players of ball-winning midfielders who end up in the cluster of holding midfielders. These players are Casemiro (Real Madrid), Remo Freuler (Atalanta Bergamo), Fernandinho (Manchester City) and Pierre Højbjerg (Southampton). Furthermore, Jorginho (Chelsea), Granit Xhaka (Arsenal), Thiago Alcántara (Bayern München), Marco Verratti (PSG) and Miralem Pjanić (Juventus) (who represent deep-lying playmakers) end up in the cluster of holding midfielders as well. Blaise Matuidi (Juventus), Leon Goretzka (Bayern München), Franck Kessié (AC Milan) and Emre Can (Juventus) end up in the cluster of ball-winning midfielders, while they represent box-to-box midfielders. Georginio Wijnaldum (Liverpool) and Arturo Vidal (FC Barcelona) also represent box-to-box midfielders, but they end up in the cluster of holding midfielders.

	Cluster								
Play style	1	2	3	4	5	6	Total	Accuracy (%)	Average similarity (%)
Ball-winning midfielders	**3**	4	0	1	1	0	9	33.3	87.5
Holding midfielders	0	**8**	1	0	0	0	9	88.9	92.7
Deep-lying playmakers	1	5	**2**	0	1	0	9	22.2	87.2
Box-to-box midfielders	4	2	0	**2**	1	0	9	22.2	81.5
Advanced playmakers	1	0	0	0	**10**	1	12	83.3	89.5
Total	9	19	3	3	13	1	48	52.1	87.8

Table 10: Confusion table which shows how many of the representing players of each play style are assigned to each cluster of central midfielders for PCA and hierarchical clustering. It also shows the accuracy and average similarity for each play style and in total.

5.1.2 Reduced k-means

We want to apply Reduced k-means to the $(N \times q)$ data matrix $\mathbf{X}$ for central midfielders to find k clusters of players in a p-dimensional subspace of the columns of $\mathbf{X}$. We however still have to decide what values we choose for k (the number of clusters) and p (the lower dimension). To distinguish between PCA and Reduced k-means, we call the p variables that are found by Reduced k-means reduced variable 1 (RV1), reduced variable 2 (RV2), and so forth (instead of PC1, PC2, and so forth).

For central midfielders we again have $N = 519$ and $q = 51$, meaning that we have 519 central midfielders and 51 statistics. To decide on the number of clusters, we again consider the number of play styles defined by Aalbers and Van Haaren (2018), the ASW and the CH index. For a given value of k, we decide on the value of p by the following rule of thumb. As we should have that $p < k$ (see Section 4.3.2), we initially take $p = k - 1$. We then take the lowest value of p for which the resulting allocation of players to clusters is exactly the same as for the case where $p = k - 1$. As a result, we get the values of the ASW and the CH index for 2 until 7 clusters that are displayed in Table 11.

Number of clusters (k)	Dimension (p)	ASW	CH index
2	1	0.614	1160
3	2	0.360	408
4	3	0.308	308
5	4	0.253	215
6	5	0.219	165
7	6	0.206	142

Table 11: Values of the average silhouette width (ASW) and the Caliński-Harabasz (CH) index for 2 until 7 clusters for central midfielders.

The ASW and the CH index both indicate that taking $k = 2$ would be optimal. However, since Aalbers and Van Haaren (2018) defined 5 play styles for central midfielders, we have strong reason to believe that there are more than 2 play styles for central midfielders. The ASW and the CH index suggest to take as few clusters as possible. As a result, we have a trade-off between taking as few clusters as possible and taking 5 clusters. As it is hard to make a decision for this trade-off, we choose to rely on the expert knowledge provided by Aalbers and Van Haaren (2018), and take 5 clusters. For $k = 5$, we get $p = 4$ according to our rule of thumb. As we did for the principal components found by PCA, we use varimax rotation to rotate the reduced variables, such that the reduced variables become easier to interpret. Again, these rotated reduced variables are only used for interpretation. After applying Reduced k-means with $k = 5$ and $p = 4$, we obtain the varimax rotated reduced variables of which the loadings are shown in Table 12. The 10 highest loadings (in absolute terms) for each varimax rotated reduced variable can be found in Appendix E.1.2, which can be used for easier interpretation of the variables. The interpretation of the four varimax rotated reduced variables is given below.

* **RV1** (offensive play and low defensive effort): can be interpreted as how offensive a central midfielder plays, in terms of where he positions himself on the football pitch, and how often he tries to score a goal. This variable also captures how low the defensive effort of a central midfielder is.

* **RV2** (involvement): can be interpreted as how involved a central midfielder is in the play of his team when they are in possession of the ball.

* **RV3** (dribbling and simple passing): can be interpreted as how much a central midfielder dribbles, and how simple his passes are. Simple passes are passes that are typically short and either wide or backwards.

∗ **RV4** (creating chances): can be interpreted as how likely a central midfielder is to create chances for teammates to score goals.

	RV1	RV2	RV3	RV4		RV1	RV2	RV3	RV4		RV1	RV2	RV3	RV4
V1	-0.13	**0.24**	**-0.21**	0.05	V18	0.19	0.11	0.06	-0.13	V35	0.19	0.10	-0.13	0.10
V2	-0.17	**0.24**	-0.19	-0.02	V19	0.16	0.04	-0.03	-0.16	V36	0.13	-0.17	0.02	-0.14
V3	-0.19	0.10	**0.22**	-0.02	V20	-0.02	0.07	0.08	0.13	V37	**0.21**	0.00	-0.03	**0.20**
V4	-0.08	0.03	-0.19	-0.19	V21	0.03	0.09	0.09	0.07	V38	0.19	0.07	-0.09	0.10
V5	-0.18	0.12	**0.21**	-0.03	V22	0.10	-0.03	-0.03	-0.17	V39	0.14	-0.17	0.01	-0.15
V6	-0.13	0.01	0.05	**0.20**	V23	-0.07	0.07	-0.15	-0.04	V40	0.04	**-0.22**	-0.08	0.04
V7	-0.08	0.06	**-0.36**	-0.12	V24	-0.19	-0.01	-0.03	0.10	V41	0.05	**0.32**	-0.12	0.02
V8	0.16	0.19	0.17	-0.15	V25	0.07	0.17	**0.33**	-0.17	V42	0.08	0.19	-0.13	0.07
V9	0.12	**0.25**	0.05	**-0.21**	V26	-0.08	-0.02	0.05	0.05	V43	**0.24**	-0.08	-0.17	**0.23**
V10	**0.21**	0.09	0.16	0.04	V27	-0.16	0.02	0.00	0.15	V44	-0.16	-0.19	-0.05	**-0.23**
V11	-0.15	0.09	**0.23**	-0.11	V28	-0.13	-0.08	-0.10	0.01	V45	-0.04	0.10	-0.01	0.09
V12	-0.17	0.11	**0.23**	0.09	V29	-0.06	-0.01	0.01	0.14	V46	0.09	0.05	0.01	0.13
V13	**0.24**	0.08	0.16	0.00	V30	-0.12	-0.07	-0.05	-0.03	V47	0.05	**-0.24**	-0.14	-0.13
V14	-0.13	0.11	**0.25**	0.19	V31	-0.12	0.04	0.01	**0.22**	V48	-0.10	-0.01	-0.04	-0.09
V15	-0.15	0.15	-0.15	0.18	V32	0.14	**0.25**	-0.02	-0.17	V49	0.11	-0.05	0.01	0.07
V16	-0.10	-0.04	-0.14	-0.19	V33	0.10	**0.31**	-0.15	**-0.21**	V50	-0.05	0.03	-0.05	0.03
V17	-0.05	**0.27**	**-0.22**	0.12	V34	**0.22**	0.03	-0.07	**0.21**	V51	**0.24**	-0.09	0.00	**0.29**

Table 12: Loadings of the varimax rotated reduced variables for central midfielders that follow from applying Reduced k-means with $k = 5$ and $p = 4$. The loadings that are greater than or equal to 0.2 (in absolute terms) are shown in bold.

In Figure 7, the resulting clusters are plotted based on the first two varimax rotated reduced variables, and for the central midfielders who are chosen as representing players the names are displayed. Furthermore, Reduced k-means with $k = 5$ and $p = 4$ gives us the average profiles of each cluster that are displayed in Table 13.

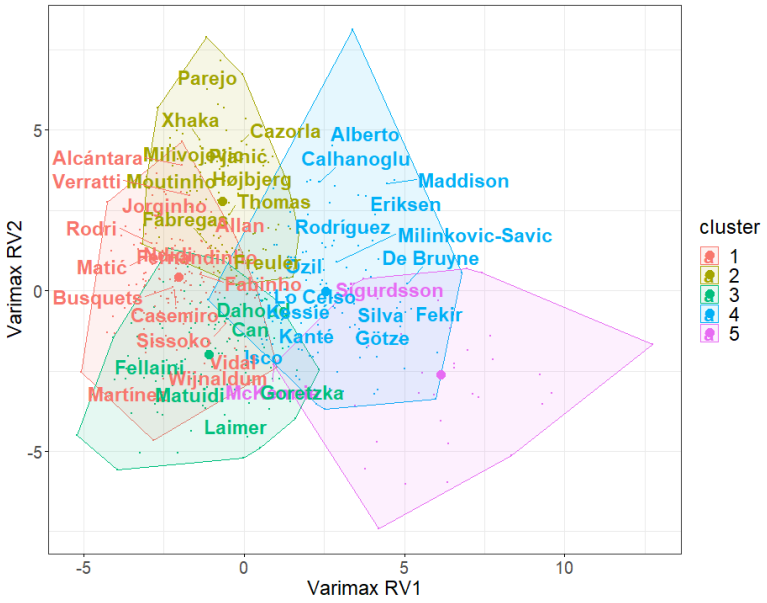

Figure 7: Plot of the found clusters based on the first two varimax rotated reduced variables for central midfielders, after applying Reduced k-means with $k = 5$ and $p = 4$.

As for PCA and hierarchical clustering, we find that there is one cluster that contains players who usually tend to play as centre-forwards. This can be recognized by the high score of the fifth clus-

	Cl. 1	Cl. 2	Cl. 3	Cl. 4	Cl. 5		Cl. 1	Cl. 2	Cl. 3	Cl. 4	Cl. 5
V1	10	11	8	9	7	V27	56	53	51	48	44
V2	11	12	9	9	6	V28	53	47	50	35	33
V3	88	84	82	81	72	V29	59	57	57	55	55
V4	52	54	52	50	47	V30	67	64	66	58	56
V5	82	78	74	74	62	V31	48	45	42	40	38
V6	20	18	18	17	16	V32	1	2	1	2	2
V7	13	16	13	12	11	V33	7	15	6	14	10
V8	1	1	1	2	1	V34	1	1	1	2	4
V9	7	12	7	15	10	V35	5	9	7	12	16
V10	48	52	51	65	69	V36	30	29	55	48	71
V11	89	86	84	84	75	V37	1	1	1	2	4
V12	87	83	79	80	71	V38	5	9	7	12	16
V13	16	21	21	34	40	V39	30	29	56	50	72
V14	82	77	73	75	68	V40	10	5	20	8	25
V15	11	12	9	9	8	V41	38	48	32	46	39
V16	54	54	55	49	45	V42	33	47	30	45	45
V17	10	12	8	10	8	V43	7	9	10	12	20
V18	3	5	4	7	8	V44	9	7	9	5	4
V19	1	2	2	3	4	V45	63	64	59	62	59
V20	27	26	19	25	22	V46	31	33	31	36	41
V21	10	12	7	14	12	V47	2	2	3	2	3
V22	44	52	60	64	72	V48	52	50	49	42	34
V23	64	71	58	55	45	V49	2	2	2	2	3
V24	56	52	51	46	42	V50	0	0	0	0	0
V25	4	4	3	6	4	V51	1	1	3	4	8
V26	6	4	5	3	2						

Table 13: Average profiles for the 5 clusters of central midfielders that are found by applying Reduced k-means with $k = 5$ and $p = 4$.

ter on variables such as `passesPercentageFinalThird`, `goalAttemptsPercentageInsidePenaltyBox` and `actionsPercentageInOpponentBox`. Also worth to note is that the fifth cluster of "False strikers" mostly contains the same players as before, who are Danny Ings (Southampton), Mario Mandžukić (Juventus), Radamel Falcao (AS Monaco), Rodrigo (Valencia CF), Yussuf Poulsen (RB Leipzig), Troy Deeney (Watford), Joshua King (Bournemouth), Aleksandar Mitrović (Fulham) and Sébastien Haller (Eintracht Frankfurt). We have thus again found a cluster that represents an additional play style compared to the 5 play styles defined by Aalbers and Van Haaren (2018). Therefore, we also again increase the number of clusters to 6, and by our rule of thumb the value of p becomes 5. As a result, we apply Reduced k-means with $k = 6$ and $p = 5$. This gives us the varimax rotated reduced variables of which the loadings are shown in Table 14. In Appendix E.1.2, the 10 highest loadings (in absolute terms) can be found for each varimax rotated reduced variable, which can be used for easier interpretation of the variables. The interpretation of the five varimax rotated reduced variables is given below.

* **RV1** (creating chances and dribbling): can be interpreted as how likely a central midfielder is to create chances for teammates to score goals, and how often he dribbles with the ball.

* **RV2** (involvement): can be interpreted as how involved a central midfielder is in the play of his team when they are in possession of the ball.

* **RV3** (simple passing): can be interpreted as how simple the passes of a central midfielder are. Simple passes are passes that are typically short and either wide or backwards.

* **RV4** (appearance in opponent box and likelihood of goal attempts): can be interpreted as how much a central midfielder appears in the penalty area of the opposing team, and how likely he is to make goal attempts.

* **RV5** (good early crossing and goal attempts mainly from outside the box): can be interpreted as how likely a central midfielder is to give good early crosses into the penalty area of the opposing team, and how much of his goal attempts are from outside the penalty area of the opposing team.

	RV1	RV2	RV3	RV4	RV5		RV1	RV2	RV3	RV4	RV5
V1	-0.13	**0.24**	-0.13	-0.15	0.12	V27	-0.17	0.03	0.01	-0.06	-0.15
V2	-0.17	**0.23**	-0.13	-0.10	0.13	V28	-0.13	-0.10	-0.11	-0.07	0.02
V3	**-0.20**	0.11	**0.28**	0.07	0.10	V29	-0.06	-0.01	-0.02	-0.02	-0.15
V4	-0.08	0.02	**-0.33**	0.07	-0.07	V30	-0.12	-0.08	-0.07	0.00	0.07
V5	-0.19	0.12	**0.23**	0.07	0.07	V31	-0.12	0.06	0.03	-0.15	**-0.22**
V6	-0.12	0.03	**0.23**	-0.19	0.09	V32	0.14	**0.25**	-0.05	0.11	0.11
V7	-0.08	0.05	**-0.30**	-0.05	**0.22**	V33	0.09	**0.30**	-0.13	0.10	**0.21**
V8	0.16	**0.20**	0.06	0.17	-0.07	V34	**0.21**	0.03	0.06	**-0.23**	0.07
V9	0.11	**0.26**	-0.05	**0.20**	0.00	V35	0.18	0.10	-0.01	-0.13	0.16
V10	**0.21**	0.10	0.11	0.05	-0.11	V36	0.12	-0.14	0.14	0.08	**0.20**
V11	-0.16	0.12	**0.27**	0.14	0.14	V37	**0.21**	0.00	0.08	**-0.22**	0.05
V12	-0.18	0.12	**0.28**	0.00	0.03	V38	0.18	0.08	0.02	-0.12	0.13
V13	**0.24**	0.09	0.10	0.08	-0.07	V39	0.13	-0.14	0.12	0.10	**0.20**
V14	-0.14	0.12	**0.29**	-0.08	-0.13	V40	0.04	**-0.20**	0.06	-0.09	0.17
V15	-0.15	0.14	-0.09	**-0.26**	0.00	V41	0.04	**0.33**	-0.04	-0.07	0.15
V16	-0.10	-0.05	**-0.25**	0.06	0.00	V42	0.08	**0.20**	0.00	-0.19	0.05
V17	-0.06	**0.26**	-0.13	**-0.23**	0.07	V43	**0.24**	-0.09	-0.11	**-0.28**	-0.10
V18	0.18	0.11	-0.12	**0.21**	-0.09	V44	-0.17	**-0.20**	-0.08	0.13	0.16
V19	0.15	0.03	-0.19	**0.24**	0.00	V45	-0.04	0.09	-0.04	-0.10	-0.09
V20	-0.02	0.06	-0.05	-0.05	**-0.35**	V46	0.08	0.07	0.10	-0.13	0.03
V21	0.03	0.07	-0.09	0.01	**-0.32**	V47	0.06	**-0.24**	-0.08	0.00	0.17
V22	0.11	-0.02	0.03	0.03	**0.20**	V48	-0.10	-0.02	-0.06	0.11	**0.22**
V23	-0.08	0.06	-0.08	-0.01	0.16	V49	0.11	-0.04	0.01	-0.04	-0.11
V24	-0.19	0.00	-0.02	-0.04	-0.11	V50	-0.05	0.03	-0.04	-0.05	-0.04
V25	0.07	0.18	0.11	**0.34**	-0.17	V51	**0.24**	-0.07	0.15	**-0.28**	-0.05
V26	-0.07	-0.01	-0.03	-0.01	-0.09						

Table 14: Loadings of the varimax rotated reduced variables for central midfielders that follow from applying Reduced k-means with $k = 6$ and $p = 5$. The loadings that are greater than or equal to 0.2 (in absolute terms) are shown in bold.

Applying Reduced k-means with $k = 6$ and $p = 5$ gives us the clusters that are plotted in Figure 8 based on the first two varimax rotated reduced variables, and for the the central midfielders who are chosen as representing players the names are displayed. The resulting average profile of each cluster is displayed in Table 15. Relating the average profiles of the clusters to the play styles discussed in Section 4.4.1 gives us the following interpretation of each cluster. We have ordered the clusters such that they are in the same order as the play styles defined in Section 4.4.1.

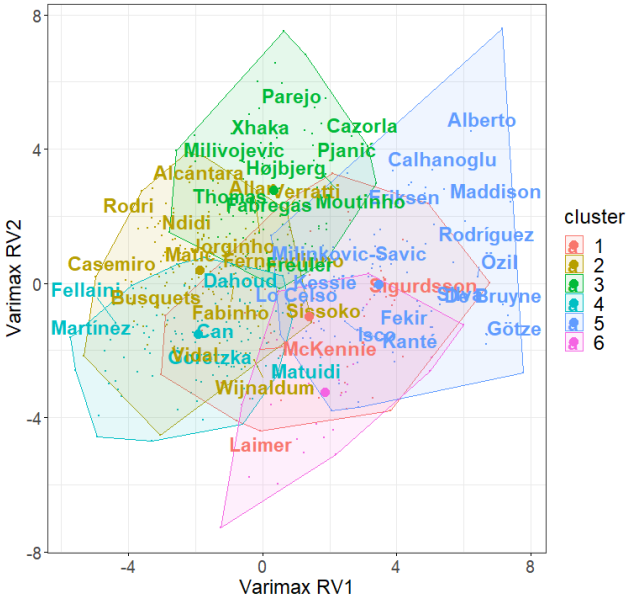

Figure 8: Plot of the found clusters based on the first two varimax rotated reduced variables for central midfielders, after applying Reduced k-means with $k = 6$ and $p = 5$.

	Cl. 1	Cl. 2	Cl. 3	Cl. 4	Cl. 5	Cl. 6		Cl. 1	Cl. 2	Cl. 3	Cl. 4	Cl. 5	Cl. 6
V1	8	10	11	8	9	7	V27	48	56	53	51	48	43
V2	9	11	12	9	9	6	V28	44	53	48	51	32	32
V3	74	88	84	83	83	73	V29	57	59	57	57	55	54
V4	55	53	55	51	48	46	V30	61	67	64	67	57	55
V5	66	82	78	76	76	63	V31	39	49	45	42	41	37
V6	15	20	18	18	18	17	V32	2	1	2	1	2	2
V7	14	13	16	13	11	9	V33	11	7	15	6	15	9
V8	1	1	1	1	2	1	V34	2	1	1	1	2	4
V9	12	7	12	6	15	10	V35	10	5	9	7	12	16
V10	59	47	52	50	66	71	V36	47	30	28	53	50	47
V11	77	89	86	85	86	75	V37	2	1	1	1	2	4
V12	72	87	83	81	82	72	V38	10	5	9	7	12	16
V13	29	16	21	20	35	41	V39	49	30	29	54	52	74
V14	67	82	77	75	78	69	V40	12	9	5	20	8	26
V15	9	11	12	9	9	8	V41	38	38	48	32	47	38
V16	55	54	55	54	46	43	V42	38	34	46	29	48	49
V17	9	10	12	8	10	8	V43	14	7	9	9	11	20
V18	8	3	5	4	7	7	V44	7	9	7	10	5	4
V19	4	1	2	2	3	3	V45	60	63	64	59	61	59
V20	27	28	25	18	24	21	V46	32	31	33	31	38	42
V21	16	11	12	6	12	11	V47	3	2	2	3	2	3
V22	61	43	52	58	65	78	V48	43	51	51	52	42	30
V23	55	64	70	60	55	42	V49	2	2	2	2	2	3
V24	47	56	52	52	46	41	V50	0	0	0	0	0	0
V25	4	4	4	3	6	3	V51	4	1	1	2	4	9
V26	4	6	5	5	3	2							

Table 15: Average profiles for the 6 clusters of central midfielders that are found by applying Reduced k-means with $k = 6$ and $p = 5$.

- **Cluster 1: 57 central midfielders (11%)**
  The average profile of this cluster scores high on `defensiveDuelsPercentageWon` and `foulsPerBallActionPercentage`, and low on `attackingDuelsPercentageWon`. This in-

dicates that players from this cluster tend to have a high defensive contribution, and are mainly focused on regaining possession of the ball. In addition, the average profile of this cluster scores low on `possessionRegainInPlayPercentageByInterception`, meaning that most of the possession regains in play are from duels. Hence, this cluster can be related to ball-winning midfielders.

* **Cluster 2: 134 central midfielders (26%)**
  The average profile of this cluster scores low on `passesPercentageFinalThird` and `actionsPercentageInOpponentBox`, and high on `shareInPassFirstInPossessionPercentage`. This indicates that players from this cluster do not play offensively, and that they have an important role in determining what the team is going to do when they gain possession of the ball. As a result, this cluster is most closely related to holding midfielders.

* **Cluster 3: 96 central midfielders (18%)**
  The average profile of this cluster scores high on `shareInBallActionsPercentage`, `shareInPassesPercentage`, `passesPercentageForward` and `shareInKeyActionsPercentage`, indicating that players from this cluster are likely to be involved in the build-up play. In addition, they try to create chances for teammates to score goals. This cluster thus most most closely resembles deep-lying playmakers.

* **Cluster 4: 115 central midfielders (22%)**
  The average profile of this cluster scores low on `shareInKeyActionsPercentage` and `shareInPossessionWithGoalAttemptsPercentage`, which indicates that a player from this cluster is not very involved in the play of his team when they are in possession of the ball. Furthermore, he scores high on `goalAttemptsPercentageInsidePenaltyBox`, meaning that he has a tendency to arrive late in the penalty area of the opposing team to make a goal attempt. This cluster can thus be related to box-to-box midfielders.

* **Cluster 5: 90 central midfielders (17%)**
  The average profile of this cluster scores high on `keyPassesPerBallActionPercentage`, `shareInKeyPassesPercentage`, `passesPercentageFinalThird`, `shareInGoalAttemptsPercentage`, `shareInPossessionWithGoalAttemptsPercentage` and `actionsPercentageInOpponentBox`. This shows that players from this cluster tend to play offensively, and that they try to create chances for teammates to score goals. Hence, this cluster is most closely related to advanced playmakers.

* **Cluster 6: 27 central midfielders (5%)**
  This is the cluster of players who usually tend to play as centre-forwards.

Table 16 is a confusion table which shows how many of the representing players of each play style are assigned to each cluster. It also shows the accuracy and average similarity for each play style and in total. We notice that many of the players who represent ball-winning midfielders end up in other clusters. Casemiro (Real Madrid), Moussa Sissoko (Tottenham Hotspur) and Fernandinho (Manchester City) end up in the cluster of holding midfielders. The cluster of deep-lying playmakers contains Remo Freuler (Atalanta Bergamo) and Pierre Højbjerg (Tottenham

Hotspur). Mahmoud Dahoud (Borussia Dortmund) is in the cluster of box-to-box midfielders, and the cluster of advanced playmakers contains N'Golo Kanté (Chelsea) and Giovani Lo Celso (Real Betis).

	Cluster								
Play style	1	2	3	4	5	6	Total	Accuracy (%)	Average similarity (%)
Ball-winning midfielders	**1**	3	2	1	2	0	9	11.1	77.6
Holding midfielders	0	**6**	2	1	0	0	9	66.7	92.5
Deep-lying playmakers	0	3	**6**	0	0	0	9	66.7	91.6
Box-to-box midfielders	1	2	0	**4**	2	0	9	44.4	86.5
Advanced playmakers	1	0	0	0	**11**	0	12	91.7	90.2
Total	3	14	10	6	15	0	48	58.3	87.8

Table 16: Confusion table which shows how many of the representing players of each play style are assigned to each cluster of central midfielders for Reduced k-means with $k = 6$ and $p = 5$. It also shows the accuracy and average similarity for each play style and in total.

5.2 Centre-forwards

For the player statistics of centre-forwards we have no missing values. The results that follow from applying PCA and hierarchical clustering for centre-forwards are presented in Section 5.2.1, and in Section 5.2.2 we present the results that follow from applying Reduced k-means for centre-forwards.

5.2.1 PCA and hierarchical clustering

We have applied PCA to the $(N \times q)$ data matrix $\mathbf{X}$ for centre-forwards to reduce the dimension q. For centre-forwards we have $N = 296$ and $q = 48$, meaning that we have 296 centre-forwards and 48 statistics. Applying PCA to the statistics of centre-forwards yields the scree plot that is displayed in Figure 9, together with the suggested number of principal components to retain for the methods Kaiser's rule, Parallel Analysis, Optimal Coordinates and Acceleration Factor. Also, in Table 17 we show the percentages of the explained variance for the 10 first principal components. As the most suggested number of principal components to retain is 5 (see Figure 9), we decide to retain the first five principal components. Hence, 51.6% of the variance of the original variables is retained.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Percentage of explained variance	20.9	9.6	8.7	6.7	5.7	4.3	4.0	3.5	2.9	2.7
Cumulative percentage of explained variance	20.9	30.5	39.2	45.9	51.6	55.9	59.9	63.4	66.2	69.0

Table 17: Percentages of explained variance for the first 10 principal components.

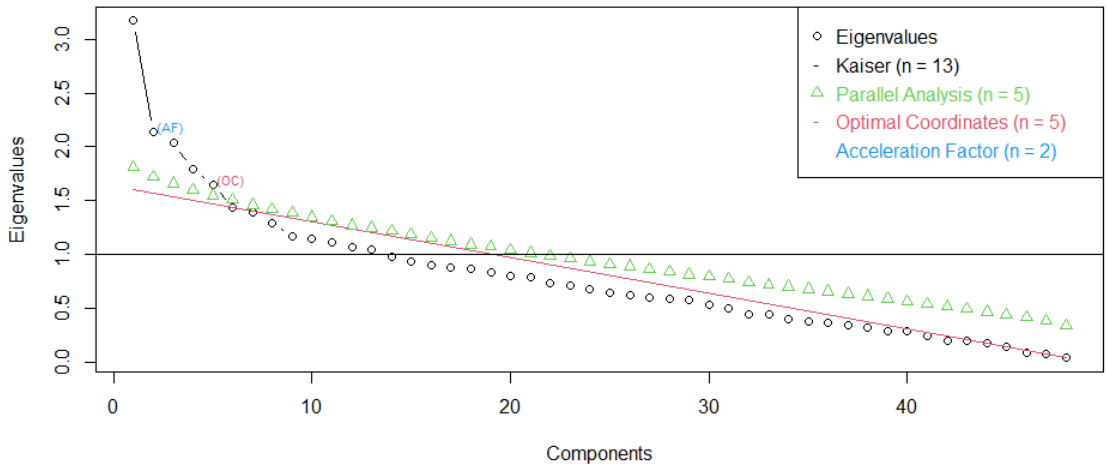

Figure 9: Scree plot of the principal components found for centre-forwards.

As we did for the retained principal components of central midfielders, we use varimax rotation to rotate the retained principal components of centre-forwards, such that the components become easier to interpret.

The loadings of the five varimax rotated principal components for centre-forwards can be found in Table 19, and in Table 18 the variables to which the loadings belong are displayed. The 10 highest loadings (in absolute terms) for each varimax rotated principal component are given in Appendix E.2.1, which can be used for easier interpretation of the components. The interpretation of the five varimax rotated principal components is given below.

* **PC1** (involvement and operating outside penalty box): can be interpreted as how involved a centre-forward is in the play of his team when they are in possession of the ball, and how much he operates outside of the penalty area of the opposing team.

* **PC2** (creating chances and offensive positioning/passing): can be interpreted as how likely a centre-forward is to create chances for teammates to score goals and how offensive he plays, both in terms of where he positions himself on the pitch and how offensive his passing is.

* **PC3** (simple passing): can be interpreted as how simple the passes of a centre-forward are. Simple passes are passes that are typically short and either wide or backwards.

* **PC4** ((involvement in) goal attempts): can be interpreted as how likely a centre-forward is to make a goal attempt or to be involved in the possession moment leading to a goal attempt.

* **PC5** (duel strength): can be interpreted as how strong a centre-forward is in duels.

Having performed PCA, we apply hierarchical clustering to the 5 principal components that are retained for the 296 centre-forwards. We again consider the three linkage criteria complete

Variable	Variable name	Variable	Variable name
V1	shareInBallActionsPercentage	V25	defensiveDuelsPercentageWon
V2	shareInPassesPercentage	V26	defensiveDuelsPercentageOwnHalf
V3	passesPercentageCompleted	V27	attackingDuelsPercentageWon
V4	passesPercentageForward	V28	keyActionsPerBallAction-Percentage
V5	passesForwardPercentage-Completed	V29	shareInKeyActionsPercentage
V6	passesPercentageWide	V30	goalAttemptsPerBallAction-Percentage
V7	passesPercentageLong	V31	shareInGoalAttemptsPercentage
V8	keyPassesPerBallAction-Percentage	V32	goalAttemptsPercentageInside-PenaltyBox
V9	shareInKeyPassesPercentage	V33	shotsPerBallActionPercentage
V10	passesPercentageOpponentHalf	V34	shareInShotsPercentage
V11	passesPercentageFinalThird	V35	shotsPercentageInsidePenaltyBox
V12	passesFinalThirdPercentage-Completed	V36	shotsPercentageWithHead
V13	shareInReceivedFirstPassIn-PossessionPercentage	V37	shareInPossessionWithGoal-AttemptsPercentage
V14	passesPercentageToBox	V38	shareInPossessionWithGoals-Percentage
V15	passesPercentageCrosses	V39	possessionLossPerBallAction-Percentage
V16	crossPassesPercentageCompleted	V40	possessionRegainInPlayPer-BallActionPercentage
V17	crossPassesPercentageToGoal-Attempt	V41	possessionRegainInPlay-PercentageByInterception
V18	crossPassesPercentageLate	V42	possessionRegainInPlay-PercentageOpponentHalf
V19	crossPassesPercentageHigh	V43	offsidesPerBallAction-Percentage
V20	duelsPercentageWon	V44	foulsPerBallActionPercentage
V21	dribblesPerBallAction-Percentage	V45	foulsPercentageOwnHalf
V22	slidingsPerDuel	V46	foulsSufferedPerBallAction-Percentage
V23	groundDuelsPercentageWon	V47	cardsPerFoul
V24	airDuelsPercentageWon	V48	actionsPercentageInOpponentBox

Table 18: Variables for centre-forwards.

linkage, average linkage and Ward linkage. For determining the number of clusters that we want to obtain, we again consider the number of play styles defined by Aalbers and Van Haaren (2018), the cluster dendrogram, the ASW and the CH index. For centre-forwards, we get that applying hierarchical clustering with complete linkage or average linkage provides us with a clustering in which one cluster contains approximately 99% of the centre-forwards, and all other clusters only contain less than 1% of the centre-forwards. Applying hierarchical clustering with Ward linkage provides us with a clustering in which the players are more evenly distributed over the clusters. Hence, we decide to use Ward linkage, which gives the cluster dendrogram that is displayed in Figure 10. The values of the ASW and the CH index for 2 until 6 clusters are displayed in Table 20.

The dendrogram, ASW and CH index each indicate that it would be optimal to set the number of clusters to 2. However, given that Aalbers and Van Haaren (2018) have defined 4 play

	PC1	PC2	PC3	PC4	PC5		PC1	PC2	PC3	PC4	PC5
V1	**0.31**	0.00	-0.08	-0.03	0.01	V25	0.02	0.12	-0.04	-0.05	**0.23**
V2	**0.30**	0.03	0.00	-0.05	0.02	V26	0.11	-0.10	0.06	-0.12	0.05
V3	0.01	0.02	**0.46**	0.03	-0.01	V27	0.01	-0.05	-0.02	0.10	**0.46**
V4	0.19	0.04	-0.19	0.00	0.04	V28	-0.04	**0.34**	0.08	0.01	0.06
V5	0.02	0.06	**0.39**	0.03	-0.02	V29	0.18	**0.21**	0.02	-0.04	0.05
V6	0.01	0.10	-0.02	0.09	-0.06	V30	-0.12	0.01	0.05	**0.41**	0.01
V7	**0.22**	-0.08	-0.13	0.09	0.02	V31	0.15	-0.06	-0.02	**0.46**	-0.02
V8	-0.12	**0.36**	0.08	0.01	0.05	V32	**-0.27**	-0.02	-0.05	-0.04	0.06
V9	0.11	**0.26**	0.00	-0.06	0.04	V33	-0.15	0.01	0.04	**0.39**	0.00
V10	-0.14	**0.32**	-0.12	0.06	-0.01	V34	0.11	-0.07	-0.04	**0.45**	-0.03
V11	-0.18	**0.35**	-0.17	0.00	-0.03	V35	**-0.26**	-0.02	-0.04	-0.04	0.07
V12	-0.02	0.04	**0.44**	-0.01	-0.02	V36	-0.15	-0.15	-0.07	0.02	0.13
V13	**0.28**	-0.04	-0.10	-0.04	0.00	V37	**0.25**	0.11	0.03	0.19	0.04
V14	0.07	**0.26**	**-0.27**	0.02	0.00	V38	0.11	0.05	0.05	0.18	0.13
V15	0.03	**0.21**	**-0.30**	-0.03	-0.05	V39	-0.10	-0.18	**-0.25**	-0.03	-0.06
V16	-0.01	0.01	0.11	-0.08	0.13	V40	-0.01	-0.01	0.01	-0.15	0.12
V17	-0.05	0.10	0.09	-0.04	0.16	V41	0.05	-0.03	0.02	0.17	**-0.20**
V18	-0.16	0.01	-0.01	-0.02	-0.02	V42	-0.10	0.11	-0.06	0.10	0.03
V19	0.13	0.00	0.01	0.05	-0.03	V43	-0.15	-0.06	-0.04	0.06	-0.05
V20	0.05	0.00	-0.01	0.05	**0.51**	V44	-0.09	**-0.23**	-0.08	-0.03	-0.01
V21	0.09	**0.21**	0.00	0.00	-0.19	V45	0.09	-0.03	0.12	-0.12	0.04
V22	0.02	0.01	0.02	-0.02	0.05	V46	-0.05	-0.10	-0.09	-0.01	0.18
V23	0.04	0.09	0.05	0.00	**0.42**	V47	0.06	0.06	0.03	0.08	0.02
V24	-0.06	-0.16	-0.11	0.07	**0.25**	V48	**-0.29**	0.08	-0.02	0.17	0.01

Table 19: Loadings of the varimax rotated principal components for centre-forwards. The loadings that are greater than or equal to 0.2 (in absolute terms) are shown in bold.

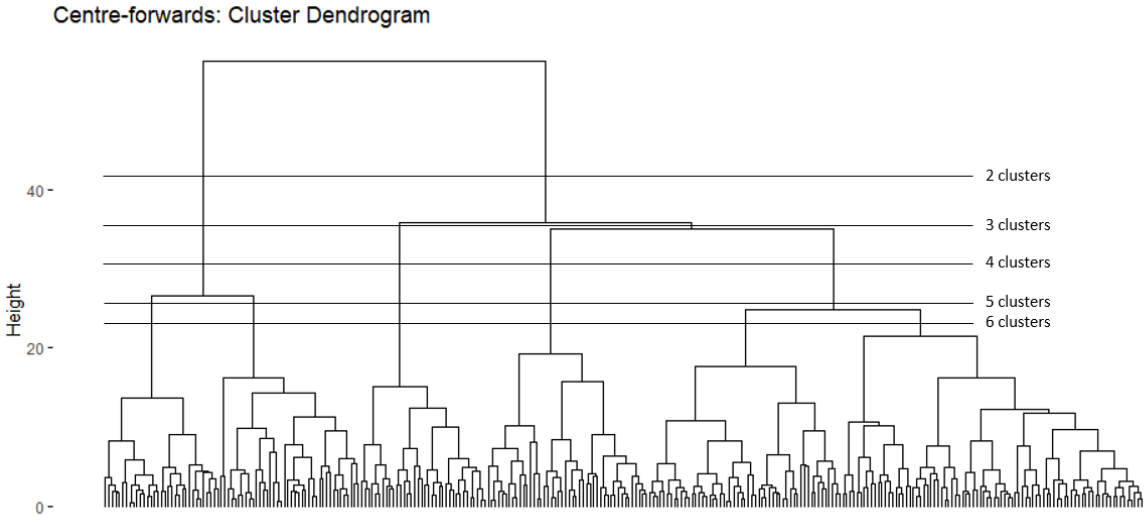

Figure 10: Cluster dendrogram for centre-forwards that follows from applying hierarchical clustering with Ward linkage.

styles for centre-forwards, we have strong reason to believe that there are more than 2 play styles for centre-forwards. When considering more than 2 clusters, we have an indication to set the number of clusters to 4 based on the dendrogram, the ASW and the CH index. The difference in heights in the dendogram is relatively high when cutting the tree at 4 clusters, and the ASW and CH index both decrease by a relatively high number when going from 4 to 5 clusters. As

Number of clusters	ASW	CH index
2	0.225	82
3	0.172	64
4	0.171	62
5	0.151	57
6	0.138	54

Table 20: Values of the ASW and the CH index for 2 until 6 clusters for centre-forwards.

Aalbers and Van Haaren (2018) have also defined 4 play styles for centre-forwards, we choose to set the number of clusters to 4. The resulting clusters are plotted in Figure 11 based on the first two varimax rotated principal components, and for the centre-forwards who are chosen as representing players the names are displayed. The corresponding obtained average profiles for each of the 4 clusters are shown in Table 21.

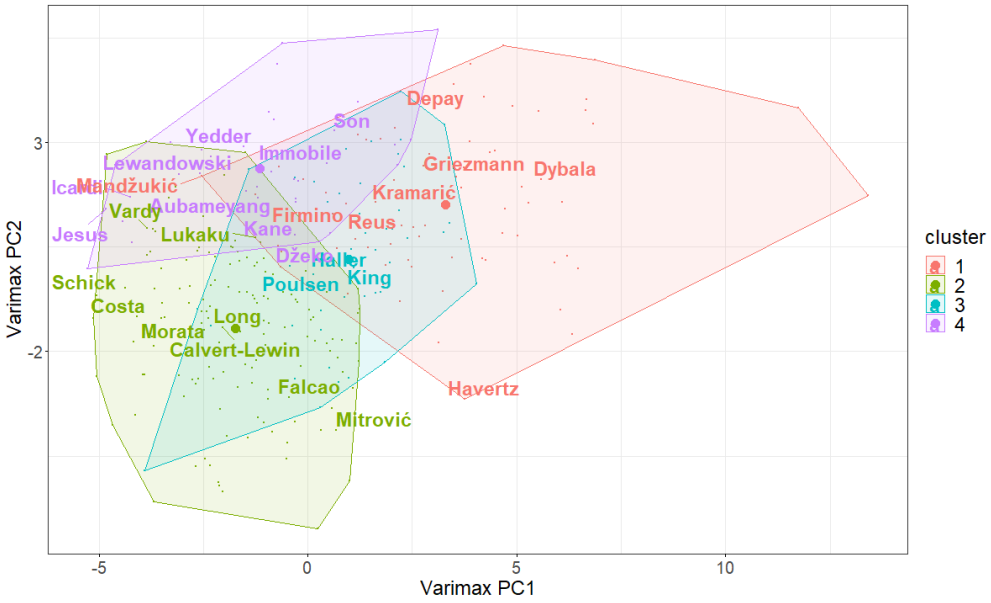

Figure 11: Plot of the found clusters based on the first two varimax rotated principal components for centre-forwards, after applying hierarchical clustering with the number of clusters set to 4.

When interpreting each of the clusters and relating them to the play styles defined in Section 4.4.2, we get the following interpretation of the clusters. We have ordered the clusters such that they are in the same order as the play styles defined in Section 4.4.2.

* **Cluster 1: 73 centre-forwards (25%)**
  The average profile of this cluster scores low on `actionsPercentageInOpponentBox` and `passesPercentageFinalThird`, indicating that players from this cluster mainly operate from outside the penalty area of the opposing team. Furthermore, players from this cluster score low on `foulsPerBallActionPercentage`, which shows that they are likely to have a low contribution in terms of defensive actions. This cluster is thus most closely related to shadow strikers.

* **Cluster 2: 141 centre-forwards (48%)**
  The average profile of this cluster scores low on `shareInGoalAttemptsPercentage` and

	Cl. 1	Cl. 2	Cl. 3	Cl. 4		Cl. 1	Cl. 2	Cl. 3	Cl. 4		Cl. 1	Cl. 2	Cl. 3	Cl. 4
V1	8	6	7	6	V17	11	7	9	15	V33	3	5	4	7
V2	8	4	6	5	V18	72	83	76	83	V34	17	18	19	22
V3	79	74	66	78	V19	51	41	44	43	V35	61	82	72	77
V4	45	39	46	41	V20	41	37	41	38	V36	7	27	19	15
V5	70	63	56	69	V21	6	4	6	6	V37	52	37	44	48
V6	18	17	18	19	V22	2	1	2	1	V38	57	43	48	56
V7	11	9	11	8	V23	43	37	40	40	V39	15	23	22	17
V8	2	2	2	3	V24	27	35	39	28	V40	3	3	3	3
V9	17	10	14	14	V25	55	46	54	52	V41	61	64	61	65
V10	76	76	79	82	V26	50	42	43	35	V42	45	51	52	63
V11	47	48	53	57	V27	37	35	38	36	V43	1	2	1	2
V12	74	69	60	74	V28	3	2	2	3	V44	2	4	3	2
V13	9	6	8	6	V29	18	9	13	13	V45	31	22	20	17
V14	10	7	12	9	V30	4	5	4	7	V46	3	3	4	2
V15	5	4	8	5	V31	18	18	19	22	V47	15	9	11	15
V16	24	20	17	24	V32	57	82	71	75	V48	8	14	11	16

Table 21: Average profiles for the 4 clusters of centre-forwards that are found by subsequently applying PCA and hierarchical clustering.

high on `shotsPercentageWithHead`, indicating that players from this cluster do not shoot a lot, but the shots they do produce are oftentimes headers. In addition, the average profile of this cluster scores high on `foulsPerBallActionPercentage`. This cluster thus most closely resembles target men.

* **Cluster 3: 46 centre-forwards (16%)**
  The average profile of this cluster scores low on `keyActionsPerBallActionPercentage` and `shareInKeyActionsPercentage`, and high on `shareInShotsPercentage`. Hence, players from this cluster tend to shoot instead of creating chances for teammates to score goals. As a result, this cluster is most closely related to poachers.

* **Cluster 4: 36 centre-forwards (12%)**
  The average profile of this cluster scores high on `shareInPossessionWithGoalAttempts-Percentage` and `shotsPercentageInsidePenaltyBox`, indicating that players from this cluster tend be involved in the build-up play, and they tend to shoot from inside the penalty area of the opposing team. Furthermore, players from this cluster are not likely to be involved in defensive actions, since they score low on `foulsPerBallActionPercentage`. Hence, this cluster most closely resembles mobile strikers.

Table 22 is a confusion table which shows how many of the representing players of each play style are assigned to each cluster. It also shows the accuracy and average similarity for each play style and in total. We notice that four of the players who represent poachers end up in another cluster. Romelu Lukaku (Manchester United) and Radamel Falcao (AS Monaco) end up in the cluster of target men, whereas Edin Džeko (AS Roma) and Mauro Icardi (Inter Milan) end up in the cluster of mobile strikers.

5.2.2 Reduced k-means

We want to apply Reduced k-means to the $(N \times q)$ data matrix $\mathbf{X}$ for centre-forwards to find k clusters of players in a p-dimensional subspace of the columns of $\mathbf{X}$. However, we still need to

	Cluster						
Play style	1	2	3	4	Total	Accuracy (%)	Average similarity (%)
Shadow strikers	**7**	0	0	1	8	87.5	79.6
Target men	1	**6**	1	0	8	75.0	79.8
Poachers	0	2	**1**	2	5	20.0	70.2
Mobile strikers	0	1	1	**6**	8	75.0	77.5
Total	8	9	3	9	29	69.0	77.4

Table 22: Confusion table which shows how many of the representing players of each play style are assigned to each cluster of centre-forwards for PCA and hierarchical clustering. It also shows the accuracy and average similarity for each play style and in total.

decide what values we choose for k (the number of clusters) and p (the lower dimension).

We again have $N = 296$ and $q = 48$ for centre-forwards, meaning that we have 296 centre-forwards and 48 statistics. For the number of clusters we consider the number of play styles defined by Aalbers and Van Haaren (2018), the ASW and the CH index. For a given value of k, we again decide on the value of p by the following rule of thumb. We initially take $p = k - 1$, and we then take the lowest value of p for which the resulting allocation of players to clusters is exactly the same as for the case where $p = k - 1$. The resulting values of the ASW and the CH index for 2 until 6 clusters are displayed in Table 23.

Number of clusters (k)	Dimension (p)	ASW	CH index
2	1	0.616	613
3	2	0.380	254
4	3	0.293	148
5	4	0.237	99
6	5	0.211	79

Table 23: Values of the ASW and the CH index for 2 until 6 clusters for centre-forwards.

The ASW and the CH index both indicate that taking $k = 2$ would be optimal. However, given that Aalbers and Van Haaren (2018) have defined 4 play styles for centre-forwards, we have strong reason to believe that there are more than 2 play styles for centre-forwards. For $k > 2$, the ASW and the CH index indicate to take as few clusters as possible. We thus have a trade-off between taking as few clusters as possible and taking 4 clusters. As it is hard to make a decision for this trade-off, we choose to rely on the expert knowledge provided by Aalbers and Van Haaren (2018), and take 4 clusters. For $k = 4$, we get $p = 3$ according to our rule of thumb. We again use varimax rotation to rotate the 3 reduced variables. The loadings of these varimax rotated reduced variables can be found in Table 24. In Appendix E.2.2, the 10 highest loadings (in absolute terms) can be found for each varimax rotated reduced variable. The interpretation of the three varimax rotated reduced variables is given below.

- **RV1** (operating inside penalty box, low involvement and likelihood of goal attempts): can be interpreted as how much a centre-forward operates in the penalty area of the opposing

team, how low his involvement is in the play of his team when they are in possession of the ball, and how likely he is to make goal attempts.

* **RV2** (creating chances): can be interpreted as how likely a centre-forward is to create chances for teammates to score goals.

* **RV3** (simple passing outside final third): can be interpreted as how simple the passes of a centre-forward are, and how much of his passes are outside of the final third. Simple passes are passes that typically short and either wide or backwards.

	RV1	RV2	RV3		RV1	RV2	RV3		RV1	RV2	RV3
V1	**-0.25**	**0.23**	-0.09	V17	-0.05	**-0.20**	-0.12	V33	0.15	**-0.24**	-0.09
V2	**-0.27**	0.15	0.00	V18	0.10	-0.06	-0.04	V34	0.01	0.00	-0.17
V3	-0.09	**-0.21**	**0.36**	V19	-0.09	0.04	0.03	V35	**0.22**	-0.05	-0.09
V4	-0.17	0.11	-0.16	V20	-0.11	0.03	-0.05	V36	**0.21**	0.06	-0.03
V5	-0.11	-0.19	**0.28**	V21	-0.17	-0.07	-0.09	V37	**-0.26**	-0.01	-0.05
V6	-0.04	0.00	-0.14	V22	-0.04	-0.03	0.02	V38	-0.14	-0.09	-0.07
V7	-0.10	0.12	-0.05	V23	-0.14	-0.08	0.01	V39	**0.23**	**0.22**	-0.13
V8	-0.08	**-0.39**	-0.12	V24	0.11	0.11	-0.15	V40	-0.03	0.01	0.03
V9	**-0.21**	-0.11	-0.10	V25	-0.11	-0.05	-0.13	V41	0.02	-0.01	0.01
V10	-0.02	-0.16	**-0.27**	V26	-0.06	0.10	0.07	V42	0.05	-0.08	-0.13
V11	0.00	-0.18	**-0.27**	V27	-0.04	0.06	-0.05	V43	0.18	-0.01	0.01
V12	-0.06	**-0.21**	**0.35**	V28	-0.15	**-0.34**	-0.03	V44	**0.21**	0.17	0.08
V13	**-0.20**	**0.23**	-0.09	V29	**-0.26**	-0.07	0.01	V45	-0.09	0.03	0.12
V14	-0.14	-0.02	**-0.34**	V30	0.13	**-0.24**	-0.07	V46	0.08	0.09	-0.01
V15	-0.07	0.03	**-0.28**	V31	-0.04	0.00	-0.12	V47	-0.10	0.04	0.01
V16	-0.03	-0.12	0.02	V32	**0.24**	-0.05	-0.11	V48	**0.22**	**-0.24**	-0.13

Table 24: Loadings of the varimax rotated reduced variables for centre-forwards that follow from applying Reduced k-means with $k = 4$ and $p = 3$. The loadings that are greater than or equal to 0.2 (in absolute terms) are shown in bold.

Furthermore, Reduced k-means with $k = 4$ and $p = 3$ gives us the clusters that are shown in Figure 12 based on the first two varimax rotated reduced variables, and for the the centre-forwards who are chosen as representing players the names are displayed. In Table 25 the resulting average profiles for each cluster are displayed, together with the (relative) size of each cluster.

Relating the average profiles of the clusters to the play styles discussed in Section 4.4.2 gives us the following interpretation of each cluster. We have ordered the clusters such that they are in the same order as the play styles defined in Section 4.4.2.

* **Cluster 1: 78 centre-forwards (26%)**
  The average profile of this cluster scores low on `actionsPercentageInOpponentBox` and `goalAttemptsPercentageInsidePenaltyBox`, indicating that players from this cluster mainly operate from outside the penalty area of the opposing team. Furthermore, players from this cluster score low on `foulsPerBallActionPercentage`, which shows that they are likely to have a low contribution in terms of defensive actions. This cluster is thus most closely related to shadow strikers.

* **Cluster 2: 93 centre-forwards (31%)**
  The average profile of this cluster scores low on `shareInGoalAttemptsPercentage` and

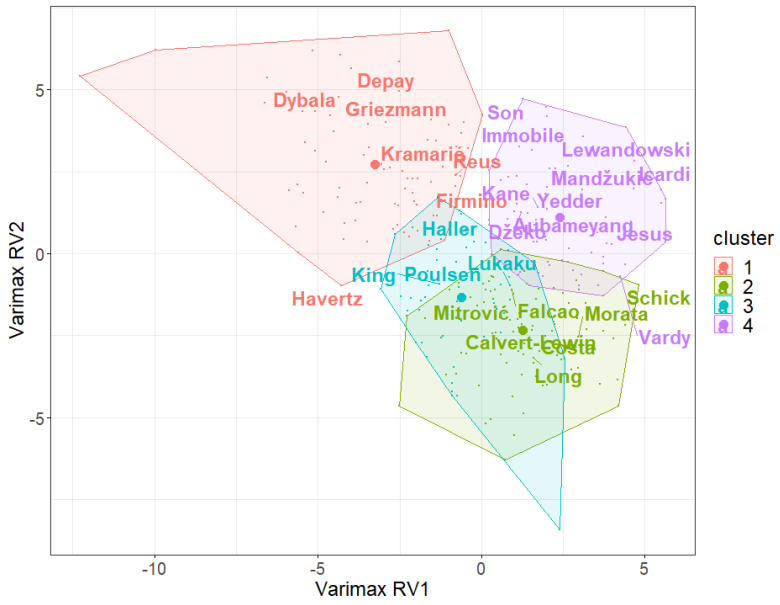

Figure 12: Plot of the found clusters based on the first two varimax rotated reduced variables for centre-forwards, after applying Reduced k-means with $k = 4$ and $p = 3$.

	Cl. 1	Cl. 2	Cl. 3	Cl. 4		Cl. 1	Cl. 2	Cl. 3	Cl. 4		Cl. 1	Cl. 2	Cl. 3	Cl. 4
V1	8	6	7	5	V17	11	5	8	15	V33	3	5	4	6
V2	8	4	6	4	V18	72	82	79	84	V34	18	18	20	19
V3	78	74	67	75	V19	52	41	44	40	V35	58	82	77	81
V4	45	38	45	40	V20	41	37	39	38	V36	6	28	22	21
V5	69	63	56	65	V21	7	3	5	5	V37	53	35	43	41
V6	18	17	19	18	V22	2	1	1	2	V38	58	40	49	51
V7	11	8	10	8	V23	43	37	38	40	V39	15	23	23	19
V8	2	1	2	3	V24	25	35	39	32	V40	3	3	3	3
V9	17	9	13	13	V25	54	45	52	51	V41	62	64	62	63
V10	76	74	79	80	V26	48	43	43	39	V42	46	50	54	56
V11	48	46	52	55	V27	37	35	37	35	V43	1	2	1	2
V12	73	70	61	71	V28	3	2	2	3	V44	2	4	3	3
V13	9	6	8	6	V29	19	9	12	12	V45	30	22	20	20
V14	10	6	11	9	V30	4	5	5	6	V46	3	3	3	3
V15	6	4	7	5	V31	19	17	20	19	V47	18	8	12	9
V16	23	18	17	25	V32	54	81	77	81	V48	7	13	12	16

Table 25: Average profiles for the 4 clusters of centre-forwards that are found by applying Reduced k-means with $k = 4$ and $p = 3$.

`shareInShotsPercentage`, and high on `shotsPercentageWithHead`, indicating that players from this cluster do not shoot a lot, but the shots they do produce are oftentimes headers. In addition, the average profile of this cluster scores high on `foulsPerBallActionPercentage`. This cluster thus most closely resembles target men.

* **Cluster 3: 54 centre-forwards (18%)**
  The average profile of this cluster scores low on `keyActionsPerBallActionPercentage` and `shareInKeyActionsPercentage`, and high on `shareInGoalAttemptsPercentage` and `shareInShotsPercentage`. Hence, players from this cluster tend to shoot instead of creating chances for teammates to score goals. As a result, this cluster is most closely related to poachers.

- **Cluster 4: 71 centre-forwards (24%)**
  The average profile of this cluster scores high on `shareInPossessionWithGoalsPercentage` and `shotsPercentageInsidePenaltyBox`, indicating that players from this cluster tend be involved in the build-up play, and they tend to shoot from inside the penalty area of the opposing team. Hence, this cluster most closely resembles mobile strikers.

Table 26 is a confusion table which shows how many of the representing players of each play style are assigned to each cluster. It also shows the accuracy and average similarity for each play style and in total.

	Cluster						
Play style	1	2	3	4	Total	Accuracy (%)	Average similarity (%)
Shadow strikers	**7**	0	0	1	8	87.5	82.1
Target men	0	**6**	1	1	8	75.0	79.9
Poachers	0	1	**2**	2	5	40.0	75.1
Mobile strikers	0	0	1	**7**	8	87.5	83.1
Total	7	7	4	11	29	75.9	80.6

Table 26: Confusion table which shows how many of the representing players of each play style are assigned to each cluster of centre-forwards for Reduced k-means with $k = 4$ and $p = 3$. It also shows the accuracy and average similarity for each play style and in total.

5.3 Comparison between methods

For central midfielders, PCA and hierarchical clustering achieves an accuracy of 52.1% and an average similarity of 87.8%. Reduced k-means achieves an accuracy of 58.3%, and an average similarity of 87.8%. For centre-forwards, PCA and hierarchical clustering achieves an accuracy of 69.0% and an average similarity of 77.4%. Reduced k-means achieves an accuracy of 75.9%, and an average similarity of 80.6%. Hence, both in terms of accuracy and average similarity, we conclude that in general Reduced k-means outperforms PCA and hierarchical clustering.

6 Conclusion

In this thesis we have identified play styles of football players based on match event data. In order to do this, we first determined the playing positions of football players in single matches based on their average locations. From the playing positions of players in single matches we were able derive to which position group each player belongs in every single match. These six position groups are goalkeepers, centre-backs, wing-backs, central midfielders, wingers and centre-forwards. Based on these position groups in single matches we determined to which position group(s) players belong over multiple matches. In this thesis we did not evaluate the play styles for all six position groups, as this would become rather extensive. Instead, we focused on the two position groups for which we think the identification of play styles is most relevant and

interesting, which are central midfielders and centre-forwards. For both of these position groups we determined which player statistics we find relevant for identifying play styles in that position group. Also, for both position groups we determined some pre-defined play styles with representing players in order to evaluate the results. Based on the player statistics we identified play styles for both position groups, for which we considered two approaches. The first approach consists of subsequently applying Principal Component Analysis (PCA) and hierarchical clustering, and the second approach is called Reduced k-means, which is a joint dimension reduction and clustering method. The found clusters for both position groups and both methods were evaluated by relating the found clusters to the pre-defined play styles, assessing whether representing players end up in the cluster that has the highest resemblance to their pre-defined play style (which results in accuracies), and assessing how similar a player who represents a certain play style is to the cluster that is related to that play style (which results in average similarities).

The results showed that when interpreting the found clusters and relating them to the pre-defined play styles, for central midfielders we find an extra cluster of players who usually tend to play as centre-forwards. However, these players are classified as central midfielders. This indicates that our approach for determining players' playing positions can be improved. For example, our approach for determining players' playing positions can possibly be improved by using median locations of players instead of average locations (since the median is a more robust statistic than the mean). It could also help to take into account the standard deviations of players' locations in matches, in both the x-direction and the y-direction.

In terms of accuracy and average similarity for the pre-defined play styles, the results showed that in general Reduced k-means performs better than subsequently applying PCA and hierarchical clustering. PCA and hierarchical clustering achieves an accuracy of 52.1% and 69.0% for central midfielders and centre-forwards, respectively. On the other hand, Reduced k-means achieves an accuracy of 58.3% and 75.9% for central midfielders and centre-forwards, respectively. The average similarity that follows from PCA and hierarchical clustering is 87.8% and 77.4% for central midfielders and centre-forwards, respectively. For Reduced k-means, the average similarity is 87.8% and 80.6% for central midfielders and centre-forwards, respectively. Hence, in general, we prefer Reduced k-means over PCA and hierarchical clustering because of the better performance in terms of accuracy and average similarity.

References

Aalbers, B. and Van Haaren, J. (2018). Distinguishing between roles of football players in play-by-play match event data. In *International Workshop on Machine Learning and Data Mining for Sports Analytics*, pages 31–41. Springer.

Achanta, S. (2020). What can Miralem Pjanić provide to Barcelona? `https://barcauniversal.com/what-can-miralem-pjanic-provide-to-barcelona/`. Accessed: 2020-12-07.

Agate, A. (2019). Nabil Fekir 2018/19 – scout report. `https://footballbh.net/2019/07/26/nabil-fekir-201819-scout-report-tactical-analysis-tactics/`. Accessed: 2020-12-07.

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420 – 434. Springer.

Baldi, R. (2019). Joao Moutinho: The most underrated player in the Premier League? `https://www.unibet.co.uk/blog/football/premier-league/joao-moutinho-the-most-underrated-player-in-the-premier-league-1.1317223`. Accessed: 2020-12-06.

Bate, A. (2019). Why Fabinho is now the Premier League's best holding midfielder. `https://www.skysports.com/football/news/15117/11812068/why-fabinho-is-now-the-premier-leagues-best-holding-midfielder`. Accessed: 2020-12-04.

Bertsekas, D. P. (1998). *Network optimization: continuous and discrete models*. Athena Scientific Belmont, MA.

Beuvink, W. (2018). Hoe een typische en atypische Braziliaanse speler bij Liverpool en Manchester City van belang zijn voor de einduitslag. `https://www.tussendelinies.nl/brazilianen-liverpool-manchestercity/`. Accessed: 2020-12-09.

Bliss, N. (2019). How good is Wissam Ben Yedder amid Man Utd transfer rumours? - Football Manager. `https://www.manchestereveningnews.co.uk/sport/football/transfer-news/wissam-ben-yedder-manchester-united-16488026`. Accessed: 2020-12-09.

Bodell, T. (2020). Loved by Pep, wanted by Mou: What is it Pierre-Emile Højbjerg does? `https://footballwhispers.com/blog/pierre-emile-hojbjerg-tottenham-hotspur-scout-report/`. Accessed: 2020-12-03.

Bourgeois, B. (2019). Ralph Hasenhüttl explains why Shane Long is so important. `https://onefootball.com/en/news/ralph-hasenhuttl-explains-why-shane-long-is-so-important-28304080`. Accessed: 2020-12-08.

Branine, M. (2008). Graduate recruitment and selection in the UK. *Career Development International*.

Caliński, T. and Harabasz, J. (1974). A Dendrite Method for Cluster Analysis. *Communications in Statistics - Theory and Methods*, 3:1–27.

Carlisle, J. (2018). Diego Costa doesn't fit Spain style, but sometimes they need a battering ram. https://www.espn.com/soccer/fifa-world-cup/4/blog/post/3541910/diego-costa-doesnt-fit-spain-style-but-sometimes-they-need-a-battering-ram. Accessed: 2020-12-08.

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, 1(2):245–276.

Chambers, R. (2020a). Charting the goal-scoring abilities of Ciro Immobile. https://www.scisports.com/charting-the-goal-scoring-abilities-of-ciro-immobile-after-a-sensational-2019-20-campaign/. Accessed: 2020-12-09.

Chambers, R. (2020b). EFL is crucial to developing Premier League talent and these examples show why. https://www.scisports.com/efl-is-crucial-to-developing-premier-league-talent-and-these-examples-show-why/. Accessed: 2020-12-09.

Chambers, R. (2020c). Scouting Europe: Finding the next David Silva through data. https://www.scisports.com/scouting-europe-finding-the-next-david-silva-through-data/. Accessed: 2020-12-07.

Cooper, M. (2020). Sergej Milinkovic-Savic: The Sergeant Who Dominates All Who Stand Before Him. https://www.90min.com/posts/sergej-milinkovic-savic-best-box-to-box-midfielders-in-the-world. Accessed: 2020-12-07.

De Soete, G. and Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. In Diday, E., Lechevallier, Y., Schader, M., Bertrand, P., and Burtschy, B., editors, *New Approaches in Classification and Data Analysis*, pages 212–219, Berlin, Heidelberg. Springer Berlin Heidelberg.

Decroos, T. and Davis, J. (2019). Player Vectors: Characterizing Soccer Players' Playing Style from Match Event Streams. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer.

El-Shaboury, Y. (2019). Premier League 2018/19 Tactical Analysis: James Maddison at Leicester City. https://footballbh.net/2019/05/31/tactical-analysis-premier-league-james-maddison-leicester-city/. Accessed: 2020-12-07.

Elliott, N. (2018). How Mauro Icardi ignored the modern way to become Europe's best poacher. https://www.dreamteamfc.com/c/news-gossip/429019/mauro-icardi-inter-poacher/. Accessed: 2020-12-08.

Fitzpatrick, R. (2019). Frozen-Out Isco: Can the Real Madrid Midfielder Find His Way in from the Cold? https://bleacherreport.com/articles/2816831-frozen-out-isco-can-the-real-madrid-midfielder-find-his-way-in-from-the-cold. Accessed: 2020-12-07.

Geerts, A., Decroos, T., and Davis, J. (2018). Characterizing Soccer Players' Playing Style from Match Event Streams. In *Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2018 workshop*, volume 2284, pages 115–126. Springer.

Goldstein, P. (2017). Splunk .Conf2017: Oakland A's Billy Beane Says Data Analytics Has Transformed Baseball. https://biztechmagazine.com/article/2017/09/splunk-conf2017-oakland-billy-beane-says-data-analytics-has-transformed-baseball. Accessed: 2020-05-13.

Harris, T. (2018). Aleksandr Mitrović: the old-school, bull-dozing, stern-faced slab of Serbian bulk. https://breakingthelines.com/opinion/aleksandr-mitrovic-the-old-school-bull-dozing-stern-faced-slab-of-serbian-bulk/. Accessed: 2020-12-08.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Hastie, T., Tibshirani, R., Narasimhan, B., and Chu, G. (2020). *impute: Imputation for microarray data*. R package version 1.64.0.

Horn, J. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185.

Hornik, K. (2019). *clue: Cluster ensembles*. R package version 0.3-57.

Jackson, R. (2020). Shane Long Deserves More Credit as the Key to Danny Ings and Southampton's Success. https://www.90min.com/posts/shane-long-deserves-more-credit-as-the-key-to-danny-ings-and-southampton-success. Accessed: 2020-12-08.

Jacob, S. (2019). Arsenal signing this 22-year-old Roma starlet would be perfect for Unai Emery's style of football. https://www.soccersouls.com/arsenal-signing-this-22-year-old-roma-starlet-would-be-perfect-for-unai-emerys-style-of-football/. Accessed: 2020-12-08.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice-Hall, Inc., USA.

Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, 20(1):141–151.

Kalenderoğlu, U. (2019). Football Player Profiling Using Opta Match Event Data: Hierarchical Clustering. Master's thesis, MEF University.

Kassambara, A. and Mundt, F. (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7.

Kaynak, K. (2019). Mesut Ozil Is the Main Example That the Classic No. 10 Is Dead in Modern Football. https://www.90min.com/posts/6476895-mesut-ozil-is-the-main-example-that-the-classic-no-10-is-dead-in-modern-football. Accessed: 2020-12-07.

Kelly, A. (2019). Why Liverpool's Fabinho Is Simply the Best Defensive Midfielder in the Business Right Now. https://www.90min.com/posts/6467257-why-liverpool-s-fabinho-is-

simply-the-best-defensive-midfielder-in-the-business-right-now. Accessed: 2020-12-04.

Keogh, F. and Rose, G. (2013). Football betting - the global gambling industry worth billions. https://www.bbc.com/sport/football/24354124. Accessed: 2020-11-27.

Kircher, L. (2020). Memphis Depay: How could he benefit Barcelona? https://totalfootballanalysis.com/article/memphis-depay-benefit-barcelona-data-analysis-statistics. Accessed: 2020-12-09.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

Lewis, M. M. (2003). *Moneyball: The Art of Winning an Unfair Game*. W. W. Norton & Company, New York City, NY.

Macdonald, M. (2020). PSG, Chelsea or Man Utd? Who might sign Juventus ace Paulo Dybala. https://www.footballtransfers.com/2020/12/dybala-not-fancied-by-pirlo-where-could-he-move. Accessed: 2020-12-09.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis*. Academic Press London; New York.

Markos, A., Iodice D'Enza, A., and Van de Velden, M. (2019). Beyond tandem analysis: Joint dimension reduction and clustering in R. *Journal of Statistical Software*, 91(10):1–24.

McLaughlin, M. (2018). How Data Analytics Is Revolutionizing Sports. https://biztechmagazine.com/article/2018/12/how-data-analytics-revolutionizing-sports. Accessed: 2020-07-02.

Mukherjee, R. (2020). Miralem Pjanic to Barcelona: How will the Bosnian midfielder fit in Setien's squad? https://footviser.com/miralem-pjanic-to-barcelona/. Accessed: 2020-12-07.

Murtagh, F. and Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, 31(3):274–295.

Nalton, J. (2020). Soccer Analysis Moves Toward Smarter Scouting And More Accessible Data. https://www.forbes.com/sites/jamesnalton/2020/02/19/soccer-analytics-smarter-scouting-and-more-accessible-data/#505258bc1557. Accessed: 2020-07-02.

Parker, I. (2018). Manchester United eyeing box-to-box midfielder. https://www.dailypost.co.uk/sport/football/transfer-news/manchester-united-eyeing-box-box-14700915. Accessed: 2020-12-07.

Pearson, G. (2020). A perfect box to box midfielder option for Tottenham. https://hotspurhq.com/2020/04/04/perfect-box-box-midfielder-option-tottenham/. Accessed: 2020-12-07.

Peña, J. L. and Navarro, R. S. (2015). Who can replace xavi? a passing motif analysis of football players. *arXiv preprint arXiv:1506.07768*.

Quint, J. (2020). What are the definitions of the 22 Player Roles? `https://scisports.zendesk.com/hc/en-gb/articles/360011564238`. Accessed: 2020-12-03.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raîche, G., Walls, T., Magis, D., Riopel, M., and Blais, J. (2013). Non-graphical solutions for cattell's scree test. *Methodology*, 9:23–29.

Roden, L. (2019). Dani Parejo: From QPR to Spain's most consistent midfielder – Valencia's captain is delivering on the promise Alfredo Di Stéfano saw in him at Real Madrid. `https://talksport.com/football/519605/dani-parejo-qpr-valencia-alfredo-di-stefano-real-madrid/`. Accessed: 2020-12-06.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65.

Sandford, D. (2019). Sebastien Haller: Profile of the Eintracht Frankfurt striker set to join West Ham for £45million and compared to a World Cup winner. `https://talksport.com/football/573631/sebastien-haller-profile-eintracht-frankfurt-striker-transfer-west-ham-world-cup-winner/`. Accessed: 2020-12-08.

Sengupta, R. (2019). Liverpool reviving interest in this 25-year-old international playmaker would be a brilliant move. `https://www.soccersouls.com/liverpool-reviving-interest-in-this-25-year-old-international-playmaker-would-be-a-brilliant-move/`. Accessed: 2020-12-07.

Shelat, N. (2020). Scout Report: Why Weston McKennie is the Right Fit for Juventus. `https://theseriea.com/scout-report-why-weston-mckennie-is-the-right-fit-for-juventus`. Accessed: 2020-12-03.

Sierksma, G. (2006). Computer Support for Coaching and Scouting in Football. In Moritz, E. F. and Haake, S., editors, *The Engineering of Sport 6*, pages 215–219, New York, NY. Springer New York.

Smith, P. (2020). Jamie Vardy: How Leicester star has evolved his game to continue to shine in the Premier League. `https://www.skysports.com/football/news/30385/12083900/jamie-vardy-how-leicester-star-has-evolved-his-game-to-continue-to-shine-in-the-premier-league`. Accessed: 2020-12-09.

Sullivan, J. (2016). Beautiful and mathematical: Football as a numbers game. `https://www.bbc.com/news/science-environment-37327939`. Accessed: 2020-12-03.

Taylor, J. B., Mellalieu, S. D., and James, N. (2004). Behavioural comparisons of positional demands in professional soccer. *International Journal of Performance Analysis in Sport*, 4(1):81–97.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays.

Van de Ven, E. (2018). Clustering soccer players to find the drivers of soccer team performance. Master's thesis, Erasmus University Rotterdam.

Volpicelli, G. (2020). Football was slow to embrace data. Now AI is eating the beautiful game. https://www.wired.co.uk/article/football-data. Accessed: 2020-07-02.

Ward Jr., J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244.

Warrier, R. (2018). How Falcao rose to the top and overcame an unlikely fall from grace. https://thesefootballtimes.co/2018/01/03/how-falcao-rose-to-the-top-and-overcame-an-unlikely-fall-from-grace/. Accessed: 2020-12-08.

Wensveen, C. J. (2016). Classification of playing styles in football: The use of ball action data. Master's thesis, Technische Universiteit Delft.

Wright, N. (2019). Rodri to Manchester City: Sergio Busquets-style midfielder can shine. https://www.skysports.com/football/news/11679/11757455/rodri-to-manchester-city-sergio-busquets-style-midfielder-can-shine. Accessed: 2020-12-03.

Wright, N. (2020). Thomas Partey: Arsenal's £45m midfielder is a 'physical marvel' who 'does everything well'. https://www.skysports.com/football/news/11095/12081408/thomas-partey-arsenals-45m-midfielder-a-physical-marvel-who-does-everything-well. Accessed: 2020-12-06.

Yamamoto, M. and Hwang, H. (2014). A General Formulation of Cluster Analysis with Dimension Reduction and Subspace Separation. *Behaviormetrika*, 41:115–129.

Zavala, S. (2019). What is the best role for Paulo Dybala in Sarri's Juventus? https://fansided.com/2019/10/22/best-role-paulo-dybala-sarris-juventus/. Accessed: 2020-12-08.

Zavala, S. (2020). How Rodri has grown as a press-resistant midfielder at Manchester City. https://www.si.com/soccer/manchestercity/match-coverage/how-rodri-has-grown-as-a-press-resistant-midfielder-at-manchester-city. Accessed: 2020-12-03.

Zide, J., Elman, B., and Shahani-Denning, C. (2014). Linkedin and recruitment: How profiles differ across occupations. *Employee Relations*, 36(5):583–604.

Appendix

A Formation data

In this thesis, we are critically analysing the correctness of position labels for players. Hence, we also check the correctness of standard position labels given by ORTEC Sports to certain formations, which is discussed in Section A.1.

A.1 Relabeling of some standard positions for certain formations

In formations with 3 defenders (that is, the formations 3-4-1-2, 3-4-3 and 3-5-2), the defenders are always labeled as a left-back (LB), centre-back (CB), and right-back (RB). If we however look at Figure 13, in which the weighted average locations are shown of the positions left-back (LB), left-centre-back (LCB), right-centre-back (RCB) and right-back (RB) for each formation, we observe that the players who play in LB and RB usually tend to play more as a LCB and RCB, respectively. As a result, for the formations 3-4-1-2, 3-4-3 and 3-5-2, we replace the position labels of the LB and RB by the labels LCB and RCB, respectively.

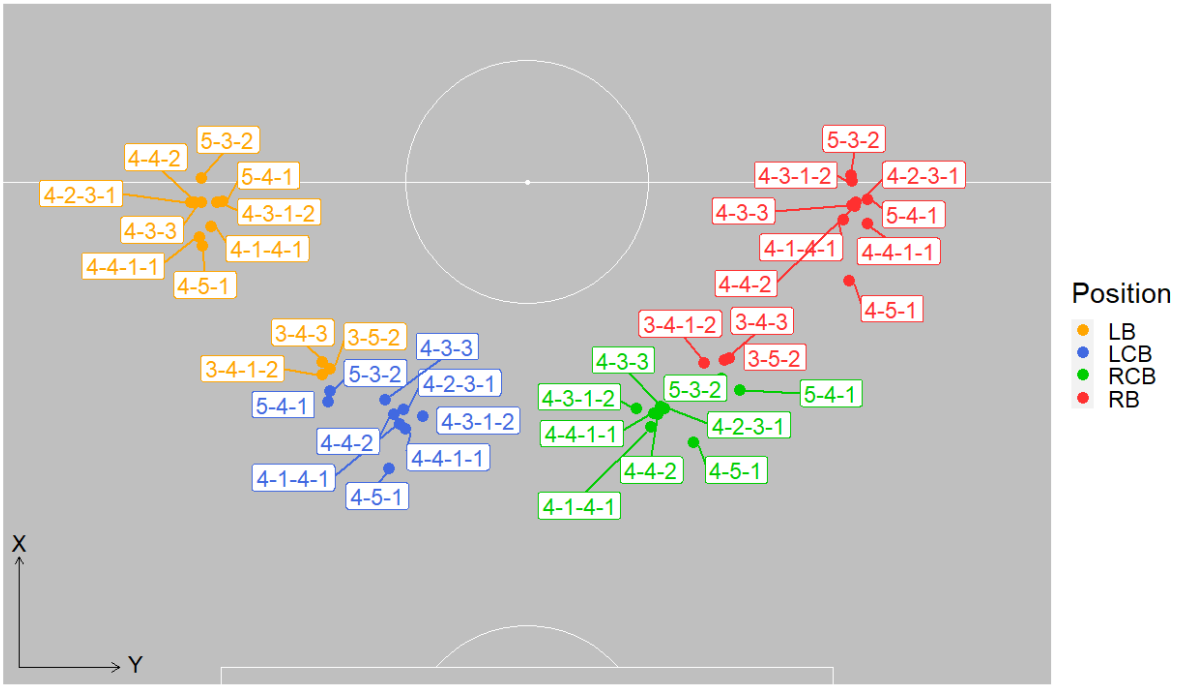

Figure 13: Weighted average locations of the positions left-back (LB), left-centre-back (LCB), right-centre-back (RCB) and right-back (RB) for each formation before relabeling some standard positions for certain formations.

When checking the correctness of standard position labels given by ORTEC Sports to players in certain formations for all other positions as well, we observe a similar situation for the formations 4-3-1-2, 4-3-3 and 5-3-2. The 3 midfielders who play next to each other in these formations are always labeled as left-midfielder (LM), central midfielder (CM), and right-midfielder (RM). However, the players that play in LM and RM in these formations actually tend to play more as a left-centre-midfielder (LCM) and right-centre-midfielder (RCM), respectively. This can be

seen from Figures 14 and 15, in which the weighted average locations are shown of the positions LM, LCM, RCM and RM for each formation.

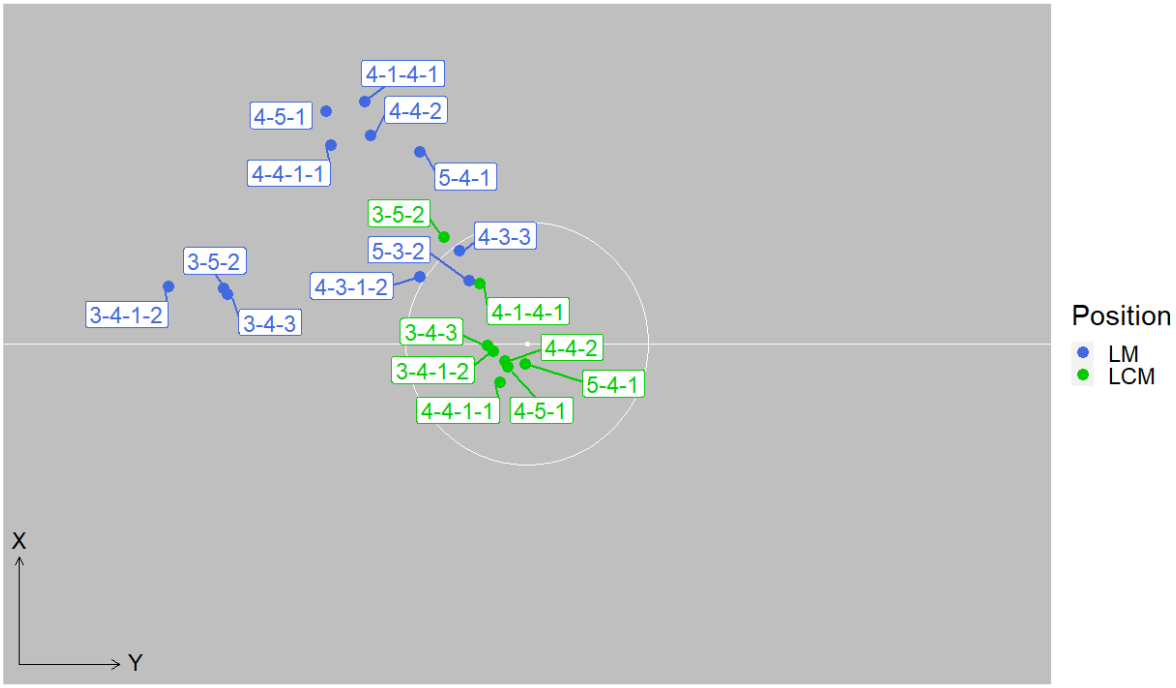

Figure 14: Weighted average locations of the positions left-midfield (LM) and left-centre-midfield (LCM) for different formations before relabeling some standard positions for certain formations (for each average location, the corresponding formation is shown).

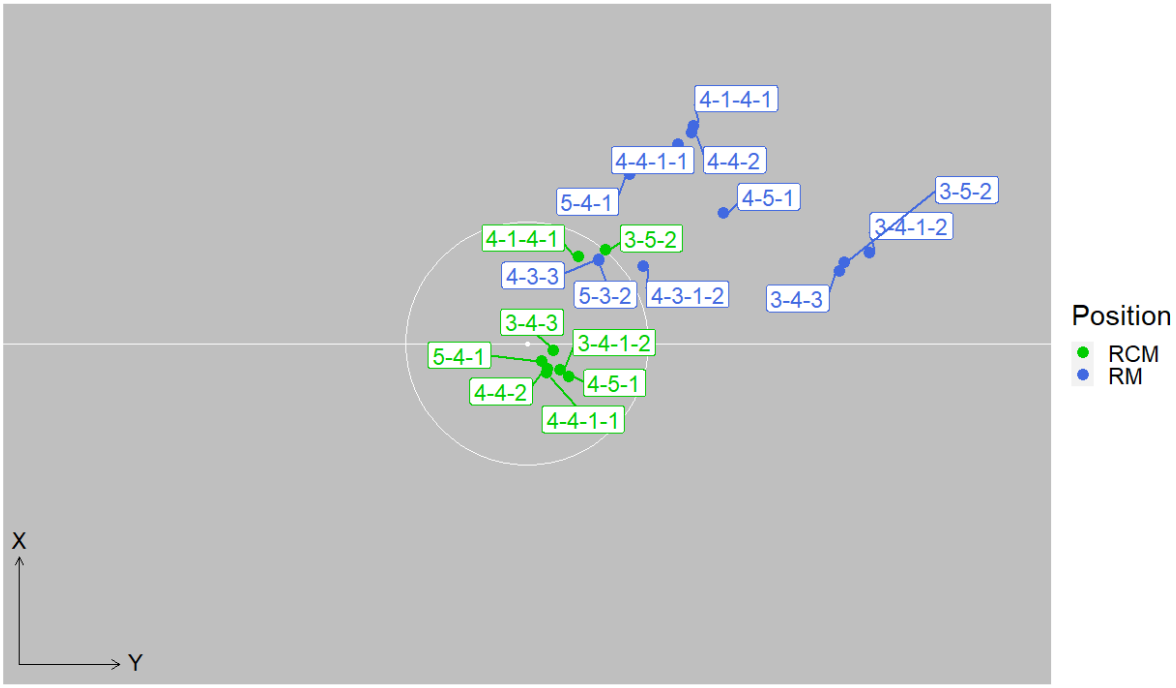

Figure 15: Weighted average locations of the positions right-centre-midfield (RCM) and right-midfield (RM) for different formations before relabeling some standard positions for certain formations (for each average location, the corresponding formation is shown).

Hence, for the formations 4-3-1-2, 4-3-3 and 5-3-2, the players that play in LM and RM are relabeled as LCM and RCM, respectively. It also appears from Figures 14 and 15 that for the positions LM and RM, there is a difference between the weighted average locations for the formations 3-4-1-2, 3-4-3 and 3-5-2, and the weighted average locations for the formations 4-1-4-1, 4-4-1-1, 4-4-2, 4-5-1 and 5-4-1. However, for these positions there is no clear other position to relabel them with. Therefore, the problem created by these differences is resolved by treating players who play in LM or RM in a 3-4-1-2, 3-4-3 or 3-5-2 formation as wing-backs, and treating players who play in LM or RM in a 4-1-4-1, 4-4-1-1, 4-4-2, 4-5-1 or 5-4-1 formation as wingers. Wing-backs and wingers are two of the used position groups in this thesis, which can be found in Table 3 (see Section 3.1.1).

For all other positions we do not observe any clear incorrectness of standard position labels given by ORTEC Sports to players in certain formations.

B Match event data

A description of the match event data is given in Appendix B.1. In Appendix B.2 we explain in detail how the match event data is collected using the computer support system Effectivity in Action (EiA). Thereafter, in Appendix B.3 an overview is given of the 180 statistics that are calculated for each player per match based on the match event data.

B.1 Description of data

The match event data is collected by ORTEC Sports, using the computer support system EiA (Sierksma, 2006). For each football match every match event is annotated, which is every event in which the ball is involved and some other events in which the ball is not involved, such as a substitution. Each observation of the resulting data set corresponds to a unique match event, yielding approximately 1500 to 2000 observations per football match. For every match event in which the ball is involved, we have the following attributes:

- `Match`: The match in which the ball-related match event occurs.
- `Team`: The team to which the player belongs who is involved in the ball-related match event.
- `Player`: The player involved in the ball-related match event.
- `Phase`: The phase of the match in which the ball-related match event occurs (first or second half).
- `Time`: The time of the ball-related match event in milliseconds.
- `Location`: The x- and y-coordinates of the ball-related match event (both on a [0,100] scale).
- `Action`: The type of action of the ball-related match event.
- `Attributes`: Additional attributes that describe the type of action of the ball-related match event in more detail.

In EiA there are 19 predefined types of `Action` for ball-related match events in football, which are listed in Table 27. The difference between "Move" (dribble) and "Attacking action" is that during a move, the player with the ball does not encounter an opponent, whereas during an attacking action the player with the ball makes an attempt to pass an opponent. An overview of the predefined `Attributes` for each `Action` is given in Table 28 (see Appendix B.2).

Note that the `Action` "Foul" is an event in which the ball is not involved. However, it is included in both Tables 27 and 28 since it is an `Action` which has some predefined `Attributes`. Next to the `Action` "Foul", some other events which are not related to ball actions are registered as well (these events however do not have predefined `Attributes`). In effect, it is also registered when a substitution is made, when a player gets booked by the referee with a yellow or red card, and when a player goes out of play and comes back (due to an injury). For each of these events,

Action			
Attacking action	Goal attempt	Offside	Referee ball
Corner	Goal kick	Out of play	Save on goal attempt
Defending action	Indirect free kick	Pass	Throw in
Direct free kick	Interception	Penalty	Touch
Foul	Move	Reception	

Table 27: Types of `Action` in Effectivity in Action (EiA).

the time of the action, the location of the action (if relevant), and the player(s) involved in the action are registered.

Besides the objective information held by the match event data, for every ball-related match event it is assessed by the analysts how effective the action is. This is done by grading the actions on a Likert scale (1 to 5). The grading of actions in EiA is further explained in Appendix B.2.

B.2 Collection of data

The match event data is collected by four ORTEC Sports analysts per match, using the computer support system EiA (Sierksma, 2006). These four analysts annotate every match event, and they are divided into groups of two analysts which are both responsible for gathering the data of one of the two teams playing. Within these groups of two analysts, one analyst observes the match events and registers the location on the field using a touch pad. The time of the event is determined by the timing of the touch. Also, this first analyst tells the second analyst what type of action the match event is and which player is involved in the match event by saying his shirt number. The second analyst then registers the type of action including possible additional attributes, the player involved in the match event, and the grade, which is explained below. The second analyst also checks if the order of the match events is correct.

The grading of ball-related match events in EiA is one of the main added values of the system. Based on its effectiveness, every ball-related match event is graded by an analyst on a Likert scale (1 to 5). The grade of a ball-related match event reflects whether the event had a positive or negative influence on the situation the team is in after the action compared to before the action. The fact that the grading is only based on the effectiveness of an event means that the result of the action is the only thing of importance. A player could thus for example give a good pass, but if it is intercepted by the opponent because the intended receiver performed poorly in the situation, the pass will be given a bad grade. In general, the grades 1 to 5 are given by the following principles:

- **Grade 1** if the action directly leads to a goal against, if a big opportunity to score is missed, if a foul is committed that leads to a red card, or if the goalkeeper concedes a goal that could be easily saved.
- **Grade 2** if the action has a negative impact on the team's situation, such as losing possession, a bad pass, losing a duel, or off target goal attempts.
- **Grade 3** if the action has no impact on the team's situation. Most of the actions are seen as neutral and given this grade, because they consist of duels with a neutral outcome or

passes that reach a teammate.

- **Grade 4** if the action has a positive effect on scoring a goal or preventing the opponent from scoring a goal. Examples are overtaking an opponent, a good forward pass to a teammate, a good defending action or winning possession.
- **Grade 5** if the action is directly related to scoring a goal or preventing the opponent from scoring a goal. Examples are a goal attempt that turns to a goal, an assist for a teammate, the goalkeeper saving a goal attempt with a great effort, a player on the pitch that manages to clear the ball and prevent the opponent from scoring a goal, or creating or preventing one on one situations with the goalkeeper.

The grades of actions are used to calculate grades for players on a 1 to 10 scale for one or multiple matches. These aggregated grades can be used to create an overview of player performances for a single match, or for a longer period consisting of more matches.

In Table 28, an overview of the predefined `Attributes` for each `Action` is given. Each `Action` can have multiple sets of `Attributes`, and one of the `Attributes` from each set is chosen to belong to that `Action`. Note that some `Action`s have multiple combinations of sets of `Attributes`, such that the many different match events that can occur during a match can be captured. Furthermore, "(No attr.)" means that it is possible to select no attribute for that set of `Attributes`.

Action	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7
Attacking action (1)	Body	Duel touched	Overtaking opponent left	Good skill			
	(No attr.)	Duel untouched	Overtaking opponent right	Big chance			
				(No attr.)			
Attacking action (2)	Head	Duel touched	Big chance				
	Sliding player	Duel untouched					
	Body						
	(No attr.)						
Attacking action (3)	Fake pass						
Attacking action (4)	Body	Duel untouched	Shield opponent				
Corner	Right foot	Curved out	High	Goal			
	Left foot	Curved in	Low	(No attr.)			

Table 28: Types of `Action` with their sets of `Attributes` in EiA (continues on next page).

Action	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7
		Straight corner					
Defending action (1)	Blocked cross/-pass/shot	Saved goal	Last line				
	Clearance	Error	(No attr.)				
		(No attr.)					
Defending action (2)	Clearance	Keeper sweeper	Error	Last line			
		Head	(No attr.)	(No attr.)			
Defending action (3)	Positioning	Offside provoked					
		Keeper sweeper					
		(No attr.)					
Defending action (4)	Body	Duel touched	Last line				
	Sliding player	Duel untouched	(No attr.)				
	Head						
	(No attr.)						
Defending action (5)	Body	Duel untouched	Shield opponent				
		Duel touched					
Defending action (6)	Blocked cross/-pass/shot	Direction unchanged					
Direct free kick	Right foot	Goal	High				
	Left foot	Saved by the keeper	(No attr.)				
		On the crossbar					
		On the left/right bar					
		Blocked					
		Off target					
		Over					
Foul	Protest	Goal disallowed					
	Hands						

Table 28: Types of `Action` with their sets of `Attributes` in EiA (continues on next page).

Action	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7
	6 second rule						
	Dangerous play						
	Obstruction						
	Time delay						
	Schwalbe/ simulation						
	(No attr.)						
Goal attempt (1)	Right foot	Goal	Volley	Direct	High	Big chance	Good skill
	Left foot	Saved by the keeper	(No attr.)	(No attr.)	(No attr.)	(No attr.)	(No attr.)
		On the crossbar					
		On the left/right post					
		Blocked					
		Off target					
		Over					
Goal attempt (2)	Head	Goal	High	Duel touched	Big chance	Good skill	
		Saved by the keeper	(No attr.)	(No attr.)	(No attr.)	(No attr.)	
		On the crossbar					
		On the left/right post					
		Blocked					
		Off target					
		Over					
Goal attempt (3)	Body	Untouched		Big chance			
	Head			(No attr.)			
	Left foot		Sliding				
	Right foot		Volley				
Goal kick	Right foot	High	Disallowed goal				
	Left foot	(No attr.)	(No attr.)				

Table 28: Types of `Action` with their sets of `Attributes` in EiA (continues on next page).

Action	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7
	Big chance						
Indirect free kick	Right foot	Cross pass	High	Disallowed goal			
	Left foot	(No attr.)	(No attr.)	Error			
				(No attr.)			
Interception (1)	On deep pass	Caught by the keeper	High	Direction unchanged (only punched)	Error		
	On cross pass	Punched by the keeper	Low		(No attr.)		
		Untouched					
		(No attr.)					
Interception (2)	On deep pass	Smother	Low				
	On cross pass						
Move	Good skill						
	Error						
Offside	Disallowed goal						
Out of play	On the crossbar						
	On the left post						
	On the right post						
	Physio entering the pitch						
	Disallowed goal						
Pass (1)	Right foot	Cross pass	Direct	High	Disallowed goal		
	Left foot	(No attr.)	(No attr.)	(No attr.)	Error		
					(No attr.)		
Pass (2)	Right foot	Launch	Direct	Disallowed goal			
	Left foot		(No attr.)	Error			
Pass (3)	Right foot	Through	Direct	Disallowed goal			
	Left foot		(No attr.)	Error			

Table 28: Types of `Action` with their sets of `Attributes` in EiA (continues on next page).

Action	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7
				(No attr.)			
Pass (4)	Head	Cross pass	High	Disallowed goal			
		(No attr.)	(No attr.)	Error			
				(No attr.)			
Pass (5)	Throw (keeper)	High	Disallowed goal				
	Body	(No attr.)	Error				
			(No attr.)				
Pass (6)	Fair play						
Penalty	Right foot	Goal	Through the center	High			
	Left foot	Saved by the keeper	Left corner				
		On the crossbar	Right corner				
		On the left/right post					
		Blocked					
		Off target					
		Over					
Reception	Right foot	Good skill					
	Left foot	Error					
	Right leg	Big chance					
	Left leg	Untouched					
	Body	(No attr.)					
	Head						
Referee ball	Physio entering the pitch						
Save on goal attempt	Ground	Opponent goal	Standing	Hands	Error		
	In the air	Caught by the keeper	Diving	Right foot	(No attr.)		
		Punched by the keeper		Left foot			
				Body			
Throw in	Error						
	Incorrect						
	Fair play						

Table 28: Types of `Action` with their sets of `Attributes` in EiA (continues on next page).

Action	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7
	(No attr.)						
Touch	Body						
	Head						
	Sliding						

Table 28: Types of `Action` with their sets of `Attributes` in EiA.

B.3 Player statistics

Here we give an overview of the 180 statistics that are calculated for each player per match based on the match event data. These 180 statistics can be divided into 9 different categories, according to which aspect of football each statistic belongs to. In Tables 29 up to 37, we give a description of goalkeeper statistics, passing statistics, duel statistics, goal attempt statistics, possession statistics, goal type and goal attempt type statistics, set play statistics, defensive statistics and other statistics, respectively.

Goalkeeper statistics		
Statistic	Description	Extra information
`saveOnGoalAttempts`	Number of attempted saves on goal attempts by the goalkeeper	This includes saves which are not successful
`savingPercentage`	Percentage of the number of saves on goal attempts by the goalkeeper that are successful	
`saveOnGoalAttemptCaught`	Number of caught saves on goal attempts by the goalkeeper	
`saveOnGoalAttemptPunched`	Number of punched saves on goal attempts by the goalkeeper	
`saveOnGoalAttemptPunched-ToCorner`	Number of punched saves on goal attempts by the goalkeeper that lead to a corner	
`saveOnGoalAttemptPunched-ToGoalAttemptOpponent`	Number of punched saves on goal attempts by the goalkeeper that lead to another goal attempt by the opponent	
`keeperInterception`	Number of interceptions by the goalkeeper	
`keeperInterceptionCaught`	Number of caught interceptions by the goalkeeper	
`keeperInterceptionCaught-OnCross`	Number of caught interceptions on crosses by the goalkeeper	
`keeperInterceptionCaught-OnDeepPass`	Number of caught interceptions on deep passes by the goalkeeper	

Table 29: Overview of the goalkeeper related player statistics that are calculated for each match based on the match event data (continues on next page).

Goalkeeper statistics		
Statistic	Description	Extra information
keeperInterception-Punched	Number of punched interceptions by the goalkeeper	
keeperInterception-PunchedOnCross	Number of punched interceptions on crosses by the goalkeeper	
keeperInterception-PunchedOnDeepPass	Number of punched interceptions on deep passes by the goalkeeper	
possessionRegainInPlayBy-KeeperInterception	Number of times possession is regained in play by an interception of the goalkeeper	
keeperThrow	Number of keeper throws	Number of passes the goalkeeper makes by throwing the ball with his hands
keeperThrowLong	Number of long keeper throws	A keeper throw is considered to be long if it is a throw over more than 30 metres
keeperThrowShort	Number of short keeper throws	A keeper throw is considered to be short if it is a throw over less than 20 metres

Table 29: Overview of the goalkeeper related player statistics that are calculated for each match based on the match event data.

Passing		
Statistic	Description	Extra information
passes	Number of passes	
completedPasses	Number of completed passes	
completedPassPercentage	Percentage of the number of passes that is completed	
passesForward	Number of forward passes	A pass is considered to be a forward pass if it has a direction angle smaller than 77.5 degrees or larger than 282.5 degrees (0 degrees is straight forward)
completedPassesForward	Number of completed forward passes	
completedPassesForward-Percentage	Percentage of the number of forward passes that is completed	
passesForwardPercentage	Percentage of the number of passes that is forward	

Table 30: Overview of the passing related player statistics that are calculated for each match based on the match event data (continues on next page).

Passing		
Statistic	Description	Extra information
passesWide	Number of wide passes	A pass is considered to be a wide pass if it has a direction angle between 77.5 and 102.5 degrees (pass to the right) or if it has a direction angle between 257.5 and 282.5 degrees (pass to the left)
passesBackward	Number of backward passes	A pass is considered to be a backward pass if it has a direction angle between 102.5 and 257.5 degrees
keyPasses	Number of key passes	A pass is considered to be a key pass if it is a key action (an action is labeled as a key action if it is the action before a goal attempt by a teammate, without a ball action of the opponent in between)
longPasses	Number of long passes	A pass is considered to be a long pass if it is a pass over 30 meters or more
passesOwnHalf	Number of passes given on own half	
passesOpponentHalf	Number of passes given on opponent half	
completedPassesOwnHalf	Number of completed passes on own half	
completedPassesOpponent-Half	Number of completed passes on opponent half	
completedPassPercentage-OwnHalf	Percentage of the number of passes given on the own half that is completed	
completedPassPercentage-OpponentHalf	Percentage of the number of passes given on the opponent half that is completed	
passesFinalThird	Number of passes in final third	The final third is the last 30 metres of the football field (this corresponds to the x-coordinate being in the range [70,100])
completedPassesFinal-Third	Number of completed passes in final third	

Table 30: Overview of the passing related player statistics that are calculated for each match based on the match event data (continues on next page).

Passing		
Statistic	Description	Extra information
completedPassPercentage-FinalThird	Percentage of the passes in the final third that is completed	
passBetweenCentral-Defenders	Number of passes between two central defenders	
passFirstInPossession	Number of first passes in a possession moment	
passFirstInPossession-Forward	Number of first passes in a possession moment that are forward passes	
passFirstInPossession-ForwardPercentage	Percentage of the number of first passes in a possession moment that is forward	
receivedFirstPassIn-Possession	Number of received first passes in a possession moment	
passToBox	Number of passes that end in the penalty area of the opponent	
crossPasses	Number of cross passes given	A cross pass is a pass from the opponent half on the flank of the field with the intention to reach a teammate in the penalty area. Cross passes that are blocked by defenders are also tagged
crossPassesCompleted	Number of completed cross passes	
crossPassesToGoalAttempt	Number of cross passes leading directly to a goal attempt	
crossPassesToGoal	Number of cross passes leading directly to a goal	
crossPassEarly	Number of early cross passes	A cross pass is considered to be given early if it is given before the horizontal line of the penalty area (this corresponds to the x-coordinate being in the range [0,83])
crossPassLate	Number of late cross passes	A cross pass is considered to be given late if it is given after the horizontal line of the penalty area (this corresponds to the x-coordinate being in the range [83,100])
crossPassHigh	Number of cross passes given through the air	

Table 30: Overview of the passing related player statistics that are calculated for each match based on the match event data (continues on next page).

Passing		
Statistic	Description	Extra information
`crossPassLow`	Number of cross passes given over the ground	
`crossPassLeftEarlyHigh`	Number of cross passes given early through the air from the left flank	A cross pass is considered to be given early if it is given before the horizontal line of the penalty area (this corresponds to the x-coordinate being in the range [0,83])
`crossPassLeftEarlyLow`	Number of cross passes given early over the ground from the left flank	
`crossPassLeftLateHigh`	Number of cross passes given late through the air from the left flank	A cross pass is considered to be given late if it is given after the horizontal line of the penalty area (this corresponds to the x-coordinate being in the range [83,100])
`crossPassLeftLateLow`	Number of cross passes given late over the ground from the left flank	
`crossPassRightEarlyHigh`	Number of cross passes given early through the air from the right flank	
`crossPassRightEarlyLow`	Number of cross passes given early over the ground from the right flank	
`crossPassRightLateHigh`	Number of cross passes given late through the air from the right flank	
`crossPassRightLateLow`	Number of cross passes given late over the ground from the right flank	

Table 30: Overview of the passing related player statistics that are calculated for each match based on the match event data.

Duels		
Statistic	Description	Extra information
`totalNrOfDuels`	Total number of duels	
`lostDuels`	Number of duels lost	
`dribbles`	Number of dribbles made	
`slidings`	Number of sliding tackles made	

Table 31: Overview of the duel related player statistics that are calculated for each match based on the match event data (continues on next page).

Duels		
Statistic	Description	Extra information
standingDuels	Number of ground duels	
standingDuelsWon	Number of ground duels won	
standingDuelsWon-Percentage	Percentage of the number of ground duels that is won	
airDuels	Number of air duels	Air duels are duels in which one of the players touches the ball with his head
airDuelsWon	Number of air duels won	
airDuelsWonPercentage	Percentage of the number of air duels that is won	If the player had at least one air duel, airDuelsWonPercentage is defined as the number of air duels won divided by the number of air duels. Otherwise, airDuelsWonPercentage is equal to 0
defensiveDuels	Number of defensive duels	Defensive duels are duels by players of the team that is not in possession of the ball
defensiveDuelsWon	Number of defensive duels won	
defensiveDuelsWon-Percentage	Percentage of the number of defensive duels that is won	If the player had at least one defensive duel, defensiveDuels-WonPercentage is defined as the number of defensive duels won divided by the number of defensive duels. Otherwise, defensiveDuelsWonPercentage is equal to 0
defensiveGroundStanding-Duels	Number of defensive ground duels	Ground duels are duels in which none of the players touches the ball with his head
defensiveGroundStanding-DuelsWon	Number of defensive ground duels won	

Table 31: Overview of the duel related player statistics that are calculated for each match based on the match event data (continues on next page).

Duels		
Statistic	Description	Extra information
defensiveGroundStanding-DuelsWonPercentage	Percentage of the number of defensive ground duels that is won	If the player had at least one defensive ground duel, defensiveGroundStanding-DuelsWonPercentage is defined as the number of defensive ground duels won divided by the number of defensive ground duels. Otherwise, defensive-GroundStandingDuelsWon-Percentage is equal to 0
defensiveAirDuels	Number of defensive air duels	Air duels are duels in which one of the players touches the ball with his head
defensiveAirDuelsWon	Number of defensive air duels won	
defensiveAirDuelsWon-Percentage	Percentage of the number of defensive air duels that is won	If the player had at least one defensive air duel, defensive-AirDuelsWonPercentage is defined as the number of defensive air duels won divided by the number of defensive air duels. Otherwise, defensive-AirDuelsWonPercentage is equal to 0
defensiveDuelsOwnHalf	Number of defensive duels on own half	
defensiveDuelsOpponent-Half	Number of defensive duels on opponent half	
defensiveDuelsOwnBox	Number of defensive duels in own penalty area	
defensiveDuelsOwnBoxWon	Number of defensive duels in own penalty area won	

Table 31: Overview of the duel related player statistics that are calculated for each match based on the match event data (continues on next page).

Duels		
Statistic	Description	Extra information
defensiveDuelsOwnBoxWonPercentage	Percentage of the number of defensive duels in the own penalty area that is won	If the player had at least one defensive duel in the own penalty area, defensiveDuelsOwnBoxWonPercentage is defined as the number of defensive duels in the own penalty area won divided by the number of defensive duels in the own penalty area. Otherwise, defensiveDuelsOwnBoxWonPercentage is equal to 0
attackingDuels	Number of offensive duels	Offensive duels are duels by players of the team that is in possession of the ball
attackingDuelsWon	Number of offensive duels won	
attackingDuelsWonPercentage	Percentage of the number of offensive duels that is won	If the player had at least one offensive duel, attackingDuelsWonPercentage is defined as the number of offensive duels won divided by the number of offensive duels. Otherwise, attackingDuelsWonPercentage is equal to 0
attackingAirDuels	Number of offensive air duels	
attackingAirDuelsWon	Number of offensive air duels won	
attackingAirDuelsWonPercentage	Percentage of the number of offensive air duels that is won	If the player had at least one offensive air duel, attackingAirDuelsWonPercentage is defined as the number of offensive air duels won divided by the number of offensive air duels. Otherwise, attackingAirDuelsWonPercentage is equal to 0
attackingGroundStandingDuels	Number of offensive ground duels	
attackingGroundStandingDuelsWon	Number of offensive ground duels won	

Table 31: Overview of the duel related player statistics that are calculated for each match based on the match event data (continues on next page).

Duels		
Statistic	Description	Extra information
attackingGroundStanding-DuelsWonPercentage	Percentage of the number of offensive ground duels that is won	If the player had at least one offensive ground duel, attackingGroundStanding-DuelsWonPercentage is defined as the number of offensive ground duels won divided by the number of offensive ground duels. Otherwise, attacking-GroundStandingDuelsWon-Percentage is equal to 0

Table 31: Overview of the duel related player statistics that are calculated for each match based on the match event data.

Goal attempts		
Statistic	Description	Extra information
goals	Number of goals scored	
assists	Number of assists	An action is labeled as an assist if it is the action before a successful goal attempt by a teammate, without a ball action of the opponent in between
keyActions	Number of key actions	An action is labeled as a key action if it is the action before a goal attempt by a teammate, without a ball action of the opponent in between
goalAttempts	Number of goal attempts	The number of goal attempts is the sum of the number of shots, the number of free kicks attempted to be shot directly on goal, and the number of penalties taken
goalAttemptConversion	Conversion rate of the goal attempts	If the player had at least one goal attempt, goalAttempt-Conversion is defined as the number of goals scored divided by the number of goal attempts. Otherwise, goalAttempt-Conversion is equal to 0

Table 32: Overview of the goal attempt related player statistics that are calculated for each match based on the match event data (continues on next page).

Goal attempts		
Statistic	Description	Extra information
goalAttemptsOnTarget	Number of goal attempts on target	
goalAttemptsInsidePenaltyBox	Number of goal attempts from inside the penalty area	
goalAttemptsOutsidePenaltyBox	Number of goal attempts from outside the penalty area	
blockedGoalAttempts	Number of blocked goal attempts	
shots	Number of shots	
shotsOnTarget	Number of shots on target	Total number of shots scored or saved by keeper
shotsWide	Number of shots that go wide	
shotsOver	Number of shots that go over	
shotsOnPost	Number of shots that hit the post	
shotsOnCrossbar	Number of shots that hit the crossbar	
shotsBlocked	Number of blocked shots	
shotsInsidePenaltyBox	Number of shots from inside the penalty area	
shotsOutsidePenaltyBox	Number of shots from outside the penalty area	
shotsInsidePenaltyBoxGoal	Number of shots from inside the penalty area that lead to a goal	
shotsOutsidePenaltyBoxGoal	Number of shots from outside the penalty area that lead to a goal	
shotWithHead	Number of headed shots	
shotWithHeadGoal	Number of headed shots that lead to a goal	
penalties	Number of penalties taken	
penaltiesScored	Number of penalties scored	The penalty is scored by the player
penaltiesMissed	Number of penalties missed	The penalty is not scored by the player, it is either saved or shot over/wide
penaltiesOnGoal	Number of penalties on goal	The penalty is shot on goal, so it is either scored or saved
freeKicks	Number of free kicks taken	Total number of free kicks taken, so the sum of the number of free kicks on goal and the number of free kicks not on goal

Table 32: Overview of the goal attempt related player statistics that are calculated for each match based on the match event data (continues on next page).

Goal attempts		
Statistic	Description	Extra information
freeKickNotOnGoal	Number of free kicks not attempted to be shot directly on goal	
freeKickOnGoal	Number of free kicks attempted to be shot directly on goal	
freeKickOnGoalOnTarget	Number of free kicks attempted to be shot directly on goal that are on target	
freeKickOnGoalWide	Number of free kicks attempted to be shot directly on goal that go wide	
freeKickOnGoalOver	Number of free kicks attempted to be shot directly on goal that go over	
freeKickOnGoalOnPost	Number of free kicks attempted to be shot directly on goal that hit the post	
freeKickOnGoalOnCrossbar	Number of free kicks attempted to be shot directly on goal that hit the crossbar	
freeKickOnGoalBlocked	Number of free kicks attempted to be shot directly on goal that are blocked	
freeKickOnGoalScored	Number of free kicks attempted to be shot directly on goal that lead to a goal	
freeKickOnGoalMissed	Number of free kicks attempted to be shot directly on goal that are missed	
totalOccurrencesOfBall-ActionInPossessionWith-Goal	Number of times the player is involved in a possession moment from which is scored	
totalOccurrencesOfBall-ActionInPossessionWith-GoalAttempt	Number of times the player is involved in a possession moment from which a goal attempt has been made	
totalTeamGoalAttempts	Total number of goal attempts made by the team	
totalTeamGoalsScored	Total number of goals scored by the team	

Table 32: Overview of the goal attempt related player statistics that are calculated for each match based on the match event data (continues on next page).

Goal attempts		
Statistic	Description	Extra information
shareInGoalPercentage	Percentage of the total number of goals scored by the team for which the player was involved in the possession moment from which the goal was scored	If the total number of goals scored by the team is at least one, shareInGoalPercentage is defined as the number of times the player is involved in a possession moment from which is scored divided by the total number of goals scored by the team. Otherwise, shareInGoal-Percentage is equal to 0
shareInGoalAttempts-Percentage	Percentage of the total number of goal attempts made by the team for which the player was involved in the possession moment from which the goal attempt was made	If the total number of goal attempts made by the team is at least one, shareInGoal-AttemptsPercentage is defined as the number of times the player is involved in a possession moment from which a goal attempt has been made divided by the total number of goal attempts made by the team. Otherwise, shareInGoal-AttemptsPercentage is equal to 0

Table 32: Overview of the goal attempt related player statistics that are calculated for each match based on the match event data.

Possession		
Statistic	Description	Extra information
possessionLoss	Number of times possession is lost	Ball possession moments are determined by an algorithm of ORTEC Sports. The last action of a possession moment is labeled as possession lost.
possessionLossByDuel	Number of times posession is lost by a duel	
possessionLossByPass	Number of times possession is lost by a pass	
possessionLossByOther	Number of times possession is lost by other actions	

Table 33: Overview of the possession related player statistics that are calculated for each match based on the match event data (continues on next page).

Possession		
Statistic	Description	Extra information
possessionRegainInPlay	Number of times possession is regained in play	Ball possession moments are determined by an algorithm of ORTEC Sports. The first action of a possession moment that is not a dead moment (corner, free kick, goal kick, kick-off or throw in) is labeled as possession regain.
possessionRegainInPlayBy-Duel	Number of times possession is regained in play by a duel	
possessionRegainInPlayBy-Interception	Number of times possession is regained in play by an interception	
possessionRegainOwnHalf	Number of times possession is regained on own half	
possessionRegainOpponent-Half	Number of times possession is regained on the opponent half	

Table 33: Overview of the possession related player statistics that are calculated for each match based on the match event data.

Goal and goal attempt types		
Statistic	Description	Extra information
goalBuildUp	Number of goals scored from a build up play	A play is considered to be a build up play if the possession of the ball starts on the own half and if the play contains 5 or more passes
goalCounter	Number of goals scored from a counter play	A play is considered to be a counter play if the possession of the ball starts on the own half and if the play contains less than 5 passes
goalOffensive	Number of goals scored from an offensive play	A play is considered to be an offensive play if the possession of the ball starts on the opponent half and if the play contains 5 or more passes
goalSetPlay	Number of goals scored from a set play	A goal is considered to be scored from a set play if the goal is scored within 2 passes after a set play

Table 34: Overview of the goal type and goal attempt type related player statistics that are calculated for each match based on the match event data (continues on next page).

Goal and goal attempt types		
Statistic	Description	Extra information
goalTurnOver	Number of goals scored from a turn over play	A play is considered to be a turn over play if the possession of the ball starts on the opponent half and if the play contains less than 5 passes
goalAttemptBuildUp	Number of goal attempts from a build up play	A play is considered to be a build up play if the possession of the ball starts on the own half and if the play contains 5 or more passes
goalAttemptCounter	Number of goal attempts from a counter play	A play is considered to be a counter play if the possession of the ball starts on the own half and if the play contains less than 5 passes
goalAttemptOffensive	Number of goal attempts from an offensive play	A play is considered to be an offensive play if the possession of the ball starts on the opponent half and if the play contains 5 or more passes
goalAttemptSetPlay	Number of goal attempts from a set play	A goal attempt is considered to be attempted from a set play if the attempt is made within 2 passes after a set play
goalAttemptTurnOver	Number of goal attempts from a turn over play	A play is considered to be a turn over play if the possession of the ball starts on the opponent half and if the play contains less than 5 passes

Table 34: Overview of the goal type and goal attempt type related player statistics that are calculated for each match based on the match event data.

Set plays		
Statistic	Description	Extra information
throwIns	Number of throw-ins taken	
corners	Number of corners taken	
cornersLeftSide	Number of corners taken from the left side	
cornersRightSide	Number of corners taken from the right side	
cornersShort	Number of corners taken short	

Table 35: Overview of the set play related player statistics that are calculated for each match based on the match event data (continues on next page).

Set plays		
Statistic	Description	Extra information
`goalKicks`	Number of goal kicks taken	
`offsides`	Number of offsides	

Table 35: Overview of the set play related player statistics that are calculated for each match based on the match event data.

Defensive statistics		
Statistic	Description	Extra information
`fouls`	Number of fouls committed	
`foulsOwnHalf`	Number of fouls committed on own half	The foul is committed in a position for which it holds that $x \in [0,50]$
`foulsSuffered`	Number of fouls suffered	The number of times a foul is made against the player
`yellowCards`	Number of yellow cards received	
`redCards`	Number of red cards received	
`directRedCards`	Number of direct red cards received	
`foulsPerCard`	Number of fouls committed per received card	If the player received at least one card, `foulsPerCard` is defined as the number of fouls committed divided by the sum of the number of yellow cards received and the number of red cards received. Otherwise, `foulsPerCard` is equal to 0
`clearances`	Number of clearances	An action is labeled as clearance if the player clears the ball without having the intention to reach a teammate.
`cleansheets`	Logical indicating whether the team has conceded no goals	
`positioningError`	Number of positional errors	A positional error is given when a player allows the opponent to freely give a cross or make a goal attempt

Table 36: Overview of the defensive related player statistics that are calculated for each match based on the match event data.

Other statistics		
Statistic	Description	Extra information
`nrOfBallActions`	Number of ball-related match events	
`actionsInOpponentBox`	Number of ball-related match events in the penalty area of the opponent	
`inStartingLineup`	Logical indicating whether the player is in the starting lineup	
`substitutionsIn`	Logical indicating whether the player was brought on as a substitute	
`substitutionsOut`	Logical indicating whether the player was substituted for another player	
`nrOfPlayedMinutes`	Number of minutes played	
`averageLocationX`	The average location of the player in the x-direction on a [0,100] scale	
`averageLocationY`	The average location of the player in the y-direction on a [0,100] scale	
`grade`	The grade of the player on a [1,10] scale, which is calculated based on the grades the player got for each ball-related match event he was involved in	The grading of players is explained in detail in Appendix B.2

Table 37: Overview of the other player statistics that are calculated for each match based on the match event data.

C Selection of player statistics used for clustering

After aggregating the statistics for each player in each position group over multiple matches by taking the sum of that statistic over the multiple matches, for each position group we select the aggregated statistics that we find relevant for describing play styles of players in that position group. Each aggregated statistic is taken relative to or as a percentage of another statistic, because we think that relative statistics better reflect the choices a player makes than absolute statistics. The resulting selected player statistics for central midfielders and centre-forwards can be found in Table 38, with for each statistic the two statistics that are taken relative to each other to calculate the resulting statistic. In addition, for central midfielders we include the following statistics:

* `passesOwnHalfPercentageCompleted`, which is defined as the number of completed passes on the own half, taken relative to the number of passes on the own half.
* `passesOpponentHalfPercentageCompleted`, which is defined as the number of completed passes on the opponent half, taken relative to the number of passes on the opponent half.
* `shareInPassFirstInPossessionPercentage`, which is defined as the number of first passes in a possession moment, taken relative to the number of passes in a possession moment by the team in total.
* `passFirstInPossessionPercentageForward`, which is defined as the number of forward first passes in a possession moment, taken relative to the number of first passes in a possession moment.

Furthermore, for centre-forwards we include the statistic `offsidesPerBallActionPercentage`, which is defined as the number of offsides divided by the number of ball-related match events (this statistic is not included for central midfielders as the number of offsides is rather low for these players).

Selected relative statistic	Original absolute statistic	Taken relative to
`shareInBallActions-Percentage`	Number of ball-related match events	Number of ball-related match events by the team in total
`shareInPassesPercentage`	Number of passes	Number of passes by the team in total
`passesPercentageCompleted`	Number of completed passes	Number of passes
`passesPercentageForward`	Number of forward passes	Number of passes
`passesForwardPercentage-Completed`	Number of completed forward passes	Number of forward passes
`passesPercentageWide`	Number of wide passes	Number of passes
`passesPercentageLong`	Number of long passes	Number of passes

Table 38: Overview of the selected player statistics for central midfielders and centre-forwards (continues on next page).

Selected relative statistic	Original absolute statistic	Taken relative to
keyPassesPerBallAction-Percentage	Number of key passes	Number of ball-related match events
shareInKeyPassesPercentage	Number of key passes	Number of key passes by team in total
passesPercentageOpponent-Half	Number of passes on opponent half	Number of passes
passesPercentageFinalThird	Number of passes in final third	Number of passes
passesFinalThirdPercentage-Completed	Number of completed passes in final third	Number of passes in final third
shareInReceivedFirstPassIn-PossessionPercentage	Number of received first passes in a possession moment	Number of received first passes in a possession moment by the team in total
passesPercentageToBox	Number of passes that end in the penalty area of the opponent	Number of passes
passesPercentageCrosses	Number of cross passes	Number of passes
crossPassesPercentage-Completed	Number of completed cross passes	Number of cross passes
crossPassesPercentageTo-GoalAttempt	Number of cross passes leading directly to a goal attempt	Number of cross passes
crossPassesPercentageLate	Number of late cross passes	Number of cross passes
crossPassesPercentageHigh	Number of cross passes given through the air	Number of cross passes
duelsPercentageWon	Number of duels won	Number of duels
dribblesPerBallAction-Percentage	Number of dribbles made	Number of ball-related match events
slidingsPerDuel	Number of sliding tackles made	Number of duels
groundDuelsPercentageWon	Number of ground duels won	Number of ground duels
airDuelsPercentageWon	Number of air duels won	Number of air duels
defensiveDuelsPercentage-Won	Number of defensive duels won	Number of defensive duels
defensiveDuelsPercentage-OwnHalf	Number of defensive duels on own half	Number of defensive duels

Table 38: Overview of the selected player statistics for central midfielders and centre-forwards (continues on next page).

Selected relative statistic	Original absolute statistic	Taken relative to
attackingDuelsPercentage-Won	Number of offensive duels won	Number of offensive duels
keyActionsPerBallAction-Percentage	Number of key actions	Number of ball-related match events
shareInKeyActions-Percentage	Number of key actions	Number of key actions by the team in total
goalAttemptsPerBallAction-Percentage	Number of goal attempts	Number of ball-related match events
shareInGoalAttempts-Percentage	Number of goal attempts	Number of goal attempts by the team in total
goalAttemptsPercentage-InsidePenaltyBox	Number of goal attempts from inside the penalty area of the opponent	Number of goal attempts
shotsPerBallAction-Percentage	Number of shots	Number of ball-related match events
shareInShotsPercentage	Number of shots	Number of shots by the team in total
shotsPercentageInside-PenaltyBox	Number of shots from inside the penalty area of the opponent	Number of shots
shotsPercentageWithHead	Number of headed shots	Number of shots
shareInPossessionWithGoal-AttemptsPercentage	Number of times the player is involved in a possession moment from which a goal attempt has been made	Number of goal attempts by the team in total
shareInPossessionWithGoals-Percentage	Number of times the player is involved in a possession moment from which a goal has been scored	Number of goals by the team in total
possessionLossPerBall-ActionPercentage	Number of times possession is lost	Number of ball-related match events
possessionRegainInPlayPer-BallActionPercentage	Number of times possession is regained in play	Number of ball-related match events
possessionRegainInPlay-PercentageByInterception	Number of times possession is regained in play by an interception	Number of times possession is regained in play

Table 38: Overview of the selected player statistics for central midfielders and centre-forwards (continues on next page).

Selected relative statistic	Original absolute statistic	Taken relative to
possessionRegainInPlay-PercentageOpponentHalf	Number of times possession is regained in play on the opponent half	Number of times possession is regained in play
foulsPerBallAction-Percentage	Number of fouls committed	Number of ball-related match events
foulsPercentageOwnHalf	Number of fouls committed on own half	Number of fouls committed
foulsSufferedPerBallAction-Percentage	Number of fouls suffered	Number of ball-related match events
cardsPerFoul	Number of cards received	Number of fouls committed
actionsPercentageIn-OpponentBox	Number of ball-related match events in the penalty area of the opponent	Number of ball-related match events

Table 38: Overview of the selected player statistics for central midfielders and centre-forwards.

D Reduced k-means

When performing Reduced k-means, we use the alternating least-squares (ALS) algorithm to minimize the objective function

$$f(\mathbf{Z}, \mathbf{L}) = \text{tr}(\mathbf{X}'\mathbf{X}) - \text{tr}(\mathbf{L}'\mathbf{X}'\mathbf{P}\mathbf{X}\mathbf{L}), \tag{15}$$

over $\mathbf{Z}$ and $\mathbf{L}$, with $\mathbf{Z}$ a binary $(N \times k)$ matrix that indicates for each of the N players to which of the k clusters he belongs, and $\mathbf{L}$ a $(q \times p)$ orthonormal matrix that contains the laodings to transform the variables in the original q-dimensional space to the reduced p-dimensional space. The ALS algorithm is given below in Algorithm 1.

Algorithm 1 Alternating least-squares (ALS) for Reduced k-means

Required parameters/functions:

- $\mathbf{X}$: An $(N \times q)$ data matrix, of which we want to cluster the N players in a subspace of the q columns.
- k: The number of clusters.
- p: The lower dimension to which we want to transform the q columns of $\mathbf{X}$ (it thus holds that $p < q$).

def `ALS`($\mathbf{X}$, k, p):

1: Initialize $\mathbf{Z}$ by randomly assigning each player to one of the k clusters.
2: Calculate the $(N \times N)$ matrix $\mathbf{P}$ as $\mathbf{P} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$.
3: **while** Convergence is not reached yet **do**
4: Perform an eigendecomposition on $\mathbf{X}'\mathbf{P}\mathbf{X}$, and let the $(q \times p)$ matrix $\mathbf{L}$ contain the orthonormal eigenvectors that correspond to the p largest eigenvalues.
5: Apply k-means to $\mathbf{X}\mathbf{L}$, and let the resulting cluster allocation be given by the $(N \times k)$ matrix $\mathbf{Z}$.
6: **if** $\mathbf{Z}$ has not changed **then**
7: Convergence is reached.
8: **else**
9: Convergence is not reached yet.
10: Update $\mathbf{P}$ by calculating $\mathbf{P} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$.
11: **end if**
12: **end while**
13: **return** $\mathbf{Z}, \mathbf{L}$

E Dimension reduction results

In this Appendix we show additional results for the approach that consists of PCA to reduce the dimension of a data set consisting of player match statistics, and for the Reduced k-means approach to reduce the dimension of a data set consisting of player match statistics. These results are shown in Section E.1 for central midfielders, and in Section E.2 for centre-forwards.

E.1 Central midfielders

In Section E.1.1 we show the resulting PCA loadings for centre-forwards, and in Section E.1.2 we show the resulting loadings of Reduced k-means for centre-forwards.

E.1.1 PCA

For central midfielders, the loadings of the 4 varimax rotated principal components are displayed in Tables 39 and 40. For each varimax rotated principal component, the loadings are shown for the 10 variables that have the highest absolute loading. The first varimax rotated principal component can be interpreted as how likely a central midfielders is to create chances for teammates to score goals, and how weak he is in duels. Hence, we label the first varimax rotated principal component as "creating chances and duel weakness". The second varimax rotated principal component can be interpreted as how involved a central midfielder is in the play of his team when they are in possession of the ball, how risky his passing is, and how much he operates outside of the penalty area of the opposing team. Risky passes are passes that are typically forward and/or long passes. The second varimax rotated principal component is therefore labeled as "involvement, risky passing and operating outside penalty box".

The third varimax rotated principal component can be interpreted as how simple the passes of a player are. Simple passes are passes that are typically short and either wide or backwards. We therefore label the third component as "simple passing". The fourth varimax rotated principal component can be interpreted as how likely a central midfielder is to make a goal attempt or to be involved in the possession moment leading to a goal attempt, and how strong he is in duels. Hence, we label the fourth varimax rotated principal component as "(involvement in) goal attempts and duel strength".

PCA Central midfielders			
PC1 (Creating chances and duel weakness)		PC2 (Involvement, risky passing and operating outside penalty box)	
Variable	Loading	Variable	Loading
keyActionsPerBallAction-Percentage	0.29	shareInPassesPercentage	0.32
keyPassesPerBallAction-Percentage	0.28	shareInBallActionsPercentage	0.32
shareInKeyPassesPercentage	0.28	shareInReceivedFirstPassIn-PossessionPercentage	0.29
shareInKeyActionsPercentage	0.27	shareInPassFirstInPossession-Percentage	0.26
duelsPercentageWon	-0.27	passesPercentageLong	0.24
possessionRegainInPlayPerBall-ActionPercentage	-0.23	passesPercentageForward	0.23
groundDuelsPercentageWon	-0.22	goalAttemptsPercentageInside-PenaltyBox	-0.23
airDuelsPercentageWon	-0.22	shotsPercentageInsidePenaltyBox	-0.23
passesPercentageOpponentHalf	0.22	shareInPossessionWithGoal-AttemptsPercentage	0.22
defensiveDuelsPercentageWon	-0.21	actionsPercentageInOpponentBox	-0.21

Table 39: Loadings of the first two varimax rotated principal components for central midfielders.

PCA Central midfielders			
PC3 (Simple passing)		PC4 ((Involvement in) goal attempts and duel strength)	
Variable	Loading	Variable	Loading
passesPercentageCompleted	0.35	shareInShotsPercentage	0.37
passesOpponentHalfPercentage-Completed	0.34	shareInGoalAttemptsPercentage	0.37
passesForwardPercentage-Completed	0.33	goalAttemptsPerBallAction-Percentage	0.33
passesOwnHalfPercentage-Completed	0.33	shotsPerBallActionPercentage	0.33
passesFinalThirdPercentage-Completed	0.32	shareInPossessionWithGoal-AttemptsPercentage	0.26
possessionLossPerBallAction-Percentage	-0.24	defensiveDuelsPercentageWon	0.25
passesPercentageWide	0.24	shareInPossessionWithGoals-Percentage	0.21
passesPercentageForward	-0.21	attackingDuelsPercentageWon	0.19
passFirstInPossessionPercentage-Forward	-0.20	actionsPercentageInOpponentBox	0.19
passesPercentageLong	-0.19	groundDuelsPercentageWon	0.18

Table 40: Loadings of the third and fourth varimax rotated principal component for central midfielders.

E.1.2 Reduced k-means

For central midfielders, the loadings of the four varimax rotated reduced variables that follow from applying Reduced k-means with $k = 5$ and $p = 4$ are displayed in Tables 41 and 42. For each varimax rotated reduced variable, the loadings are shown for the 10 original variables that have the highest absolute loading. The first varimax rotated reduced variable can be interpreted as how offensive a central midfielder plays, in terms of where he positions himself on the football pitch, and how often he tries to score a goal. It also captures how low the defensive effort of a central midfielder is. Hence, we label the first varimax rotated reduced variable as "offensive play and low defensive effort". The second varimax rotated reduced variable can be interpreted as how involved a central midfielder is in the play of his team when they are in possession of the ball. The second varimax rotated reduced variable is therefore labeled as "involvement".

Reduced k-means with $k = 5$ and $p = 4$ Central midfielders			
RV1 (Offensive play and low defensive effort)		RV2 (Involvement)	
Variable	Loading	Variable	Loading
actionsPercentageInOpponentBox	0.35	shareInReceivedFirstPassIn-PossessionPercentage	0.37
possessionRegainInPlayPerBall-ActionPercentage	-0.31	shareInBallActionsPercentage	0.34
goalAttemptsPerBallAction-Percentage	0.29	shareInPassesPercentage	0.33
possessionLossPerBallAction-Percentage	0.28	shareInPossessionWithGoal-AttemptsPercentage	0.31
shotsPerBallActionPercentage	0.28	shareInKeyActionsPercentage	0.27
passFirstInPossessionPercentage-Forward	-0.22	shareInPassFirstInPossession-Percentage	0.26
passesPercentageOpponentHalf	0.21	passesPercentageLong	0.25
passesPercentageForward	-0.21	shareInPossessionWithGoals-Percentage	0.22
passesPercentageFinalThird	0.20	goalAttemptsPercentageInside-PenaltyBox	-0.19
shareInShotsPercentage	0.20	shotsPercentageInsidePenaltyBox	-0.19

Table 41: Loadings of the first two varimax rotated reduced variables for central midfielders that result from applying Reduced k-means with $k = 5$ and $p = 4$.

The third varimax rotated reduced variable can be interpreted as how much a central midfielder dribbles, and how simple his passes are. We therefore label the third varimax rotated reduced variable as "dribbling and simple passing". The fourth varimax rotated reduced variable can be interpreted as how likely a central midfielder is to create chances for teammates to score goals. Hence, the fourth varimax rotated principal component is labeled as "creating chances".

For central midfielders, the loadings of the five varimax rotated reduced variables that follow from applying Reduced k-means with $k = 6$ and $p = 5$ are displayed in Tables 43, 44 and 45. For each varimax rotated reduced variable, the loadings are shown for the 10 original variables that have the highest absolute loading. The first varimax rotated reduced variable can be interpreted

Reduced k-means with $k = 5$ and $p = 4$ Central midfielders			
RV3 (Dribbling and simple passing)		RV4 (Creating chances)	
Variable	Loading	Variable	Loading
dribblesPerBallActionPercentage	0.32	shareInKeyPassesPercentage	0.31
passesFinalThirdPercentage-Completed	0.30	shareInKeyActionsPercentage	0.31
passesOpponentHalfPercentage-Completed	0.30	keyActionsPerBallAction-Percentage	0.29
passesPercentageCompleted	0.29	keyPassesPerBallAction-Percentage	0.28
passesForwardPercentage-Completed	0.28	passesPercentageToBox	0.25
passesOwnHalfPercentage-Completed	0.27	dribblesPerBallActionPercentage	0.24
possessionLossPerBallAction-Percentage	-0.25	passesPercentageCrosses	0.22
foulsPerBallActionPercentage	-0.25	passesPercentageWide	-0.21
passesPercentageLong	-0.24	attackingDuelsPercentageWon	-0.21
shotsPercentageWithHead	-0.19	groundDuelsPercentageWon	-0.20

Table 42: Loadings of the third and fourth varimax rotated reduced variable for central midfielders that result from applying Reduced k-means with $k = 5$ and $p = 4$.

as how likely a central midfielder is to create chances for teammates to score goals, and how often he dribbles with the ball. Hence, we label the first varimax rotated reduced variable as "creating chances and dribbling". The second varimax rotated reduced variable can be interpreted as how involved a central midfielder is in the play of his team when they are in possession of the ball. The second reduced variable is therefore labeled as "involvement".

The third varimax rotated reduced variable can be interpreted as how simple the passes of a central midfielder are. Simple passes are passes that are typically short and either wide or backwards. We therefore label the third varimax rotated reduced variable as "simple passing". The fourth varimax rotated reduced variable can be interpreted as how much a central midfielder appears in the penalty area of the opposing team, and how likely he is to make goal attempts. Hence, we label the fourth varimax rotated reduced variable as "appearance in opponent box and likelihood of goal attempts".

The fifth varimax rotated reduced variable can be interpreted as how likely a central midfielder is to give good early crosses into the penalty area of the opposing team, and how much of his goal attempts are from outside the penalty area of the opposing team. As a result, we label the fifth varimax rotated reduced variable as "good early crossing and goal attempts mainly from outside the box".

Reduced k-means with $k = 6$ and $p = 5$ Central midfielders			
RV1 (Creating chances and dribbling)		RV2 (Involvement)	
Variable	Loading	Variable	Loading
dribblesPerBallActionPercentage	0.39	shareInBallActionsPercentage	0.35
shareInKeyPassesPercentage	0.32	shareInReceivedFirstPassIn-PossessionPercentage	0.34
keyPassesPerBallAction-Percentage	0.32	shareInPassesPercentage	0.34
passesPercentageToBox	0.28	shareInPossessionWithGoal-AttemptsPercentage	0.32
keyActionsPerBallAction-Percentage	0.25	shareInKeyActionsPercentage	0.31
passesPercentageFinalThird	0.25	passesPercentageLong	0.29
passesPercentageOpponentHalf	0.23	shareInPassFirstInPossession-Percentage	0.24
shareInKeyActionsPercentage	0.23	shareInPossessionWithGoals-Percentage	0.19
passesPercentageCrosses	0.22	keyActionsPerBallAction-Percentage	0.17
shotsPercentageWithHead	-0.19	crossPassesPercentageHigh	0.17

Table 43: Loadings of the first two varimax rotated reduced variables for central midfielders that result from applying Reduced k-means with $k = 6$ and $p = 5$.

Reduced k-means with $k = 6$ and $p = 5$ Central midfielders			
RV3 (Simple passing)		RV4 (Appearance in opponent box and likelihood of goal attempts)	
Variable	Loading	Variable	Loading
passesPercentageCompleted	0.37	actionsPercentageInOpponentBox	0.39
passesOpponentHalfPercentage-Completed	0.35	shotsPerBallActionPercentage	0.31
passesOwnHalfPercentage-Completed	0.35	goalAttemptsPerBallAction-Percentage	0.31
passesForwardPercentage-Completed	0.33	possessionLossPerBallAction-Percentage	0.30
passesFinalThirdPercentage-Completed	0.31	possessionRegainInPlayPerBall-ActionPercentage	-0.25
passesPercentageWide	0.26	shareInShotsPercentage	0.21
possessionLossPerBallAction-Percentage	-0.25	shareInGoalAttemptsPercentage	0.21
passesPercentageForward	-0.25	shareInPossessionWithGoals-Percentage	0.21
passesPercentageCrosses	-0.21	passesPercentageForward	-0.19
passFirstInPossessionPercentage-Forward	-0.18	passFirstInPossessionPercentage-Forward	-0.19

Table 44: Loadings of the third and fourth varimax rotated reduced variable for central midfielders that result from applying Reduced k-means with $k = 6$ and $p = 5$.

Reduced k-means with $k = 6$ and $p = 5$ Central midfielders	
RV5 (Good early crossing and goal attempts mainly from outside the box)	
Variable	Loading
crossPassesPercentageCompleted	0.34
shotsPercentageInsidePenaltyBox	-0.29
goalAttemptsPercentageInsidePenaltyBox	-0.29
attackingDuelsPercentageWon	0.28
crossPassesPercentageToGoalAttempt	0.28
crossPassesPercentageLate	-0.23
foulsPerBallActionPercentage	-0.21
groundDuelsPercentageWon	0.21
shotsPercentageWithHead	-0.19
shareInPassFirstInPossessionPercentage	0.18

Table 45: Loadings of the fifth varimax rotated reduced variable for central midfielders that result from applying Reduced k-means with $k = 6$ and $p = 5$.

E.2 Centre-forwards

In Section E.2.1 we show the resulting PCA loadings for centre-forwards, and in Section E.2.2 we show the resulting loadings of Reduced k-means for centre-forwards.

E.2.1 PCA

For centre-forwards, the loadings of the 5 varimax rotated principal components are displayed in Tables 46, 47 and 48. For each varimax rotated principal component, the loadings are shown for the 10 variables that have the highest absolute loading. The first varimax rotated principal component can be interpreted as how involved a centre-forward is in the play of his team when they are in possession of the ball, and how much he operates outside of the penalty area of the opposing team. Hence, we label the first varimax rotated principal component as "involvement and operating outside penalty box". The second varimax rotated principal component can be interpreted as how likely a centre-forward is to create chances for teammates to score goals and how offensive he plays, both in terms of where he positions himself on the pitch and how offensive his passing is. The second varimax rotated principal component is therefore labeled as "creating chances and offensive positioning/passing".

PCA Centre-forwards			
PC1 (Involvement and operating outside penalty box)		PC2 (Creating chances and offensive positioning/passing)	
Variable	Loading	Variable	Loading
shareInBallActionsPercentage	0.31	keyPassesPerBallAction-Percentage	0.36
shareInPassesPercentage	0.30	passesPercentageFinalThird	0.35
actionsPercentageInOpponentBox	-0.29	keyActionsPerBallAction-Percentage	0.34
shareInReceivedFirstPassIn-PossessionPercentage	0.28	passesPercentageOpponentHalf	0.32
goalAttemptsPercentageInside-PenaltyBox	-0.27	shareInKeyPassesPercentage	0.26
shotsPercentageInsidePenaltyBox	-0.26	passesPercentageToBox	0.26
shareInPossessionWithGoal-AttemptsPercentage	0.25	foulsPerBallActionPercentage	-0.23
passesPercentageLong	0.22	shareInKeyActionsPercentage	0.21
passesPercentageForward	0.19	dribblesPerBallActionPercentage	0.21
passesPercentageFinalThird	-0.18	passesPercentageCrosses	0.21

Table 46: Loadings of the first two varimax rotated principal components for centre-forwards.

The third varimax rotated principal component can be interpreted as how simple the passes of a centre-forward are. Simple passes are passes that are typically short and either wide or backwards. We therefore label the third component as "simple passing". The fourth varimax rotated principal component can be interpreted as how likely a centre-forward is to make a goal

attempt or to be involved in the possession moment leading to a goal attempt. As a result, we label the fourth varimax rotated principal component as "(involvement in) goal attempts".

PCA Centre-forwards			
PC3 (Simple passing)		PC4 ((Involvement in) goal attempts)	
Variable	Loading	Variable	Loading
passesPercentageCompleted	0.46	shareInGoalAttemptsPercentage	0.46
passesFinalThirdPercentage-Completed	0.44	shareInShotsPercentage	0.45
passesForwardPercentage-Completed	0.39	goalAttemptsPerBallAction-Percentage	0.41
passesPercentageCrosses	-0.30	shotsPerBallActionPercentage	0.39
passesPercentageToBox	-0.27	shareInPossessionWithGoal-AttemptsPercentage	0.19
possessionLossPerBallAction-Percentage	-0.25	shareInPossessionWithGoals-Percentage	0.18
passesPercentageForward	-0.19	actionsPercentageInOpponentBox	0.17
passesPercentageFinalThird	-0.17	possessionRegainInPlay-PercentageByInterception	0.17
passesPercentageLong	-0.13	possessionRegainInPlayPerBall-ActionPercentage	-0.15
passesPercentageOpponentHalf	-0.12	foulsPercentageOwnHalf	-0.12

Table 47: Loadings of the third and fourth varimax rotated principal component for centre-forwards.

The fifth varimax rotated principal component can be interpreted as how strong a centre-forward is in duels. Hence, we label the fifth component as "duel strength".

PCA Centre-forwards	
PC5 (Duel strength)	
Variable	Loading
duelsPercentageWon	0.51
attackingDuelsPercentageWon	0.46
groundDuelsPercentageWon	0.42
airDuelsPercentageWon	0.25
defensiveDuelsPercentageWon	0.23
possessionRegainInPlayPercentageBy-Interception	-0.20
dribblesPerBallActionPercentage	-0.19
foulsSufferedPerBallActionPercentage	0.18
crossPassesPercentageToGoalAttempt	0.16
shotsPercentageWithHead	0.13

Table 48: Loadings of the fifth varimax rotated principal component for centre-forwards.

E.2.2 Reduced k-means

The loadings of the three varimax rotated reduced variables that follow from applying Reduced k-means with $k = 4$ and $p = 3$ for centre-forwards are displayed in Tables 49 and 50. For each varimax rotated reduced variable, the loadings are shown for the 10 original variables that have the highest absolute loading. The first varimax rotated reduced variable can be interpreted as how much a centre-forward operates in the penalty area of the opposing team, how low his involvement is in the play of his team when they are in possession of the ball, and how likely he is to make goal attempts. Hence, we label the first reduced variable as "operating inside penalty box, low involvement and likelihood of goal attempts". The second varimax rotated reduced variable can be interpreted as how likely a centre-forward is to create chances for teammates to score goals. We therefore label the second varimax rotated reduced variable as "creating chances".

Reduced k-means with $k = 4$ and $p = 3$ Centre-forwards			
RV1 (Operating inside penalty box, low involvement and likelihood of goal attempts)		RV2 (Creating chances)	
Variable	Loading	Variable	Loading
actionsPercentageInOpponentBox	0.34	keyActionsPerBallAction-Percentage	0.35
shareInBallActionsPercentage	-0.32	keyPassesPerBallAction-Percentage	0.35
shareInPassesPercentage	-0.30	possessionLossPerBallAction-Percentage	-0.32
shotsPerBallActionPercentage	0.28	foulsPerBallActionPercentage	-0.27
shareInReceivedFirstPassIn-PossessionPercentage	-0.28	shareInKeyPassesPercentage	0.22
goalAttemptsPerBallAction-Percentage	0.26	shareInKeyActionsPercentage	0.22
goalAttemptsPercentageInside-PenaltyBox	0.22	passesForwardPercentage-Completed	0.21
keyPassesPerBallAction-Percentage	0.22	passesPercentageCompleted	0.21
shotsPercentageInsidePenaltyBox	0.21	passesFinalThirdPercentage-Completed	0.19
shareInPossessionWithGoal-AttemptsPercentage	-0.17	crossPassesPercentageToGoal-Attempt	0.18

Table 49: Loadings of the first two varimax rotated reduced variables for centre-forwards that result from applying Reduced k-means with $k = 4$ and $p = 3$.

The third varimax rotated reduced variable can be interpreted as how simple the passes of a centre-forward are, and how much of his passes are outside of the final third. As a result, we label the third varimax rotated reduced variable as "simple passing outside final third".

Reduced k-means with $k = 4$ and $p = 3$ Centre-forwards	
RV3 (Simple passing outside final third)	
Variable	Loading
passesPercentageCompleted	0.37
passesFinalThirdPercentageCompleted	0.36
passesPercentageToBox	-0.35
passesPercentageCrosses	-0.29
passesForwardPercentageCompleted	0.29
passesPercentageFinalThird	-0.25
passesPercentageOpponentHalf	-0.24
passesPercentageForward	-0.19
shareInShotsPercentage	-0.16
airDuelsPercentageWon	-0.15

Table 50: Loadings of the third varimax rotated reduced variable for centre-forwards that result from applying Reduced k-means with $k = 4$ and $p = 3$.