

Valuing passes in football using ball event data

Master Thesis Econometrics and Management Science
Lotte Bransen, 418975

Supervised by
Michel van de Velden, Erasmus University Rotterdam
Jan Van Haaren, SciSports

October 12, 2017

Abstract

This master thesis introduces and evaluates several models for valuing passes in football. We use event data from five seasons of matches from the top five leagues in Europe. The most simplistic model considers the value of possessing the ball in a certain area of the pitch. The other models use clustering methods to find similar passes and similar attacks to value passes.

The comparison of attacks also yields the opportunity to find teams with similar playing styles. The proposed pass valuing models make it possible to rank players based on the values that were assigned to their passes. This ranking of players may help clubs to scout potential new players as well as to analyze the upcoming opponent. We show that the pass values can be used to estimate player market values and to predict match outcomes.

Acknowledgments

Foremost, I would like to thank my university supervisor dr. Michel van de Velden for supervising my master thesis, providing detailed feedback and advice. I would also like to thank my co-reader prof. dr. Philip Hans Franses for reading my thesis. Special thanks go to Jan Van Haaren who supported me throughout my thesis and provided me with his detailed feedback. I would also like to thank Lars van Hove for helping me with the visualizations presented in this thesis and for reading parts of this thesis. I would also like to express my gratitude to all my colleagues from SciSports for their help and support. I also like to thank my father, Jan Bransen, for providing feedback on the text.

Contents

Contents

1	Introduction	1
1.1	Research context	1
1.2	Problem description	1
1.3	Research questions	2
1.4	Literature review	2
1.5	Contributions	4
1.6	Thesis structure	5
2	Data description	6
2.1	Data sets	6
2.1.1	Match event data set	6
2.1.2	Match sheet data set	7
2.1.3	Player data set	7
2.1.4	Club ELO ratings data set	7
2.2	Linking the data sets	8
2.3	Division of data set in training, validation and test set	8
2.4	Limitations of the available data	9
3	Football analytics	10
3.1	Pitch partitioning	10
3.2	Possession sequences	11
3.3	Expected goals model	12
4	Machine learning	16
4.1	Similarity measures	16
4.1.1	Dynamic time warping	16
4.1.2	Fréchet distance measure	17
4.1.3	Longest common subsequence distance measure	18
4.1.4	Comparison of the distance measures	19
4.2	K-nearest neighbors algorithm	19
5	Approaches	20
5.1	Zone-oriented pass value (ZPV)	20
5.1.1	Algorithm	20
5.1.2	Step 1: Partitioning of the pitch	21
5.1.3	Step 2: Valuing the zones using the training data	21
5.1.4	Step 3: Determining the pass value	22
5.2	Pass-oriented pass value (PPV)	24
5.2.1	Algorithm	24
5.2.2	Step 1: Defining and evaluating the distance measure	25
5.2.3	Step 2: Determining the outcome of the possession sequences	27
5.2.4	Step 3: Pre-clustering the passes	27
5.2.5	Step 4: Determining the pass value	29
5.3	Sequence-oriented pass value (SPV)	30

5.3.1	Algorithm	31
5.3.2	Step 1: Determine which distance measures to use	32
5.3.3	Step 2: Break up the sequences into subsequences	34
5.3.4	Step 3: Pre-cluster the possession (sub)sequences	34
5.3.5	Step 4: Calculate the distances and find the k nearest neighbors	35
5.3.6	Step 5: Assigning values to the passes	35
5.3.7	How does DTW compare to the Fréchet distance?	35
5.3.8	Application: identifying teams' tactics	35
5.4	Comparison of the approaches	38
6	Experimental evaluation	41
6.1	Player ECOM metric	41
6.2	Evaluation criteria	41
6.2.1	Correlation FIFA 16 passing and vision skills	41
6.2.2	Estimating market values	42
6.2.3	Predicting football match outcomes	42
6.3	Evaluation results	43
6.3.1	Correlation FIFA 16 passing and vision skills	43
6.3.2	Estimating market values	44
6.3.3	Predicting football match outcomes	46
6.4	Conclusion evaluation	47
7	Results PPV approach	48
7.1	Q1: What do the high-valued passes look like?	48
7.2	Q2: Who are the top-ranked players based on their ECOM in the 2016/2017 season?	49
7.3	Q3: What are the top-ranked teams based on their players' ECOM values in the 2016/2017 season?	52
8	Conclusion	54
8.1	Summary	54
8.2	Answering the research questions	54
8.3	Discussion	55
8.4	Future work	56
A	Results ZPV - 15 seconds rule approach	58
B	Results ZPV - expected goals approach	59
C	Results PPV approach	61
D	Results SPV - DTW approach	62
E	Results SPV - Fréchet approach	64
F	Top 5 teams	65
G	High-valued passes	66
	Bibliography	67

1 Introduction

This chapter provides the main motivation and goals for this research, as well as the contributions to the field of football analytics.

1.1 Research context

In recent years, the use of data has strongly increased in the sports sector. In this context, football is one of the most popular and exciting sports and recently has been considered as a research area in mathematics and computer science (Rahnamai Barghi (2015)). Nowadays, vast amounts of data are generated during and after a game of football. This gives data intelligence companies potentially the opportunity to measure the qualities of football players.

Passing is the most common type of event that occurs during a football match (Power et al. (2017)), which makes it an interesting event to investigate. However, so far there has only been done little research on passes and the passing skills of football players. The passing skills of players are difficult to assess using only ball event data. The available data do not give an indication of how good a pass is and only denote the location of the pass and whether it arrived successfully at its destination.

Most current approaches to assess players' passing skills either focus on the completion rate of the players' passes or on the players' number of assists or keypasses. This definitely leaves room for improvement regarding measuring the passing skills of players. For example, passing the ball from one defender to another is considered relatively easy, whereas passing the ball from central midfield to a striker in the box requires more advanced passing skills and has a higher probability of failing. When considering a player's completion rate those two passes get rated equally and when both passes are intercepted by the opponent the penalty value is equal for losing possession. However, intuitively these two passes should not get the same value when succeeding and the penalty cost should be lower for the more risky pass.

Assists (i.e. passes that directly lead to a goal being scored) and keypasses (i.e. passes that directly lead to a goal attempt) are rare events in football. Therefore, when only considering a player's number of assists or keypasses to describe the player's passing skills a lot of information is lost.

This thesis uses machine learning techniques to create a new metric to value passes in football.

1.2 Problem description

The goal of this thesis is to build a model that assigns a value to each pass, which can be used to rank players on their passing quality. We use event data of almost 8000 matches in the five top leagues in Europe - the English Premier League, the Spanish Primera División, the Italian Serie A, the French Ligue 1 and the German Bundesliga - to build a model for valuing passes. Passes are clustered using different criteria and similar passes are valued equally. The resulting model values passes on their expected contribution to the outcome of the match (ECOM). This model makes it possible to rank players based on their contribution to the outcome of the

match. The model may help football clubs when scouting new players for their club as well as when analyzing the opponent in the upcoming match, by knowing which of the opponent's players is most important in the opponent's play.

This thesis focuses on the influence a pass has on the scoring opportunities of a team. The difficulty and precision of the passes is not taken into account due to a lack of accurate tracking data on the positions of all players on the pitch. In general, for passes that reach a teammate it is complicated to determine whether a pass is good or bad. It can be said that a pass that leads to a goal (i.e. an assist or pre-assist) is a good pass. However, if a team's strikers do not finish excellent scoring chances resulting from a pass, such a pass is not rewarded the way it should be. Therefore, passes are valued on their ECOM.

Scoring is a rare event in football, therefore it is a challenge to measure the influence of passes on the outcome of the game. We use an expected goals model to measure the value of goal attempts. This model determines the probability that a goal attempt turns into a goal and is used to value the expected outcomes of attacks.

1.3 Research questions

This thesis addresses four strongly related research questions. The main research question is: "To what extent are machine learning techniques capable of determining which players in Europe's top five leagues perform the passes that contribute most to the outcomes of their teams' matches?". To address this question, we first assign a value to all individual passes based on historical match data and then aggregate these pass values for each player to obtain a player ranking. We explore three different approaches to assigning values to individual passes. These approaches require comparing possession sequences in terms of their spatial similarities, which can be done in several ways. Hence, the second research question is: "How can we best compare possession sequences and compute their spatial similarities?".

A model for valuing passes and a similarity measure for possession sequences enables several applications. We explore the task of distinguishing between playing styles of players, teams and leagues, which is useful for the tactical analysis of matches. Hence, the third research question is: "Can we use our pass valuation model to detect playing styles. And, if so, which teams in Europe's top five leagues have similar playing styles?". Furthermore, this thesis explores the task of estimating the market values of players, which is relevant to the player recruitment process. Hence, the fourth research question is: "To what extent can the players' passing values be used to estimate their market values?".

1.4 Literature review

In football two different types of data are generated: event data that captures all events that happened on the pitch and tracking data that tracks all players and the ball during the match. Event data denotes the player, its x,y-coordinates, the time of the event in the match and the type of interaction with the ball. When a

passing event occurs, the data only entails information about the position of the player passing the ball and the position of the destination of the pass.

Decroos et al. (2017) use historical event data to automatically value actions in football. Their paper has a similar goal as my thesis, i.e. valuing actions in football. However, the way actions are valued is different from my approach. Decroos et al. (2017) cluster action sequences are clustered using dynamic time warping in combination with the k-nearest neighbor algorithm. Action sequences are sequences of consecutive actions performed by the same team without losing possession. These sequences are valued by taking the average of the values of the 100 nearest (using dynamic time warping as a distance measure) action sequences. Each pass is assigned a fraction of this sequence value by taking into account the relevance of this pass to the sequence.

Another work that focuses on assigning values to passes in football is the paper by Gyarmati & Stanojevic (2016). They assign a value to possessing the ball in a particular part of the pitch, which is team-specific and based on historical data. The pitch is partitioned taking into account the start and end location of all passes of a team. When each cluster is given a value, the value of a pass can be calculated on the basis of the value of the part of the pitch the pass ends and the value of the part of the pitch from where the pass was given. When the ball is lost the pass gets a negative value. The authors conclude that it is sometimes useful to lose the ball. For defenders this may be clearances, whereas for attackers this may be very risky passes leading to a loss of possession. Incorporating the tactics of the teams into the model by determining the possession values for each team separately is a very interesting approach.

Mackay (2016) uses football event data to build a model to measure players' passing skills. An expected goals value, which estimates the probability that a goal attempt turns into a goal being scored, is assigned to the start and end location of each pass. For the start location of a pass the expected goals model only takes into account the location (distance to the goal and angle to the goal) of the pass. A different expected goals model is used for the end location of the pass as the type of pass (e.g. cross, through ball, etc) is included. The pass is valued as the added expected goals value, so the end value minus the start value. Mackay concludes that Mesut Özil, Alexis Sánchez and Cesc Fàbregas have the best passing skills of the English Premier League in the 2015/2016 season.

Horton et al. (2014) introduce a model to classify passes in football by using spatiotemporal data. The dominant region of a player (first introduced by Taki & Hasegawa (2000)), which is defined as the region on the pitch that a player can reach before any other player can, plays an important role in their model. They use the dominant region of a player to incorporate the pressure the player in possession of the ball experiences.

Gyarmati & Anguera (2015) introduce an algorithm based on dynamic time warping which clusters pass strategies of football teams. They use only event data and aim to find differences in passing strategies between teams. Cintia, Giannotti, Pappalardo, Pedreschi & Malvaldi (2015) introduce indicators for the passing performances of

players and teams. Their model predicts quite well which teams will perform best in the league. Cintia, Rinzivillo & Pappalardo (2015) model the passing game of a team as a network, in which the nodes represent the players and the edges represent the passes between the players. This passing network is then used to measure the performance of the team using three network measures. These network measures are: the passing volume, the passing heterogeneity and the harmonic mean of the passing volume and heterogeneity.

Comparable problems are investigated in other sports, such as ice hockey, basketball and American football. Schulte et al. (2015) learn action-value functions that quantify the impact of actions on scoring a goal in ice hockey. In their paper the game of ice hockey is seen as a Markov game, where each performed action in the game is seen as a Markov state. The state transition probabilities can be calculated using historical data and goals are the only absorbing states. Each random walk through this state space leads to either a home team goal or an away team goal. This makes it possible to calculate the expected reward for each state in the state space. The expected reward for each state can be computed by using a dynamic programming algorithm. This is a very interesting approach of valuing actions that could also be used in football. However, the ice hockey game is a more continuous game than football as the ball cannot go out of play and goals are scored more frequently. Furthermore, the lack of information on the position of the other players may be a bigger problem in football than it is in ice hockey.

Some interesting papers are written about action-valuing in basketball. Cervone et al. (2014) introduce the Expected Possession Value, which assigns a value to each state of the game and uses all options the player in possession of the ball has. Player tracking data are used to evaluate the decisions players make and the action is valued by the increase of the Expected Possession Value after the action. It is a very interesting approach, however in football player tracking data are scarce and not very accurate yet. Moreover, football teams consist of eleven players, which makes it a time-consuming task to evaluate all options in each state of the game. However, a simplified model could be built for football as a start.

Gudmundsson & Wolle (2014) use the Fréchet distance measure and Longest Common Subsequence distance measure to compare the movement of the players on the pitch. In contrast with this thesis, Gudmundsson & Wolle (2014) dispose of tracking data of the positions of the players.

1.5 Contributions

This thesis introduces new approaches to value passes in football. This research project is, in the context of passing skills analysis in football, amongst the first to use clustering methods to value passes in football. We introduce a new distance measure for passes that makes it possible to define the similarity of passes. Furthermore, similar to Decroos et al. (2017), we use dynamic time warping to compare sequences of passes. In addition, we also use the Fréchet distance measure and Longest Common Subsequence distance measure to compare sequences of passes.

This has never been done by using only football event data. Although Decroos et al. (2017) also compare passing sequences to value individual passes, this thesis introduces a new way of valuing the individual passes these sequences consist of.

1.6 Thesis structure

This thesis is divided into eight chapters. In Chapter 1, we introduce the topic of research and the research questions. In Chapter 2, we describe the data used for this research. In Chapter 3, we describe known background in football analytics and in Chapter 4 we describe known background in machine learning. In Chapter 5, we introduce three new approaches to value passes. In Chapter 6, we evaluate the three approaches on pre-determined evaluation criteria. In Chapter 7, we present the results of the preferred model. In Chapter 8, we discuss the conclusions of the research and future work.

2 Data description

In this chapter, we explain which data sets are used for this research and how we combine these data sets. In addition, we explain how the data is split into a training data set, a validation training set and a test data set. Lastly, we discuss the limitations of the available data.

2.1 Data sets

We use five data sets in this research. The most important one is an event data set which denotes all events that occur in a match. Together with this event data set comes a match data set which includes match and player features of the match. The third data set includes the predicted transfer values for players and the fourth data set includes FIFA 16 skills for players. The fifth data set includes the Football Club ELO ratings of all clubs in the top five leagues in Europe as of August 2017.

2.1.1 Match event data set

This research uses an event data set consisting of more than sixteen million events (e.g. passes, shots, interceptions, goalkeeper saves, etc) of 9113 matches. These matches took place in the seasons 2012-2013, 2013-2014, 2014-2015, 2015-2016 and 2016-2017 in the five top leagues of Europe, the English premier League, the Spanish Primera División, the Italian Serie A, the French Ligue 1 and the German Bundesliga. Only 17 matches of those five seasons are missing in the data set. Table 1 shows the most important variables in the event data set.

Variable	Explanation
Event ID	Unique ID for the event
Match ID	Unique ID for the match
Team ID	Unique ID for the team executing the event
Player ID	Unique ID for the player executing the event
Type ID	The type of event, e.g. pass, shot, interception, etc
Time	The time in minutes and seconds of the start of the event
(x,y)	The position on the pitch where the event started
Outcome	A binary variable for the outcome of the event, e.g. succeeded or not

Table 1: The most important variables of the event data set

For each of the different type ID's some extra information is given. For a pass the most important extra information for this thesis is its end position. It is also denoted whether a pass is a keypass, an assist or neither. A keypass is a pass that is directly followed by a goal attempt. An assist is a pass that is directly followed by a goal, so an assist is also a keypass. For goal attempts the following aspects are denoted: the body part used, the type of play (e.g. regular play, corner kick, etc), whether the goal attempt results in a goal and if not whether the goalkeeper stops the ball, the ball hits the woodwork or the goal attempt is off target. These data are collected manually during the matches and checked manually shortly after the match. As all events are entered by people and people make mistakes, it is very likely that the

data set is not perfect. However, as the data set is very large (over sixteen million events) we assume that the effect of possible errors in the data is negligible.

2.1.1.1 Location of the passes

Not all pitches in the football leagues have the exact same size and that is why the x and y variables in the data are normalized to a scale from zero to 100. However, as some of the pass valuing approaches used in this thesis require the Euclidean distance between two points on the pitch, we rescale these x and y variables to the standard pitch size. The standard size of a football pitch in international matches is 68 meters wide and 105 meters long. There may be some differentiation among the pitches in all stadiums, but in this thesis we use the standard pitch size.

2.1.2 Match sheet data set

For each match the event data set is coupled to a match sheet data set which includes the line-up, the formation (e.g. 4-4-2, 4-3-3, etc), the substitutions and the final score of the match. The position of each individual player is calculated based on the average position from which the player performs his actions. The combination of the formation, line-up and substitutions is used to extract for each player the number of minutes he has played on which position. We compute for each player on which positions he has played and how many minutes he has played on each of these positions. This makes it possible to compare the passing skills from players that play on the same position and to find out on which position a player plays best concerning his passes.

2.1.3 Player data set

In this thesis we build a model to predict transfer values of players taking into account their passing skills generated by the proposed models. Transfermarkt¹ is the main site concerning transfer values of players. Data from this site are used to retrieve the estimated transfer values of the players that played at least 450 minutes in the 2016-2017 season. Transfermarkt predicts the transfer values by using the actual transfer fees paid for players, some performance features and personal assessments. As it is partly a community-driven website, transfer values are often adjusted when many people think they should be adjusted.

In the results chapter, we compare the players' passing skills generated by the proposed models to the players' passing and vision skills in the FIFA 16 videogame. We use a data set² of the passing and vision skills of all players in FIFA 16. This data contains the skills that FIFA assigned to the players on May 26th 2016.

2.1.4 Club ELO ratings data set

The club ELO ranking is a ranking of all football clubs in the world. A club's ELO rating is updated after each match and takes into account the result of the match

¹transfermarkt.de

²from sofifa.com

and the strength of the opponent. The ELO rating of the clubs are linked to the players, which is used as a measure of the strength of the team the player is playing for. This variable is used in the estimation of the players' market values and turns out to be an important one when estimating market values of football players.

2.2 Linking the data sets

We link the players in the event data set to their market value and their FIFA 16 skills using the Levenshtein distance between the names in the data sets. The Levenshtein distance is a distance measure to measure the distance between two strings. It counts the minimal number of adjustments (e.g. delete a letter, insert a letter or replace a letter by another one) to make a string from another string. We check each linking of a player between the data sets by checking whether the date of birth is equal in the data sets. When the birth date is not the same in the data sets or when there is no close match for a certain player, we do the linking manually. About 5% of the players is linked manually.

We link the club ELO ratings to the players as we use these ratings for one of the applications of this research project. To link the clubs of both data sets, we again use the Levenshtein distance.

2.3 Division of data set in training, validation and test set

We divide the event data set of all leagues into a training data set, a validation data set and a test data set. We train the proposed models on the data of all leagues of the seasons 2012/2013, 2013/2014 and 2014/2015. We use the data of the 2015/2016 season to test these models and to compare the results of the models. Based on pre-determined evaluation criteria we choose the best model. We use the data of the four seasons 2012/2013, 2013/2014, 2014/2015 and 2015/2016 to train this best model and we present the results on the 2016/2017 season. We choose to split the data in this way as there is only data available on the market values of the players from August 2017. As we want to check whether the pass values correlate with the market values it is best to use the pass values of the most recent season. Figure 1 visualizes the splitting of the data sets and Table 2 shows the number of passes, goal attempts and sequences in the three data sets.

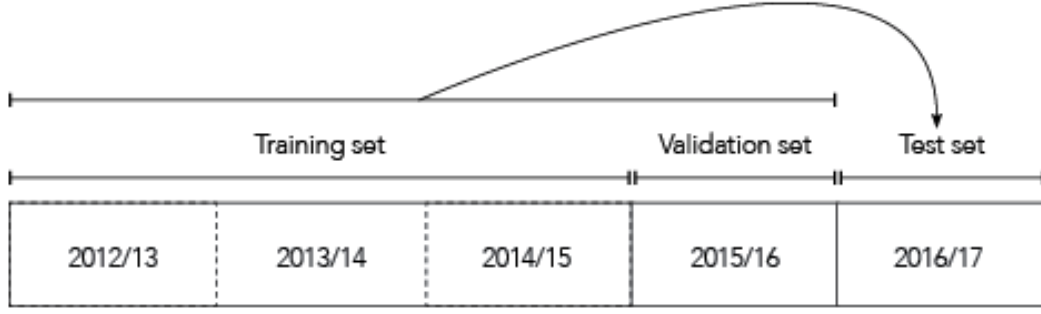


Figure 1: The data are split into a training, validation and test set. When the best model is chosen based on the results on the validation set, both the training and validation set are used as an input for valuing the passes of the test set.

	Training data set (seasons 2012-2015)	Validation data set (season 2015/2016)	Test data set (season 2016/2017)
Matches	5,462	1,826	1,825
Events	9,428,347	3,229,972	3,228,683
Passes	5,110,918	1,771,736	1,788,267
Successful passes	3,906,151	1,354,429	1,381,978
Goal attempts	138,446	45,453	45,925
Goals	14,562	4,870	5,173

Table 2: The number of matches, passes, successful passes, goal attempts and goals in the three data sets

2.4 Limitations of the available data

Two important features missing in the data are the positions of the other players on the pitch at the time of the event and the time at which a pass reaches its destination. The exact speed of a pass cannot be determined. However, the time between two passes can be calculated, so this lack of information is not a big limitation. A player dribbling with the ball across the pitch is not denoted as an event. When the end and start location of two sequential passes are not equal, it is most likely that the player receiving the first pass and executing the second pass dribbles this distance. The exact path is not known, and therefore we see these dribbles as actions in the possession sequences with a straight line between the two positions.

3 Football analytics

This chapter describes anything related to football analytics the reader should know to understand the remainder of the thesis. We explain how the pitch is partitioned into different zones and we define possession sequences. Furthermore, we explain what an expected goals model is and which expected goals model we use in this research.

3.1 Pitch partitioning

In the data description, we explained how the normalized (x, y) -values of the actions are turned into realistic (x, y) -coordinates on a standard-sized pitch. In multiple stages of this research project a partitioning of the pitch is required. The pitch is partitioned into a number of equal-sized zones. In this research project only partitionings of equal-sized zones are considered as this makes the results easier to interpret and work with. As the pitch has a length of 105 meters and a width of 68 meters, the zones have a length with a denominator of 105 and a width with a denominator of 68. Let x_{step} be the length of the zones and let y_{step} be the width of the zones. Then the total number of zones is equal to $n = \frac{105}{x_{step}} \cdot \frac{68}{y_{step}}$. Figure 2 shows how the zones are numbered.

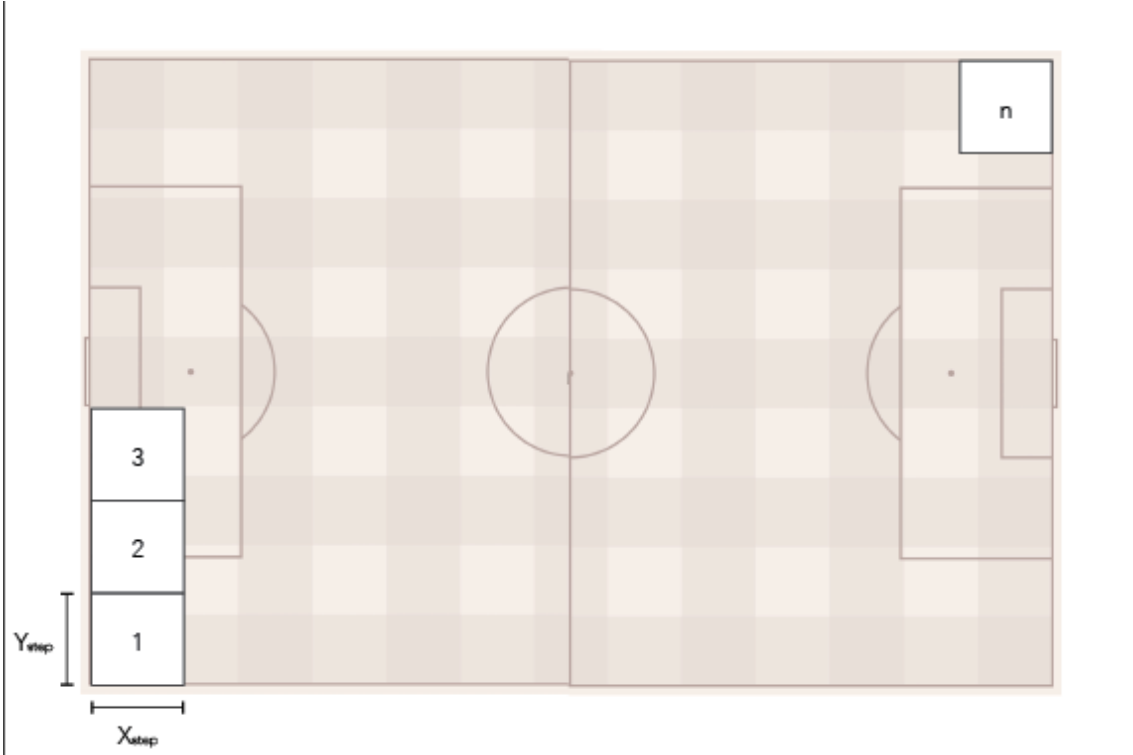


Figure 2: The pitch is partitioned into n equal-sized zones, numbered from bottom to top, from left to right

For each action, with coordinates (x, y) , we determine the corresponding zone number i as follows:

$$x_{adj} = \lfloor \frac{x}{x_{step}} \rfloor \cdot x_{step}$$

$$y_{adj} = \lfloor \frac{y}{y_{step}} \rfloor \cdot y_{step}$$

$$i = \frac{68}{y_{step}} \cdot \frac{x_{adj}}{x_{step}} + \frac{y_{adj}}{y_{step}}$$

This approach increases the speed of determining the corresponding zone for a very large number of points on the pitch.

3.2 Possession sequences

A football match can be seen as a sequence of consecutive actions such as passes, shots, the ball going out of play, throw-ins, fouls, free kicks, et cetera. The action sequence of match i is denoted as $A_i = [a_1, \dots, a_{N_i}]$ with N_i being the total number of actions in match i . This action sequence of a match can now be partitioned into possession sequences. Possession sequences start and end with one of the following events:

- The start or end of a period of the match (first half, second half, overtime)
- The ball goes out of play
- The opposing team touches the ball (one touch is sufficient)
- A goal has been scored

The match i can thus be denoted as a sequence of possession sequences S_j , $j = 1, \dots, M_i$, $A_i = [S_1, \dots, S_{M_i}]$, where M_i is the number of possession sequences of match i . Furthermore, the possession sequence S_j is denoted by $[a_{k_j+1}, \dots, a_{k_j+l_j}]$, where l_j is the number of actions in possession sequence S_j and $k_j = \sum_{m=1}^{j-1} l_m$.

We assign all these possession sequences a value u_j which represents the outcome of the sequence. The outcome of a possession sequence is one of the following:

- Possession is lost
- A period of the match has ended
- A goal attempt is being made (either a goal or not)
- A foul is committed and a free kick is awarded
- The ball goes out of play via the opponent and the team that had possession of the ball is awarded a corner kick or a throw-in

Figure 3 shows such a possession sequence which started with a throw-in and ended in a goal being scored. This possession sequence corresponds to the 2-0 scored by Alexis Sanchez in the match between Barcelona and Osasuna in the 2013/2014 season. The black solid lines represent passes, the black dotted lines represent dribbles and the red solid line is the shot that finds its way into goal. The exact path

of the players dribbling the ball is not known, and therefore dribbles are represented by a straight line.

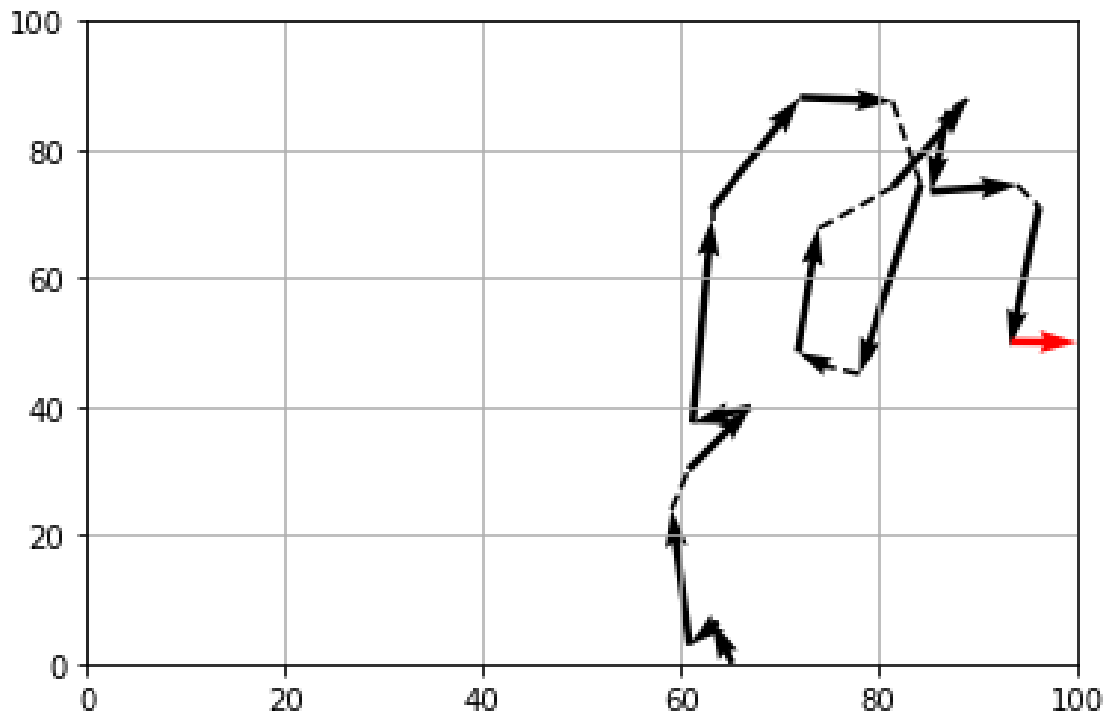


Figure 3: Possession sequence starting with a throw-in and resulting in a goal being scored

When possession gets lost a sequence we assign a value of zero to the possession sequence. The sequences that end because a period of the match has ended are not taken into account. In addition, when the ball goes out of play, or a foul is committed, the outcome of this sequence is also zero as the sequence ended and a new sequence starts. When the outcome of a sequence is a free kick, corner or a throw-in the value of the sequence is thus zero as well.

The most valuable outcome of a possession sequence is a goal attempt. We value goal attempts using the Expected Goals metric which is a frequently used metric in the football analytics world and was introduced by Caley (2013). This metric takes into account different parameters in order to calculate the expectation that a certain scoring attempt results into a goal being scored. We explain the expected goals model in more detail in the following section.

3.3 Expected goals model

Two of the most commonly used parameters for the Expected Goals model are the distance from the goal and the angle to the goal from where the shot is taken. It turns out that those two parameters have a huge influence on the probability that a goal attempt from a certain position becomes a goal. However, SciSports (2016) added some other parameters to the model, such as the match situation (open play,

corner, free kick or penalty), whether the goal attempt is a rebound and what kind of rebound, the attempt type (header, shot from dribble or shot from pass) and the difference in goals between the two teams at the time of the goal attempt. Soccermetrics (2017) also takes into account the match time, the league and the body part used to execute the goal attempt when determining the expected goal value of a goal attempt.

In pena.lt/y (2015) a model using support vector machines is introduced. Unfortunately, it is not explained in depth how this model is estimated exactly.

To create a model for assigning values to passes in football, first an expected goals model needs to be estimated in such a way that all scoring opportunities in the given data can be valued. The passes that led to these scoring possibilities can then be given a value. When only the goals are taken into account important information is lost as players who give perfect passes, but have failing team mates, do not get the credit they deserve. Also the opposite holds, when a player gives a very simple pass and his teammate scores from an almost impossible position, the player assisting the goal gets too much credit when you only take goals into account.

The core of this thesis is to value passes and not to value goal attempts based on an expected goals model and therefore we use a relatively simple expected goals model in this thesis. The expected goals model just includes the location of the goal attempt and uses data from previous seasons to determine the probability that a goal attempt from a certain position finds the net.

Figure 4 shows a heat map with for each 1 by 1 meter zone the probability of a goal being scored when a goal attempt is executed from that zone. We use the complete data set containing all goal attempts of the three seasons (2012-2015) of the top five leagues of Europe to calculate these probabilities. As expected, the closer the ball is to the goal, the higher the probability of a goal being scored. The values are simply calculated by counting the number of goal attempts from each zone and the percentage of those goal attempts that find their way into the goal is the probability that is represented in the figure.

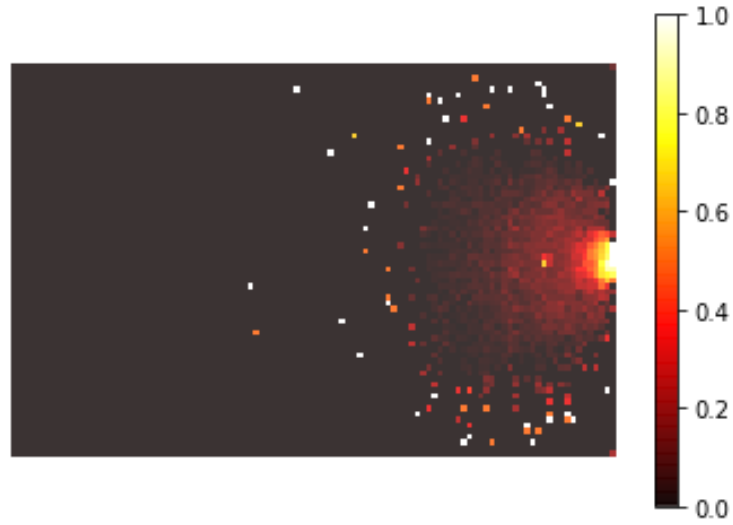


Figure 4: Heat map showing for each square the probability of scoring a goal when shooting from that square. The direction of attack is from left to right.

In Figure 4 it can be observed that there are some outliers, as for some positions far away from goal the probability of scoring a goal is equal to one which is highly unlikely. The zones are assigned those values as one lucky shot was taken from that zone and this was the only shot ever taken from that zone. Therefore all zones from which 10 or less shots were taken in all 20 seasons were ignored and those zone are assigned a probability of zero of scoring a goal. This can be supported by the fact that it can be assumed that players do rarely shoot from this range as they know the probability of scoring is very low.

It can also be observed that the penalty spot stands out in this figure and as we do not want to include the scoring possibility of penalty's in the model, all 1088 penalty's were removed from this model. These changes result in the expected goals values as shown in Figure 5. In this heat map the penalty spot sill stands out, this is probably caused by the fact that the people who denote the data are biased towards the lines on the pitch. This is the expected goals model that we use in this research to assign values to goal attempts.

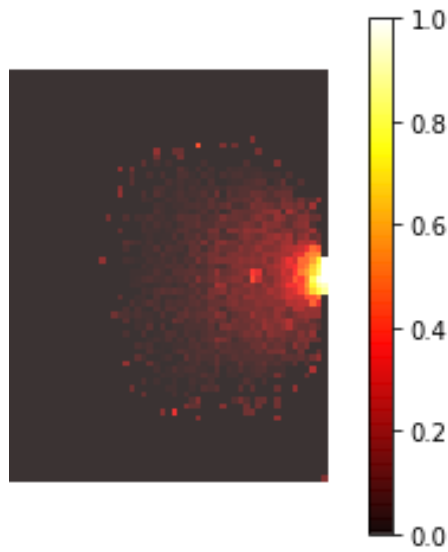


Figure 5: Heat map showing for each square the probability of scoring a goal when shooting from that square. The direction of attack is from left to right and only the right half of the pitch is shown as the left half of the pitch solely contains of black (zero probability) zones. All penalty's are removed and squares with 10 shots or less are given a probability of zero. All goal attempts from the 2012/2013, 2013/2014 and 2014/2015 seasons are used.

4 Machine learning

This chapter presents three different similarity measures to measure the similarity of possession sequences. Next to that, we describe the k-nearest neighbors algorithm.

4.1 Similarity measures

One of the proposed methods of this thesis requires a measurement of similarity between possession sequences in football. We use three distance measures to measure the 'distance' between possession sequences: the dynamic time warping distance measure, the Fréchet distance measure and the longest common subsequence distance measure. We explain each of these measures in the following paragraphs.

4.1.1 Dynamic time warping

Dynamic time warping is a commonly used technique in time series analysis (Müller (2007)). Dynamic time warping makes it possible to measure the distance between two sequences that vary in speed and length. The dynamic time warping distance measure requires an optimal coupling, the optimal warping path, between two sequences. The dynamic time warping distance measure is the sum of the Euclidean distances between the coupled points.

To explain the distance measure in more depth, let's suppose we have two sequences $S_1 = [a_1, \dots, a_{l_1}]$ and $S_2 = [b_1, \dots, b_{l_2}]$. Let $C \in \mathbb{R}^{l_1 \times l_2}$ be the cost matrix, where $C(a_n, b_m) := \sqrt{(a_n - b_m)^2}$, $n = 1, \dots, l_1$ and $m = 1, \dots, l_2$. The cost matrix thus represents the Euclidean distances between all combinations of points of both sequences. A so-called warping path $W = [(w_1^1, w_1^2), \dots, (w_L^1, w_L^2)]$ is a path through this cost matrix for which the following three conditions hold:

- Boundary condition: $(w_1^1, w_1^2) = (1, 1)$ and $(w_L^1, w_L^2) = (l_1, l_2)$
- Monotonic condition: $w_1^1 \leq w_2^1 \leq \dots \leq w_L^1$ and $w_1^2 \leq w_2^2 \leq \dots \leq w_L^2$
- Continuity condition: $w_{l+1} - w_l \in \{(1, 0), (1, 1), (0, 1)\}$, $l = 1, \dots, L - 1$

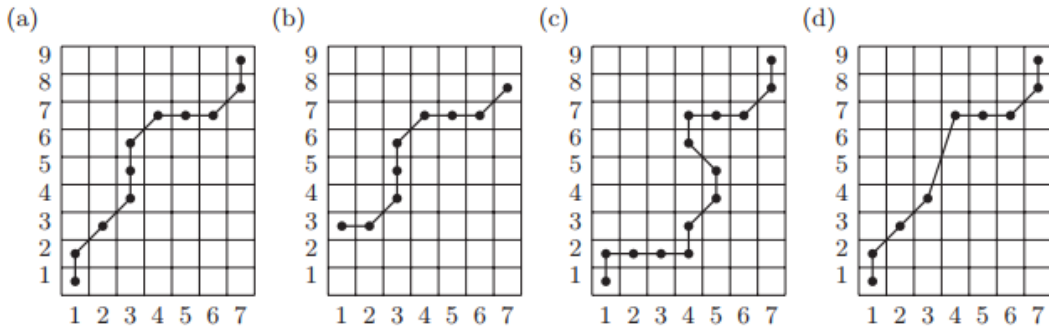


Figure 6: Illustration from Müller (2007) showing four paths of index pairs for a sequence of length seven and a sequence of length nine. (a) is a warping path satisfying all conditions. (b) does not satisfy the boundary condition, (c) does not satisfy the monotonic condition and (d) does not satisfy the continuity condition.

The optimal warping path is the path through the cost matrix that minimizes the total costs. The DTW distance is now defined as the costs of the optimal warping path:

$$DTW(S_1, S_2) = \min\left\{\sum_{l=1}^L \sqrt{(a_{w_l^1} - b_{w_l^2})^2} \mid W \text{ is a warping path}\right\}$$

The optimal warping path can be found using dynamic programming. Let $D(n, m) := DTW(S_1(1 : n), S_2(1 : m))$, where $S_1(1 : n) = (a_1, \dots, a_n)$ for $n = 1, \dots, l_1$ and $S_2(1 : m) = (b_1, \dots, b_m)$ for $m = 1, \dots, l_2$. Then the warping path can be found using the following recursive equation:

$$D(n, m) = c(a_n, b_m) + \min\{D(n-1, m-1), D(n-1, m), D(n, m-1)\}$$

for $1 < n \leq l_1$ and $1 < m \leq l_2$. It follows that $DTW(S_1, S_2) = D(l_1, l_2)$ and thus when all paths have to be considered this leads to a complexity of $O(l_1 l_2)$ for each pair of sequences.

Figure 7 illustrates the dynamic time warping coupling as compared to the Euclidean coupling. The black lines represent the matching between the two sequences.

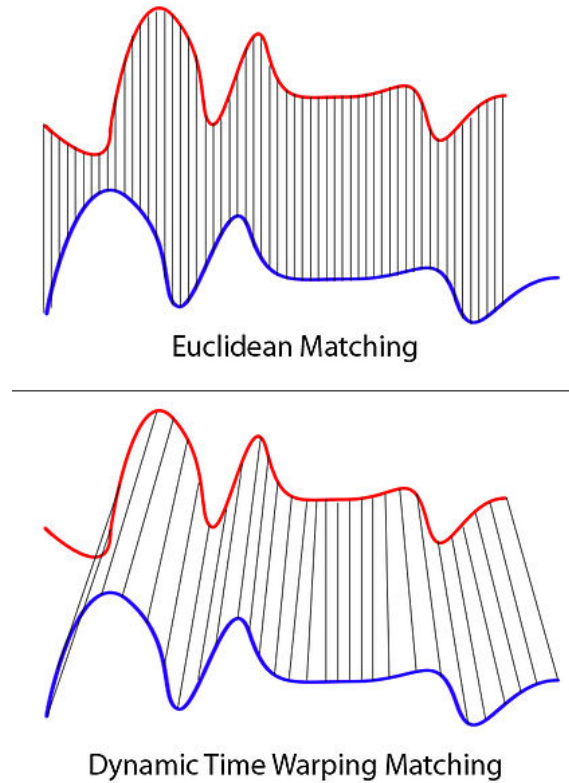


Figure 7: Illustration from Commons (2016) showing the difference between Euclidean matching and dynamic time warping matching.

4.1.2 Fréchet distance measure

The Fréchet distance measure is a distance measure which measures the similarity between curves. Gudmundsson & Wollé (2014) use the Fréchet distance measure to

compare the movement of football players on the pitch. In this thesis it is tested whether this distance measure could also be useful when measuring the similarity of possession sequences in football. Intuitively, the Fréchet distance can be compared to a person walking with its dog. Both the owner and the dog walk their own path, where the Fréchet distance is the minimum length of the leash between the dog and its owner.

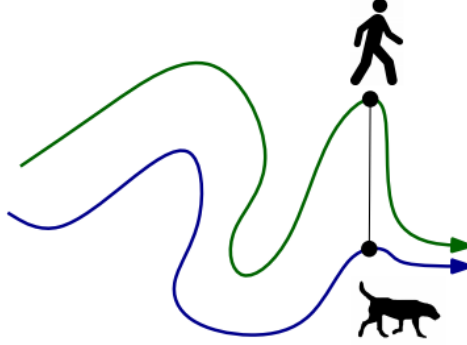


Figure 8: Illustration from Driemel (2016) illustrating the Fréchet distance.

As the position of the ball is not known continuously but only when a player touches the ball, the discrete Fréchet distance can be used which gives an approximation of the continuous Fréchet distance. Formally the discrete Fréchet distance between two curves $P = (p_1, \dots, p_n)$ and $Q = (q_1, \dots, q_m)$ can be defined as follows. First let us define a coupling L between P and Q as a sequence of distinct pairs from $P \times Q$ such that the begin points of both curves are coupled, the end points of both curves are coupled and that the order of the points in P and Q are respected. The rules for the coupling are thus the same as for the dynamic time warping distance. Then the length of the coupling L , $||L||$, is the length of the longest distance between two points in the coupling in L . Now, the discrete Fréchet distance can be defined as follows:

$$\delta_{dF}(P, Q) = \min\{||L|| \mid L \text{ is a coupling between } P \text{ and } Q\}$$

4.1.3 Longest common subsequence distance measure

Another way of measuring the similarity of two sequences is the Longest Common Subsequence (LCSS) model. This distance measure is described in detail by Vlachos et al. (2002). In a nutshell, the measure counts the number of "common" points in the two sequences. A certain value of time stretching is allowed and points are seen as common when the distance between them is smaller than a chosen threshold. The number of common points is then divided by the length of the shortest sequence of the two sequences. In this way, a value between 0 and 1 is obtained and the higher this value, the more similar the two sequences are. Therefore, the distance measure for two sequences P and Q is defined as follows:

$$D(P, Q) = 1 - \frac{\# \text{ "common" points in P and Q}}{\min\{length(P), length(Q)\}}$$

Figure 9 illustrates the Longest Common Subsequence.

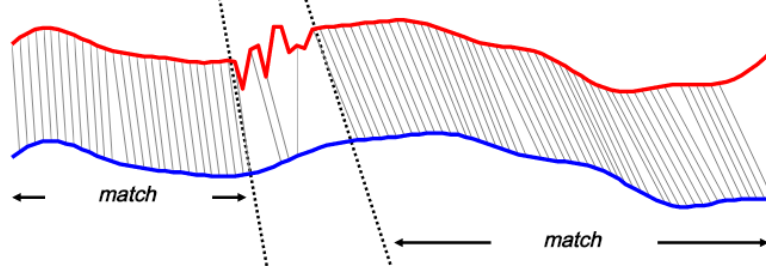


Figure 9: Illustration from Zeinalipour (2007) illustrating the Longest Common Subsequence distance. It can be observed that the Longest Common Subsequence ignores outliers.

4.1.4 Comparison of the distance measures

Both the DTW and Fréchet distance measures require an optimal coupling between both sequences. However, dynamic time warping sums up all distances between the coupled points and the Fréchet distance only looks at the maximum distance between coupled points. The Fréchet distance is very sensitive to outliers as it is equal to the maximum distance between the coupled points. The dynamic time warping distance is also sensitive to outliers as it also forces all elements to be matched.

Contrary to the DTW distance measure and the Fréchet distance, the LCSS method does not force all elements to be matched. This makes it possible to handle outliers better (Müller (2007)). The LCSS method focuses on similar points in the sequences, the Fréchet distance focuses on the pair of most dissimilar points and the DTW distance measure focuses on the dissimilarity of the optimal coupling.

4.2 K-nearest neighbors algorithm

The k-nearest neighbors algorithm is a non-parametric algorithm that for each object searches for the k nearest objects. A prediction or classification is done based on these k nearest neighbors. The k-nearest neighbors algorithm does not have a training phase. However, in the testing phase more computations may be required and also all training data need to be stored. The k-nearest neighbors algorithm is not fast when all distances have to be calculated between all objects in the test and training data sets. Fortunately, the speed can be improved by using indexing and only calculate those distances necessary to calculate. In this thesis pre-clustering is applied on both the training data set and test data set, such that only the similarity is measured between objects that are expected to be quite similar. All objects are assigned to a pre-cluster and for each of the objects of the test set, the nearest neighbors are only searched for in their own pre-cluster. In this way, the computation time is limited.

5 Approaches

In this thesis we consider three approaches of assigning values to passes. In this chapter we explain these three approaches, the zone-oriented pass value (ZPV) approach, the pass-oriented pass value (PPV) approach and the sequence-oriented pass value (SPV) approach, in more detail.

Firstly, the zone-oriented pass value (ZPV) uses the value of possessing the ball in a certain zone of the pitch to value passes. Secondly, the pass-oriented pass value (PPV) measures the similarity of individual passes using a newly introduced distance measure for passes. Lastly, the sequence-oriented pass value (SPV) relies on finding similar patterns in play, and passes are valued on their influence on the outcome of their possession sequences.

5.1 Zone-oriented pass value (ZPV)

The ZPV approach divides the pitch into a number of equal sized zones. Using historical data, we give each zone a value for possessing the ball in that zone. We calculate the value for the zones in two different ways: the 15 seconds rule and the expected goals model.

1. The 15 seconds rule: each zone is valued as the expectation that a goal is being scored from that zone within a time frame of 15 seconds
2. The expected goals model: each zone is valued as the expectation that if a shot is taken from that zone that it results into a goal being scored

In both ways we value each pass as follows: the possession value of the zone it ends in minus the possession value of the zone it came from. In this way, the pass is valued as the increase in the zone possession value. This approach is simple and is therefore a good one to use as a standard to which to compare the other approaches. The other approaches require more computations and are therefore more time-consuming.

5.1.1 Algorithm

There are three steps to be taken to calculate the ZPV for each of the passes of the test set.

1. Partition the pitch into zones.
2. Determine the zone values using the training data, either by the 15 seconds rule or with the expected goals model.
3. For each of the passes of the test data set determine its ZPV by subtracting the zone value of the origin of the pass from the zone value of the destination of the pass.

In the following subsections we explain each of these steps in more detail. Figure 10 illustrates the ZPV approach.

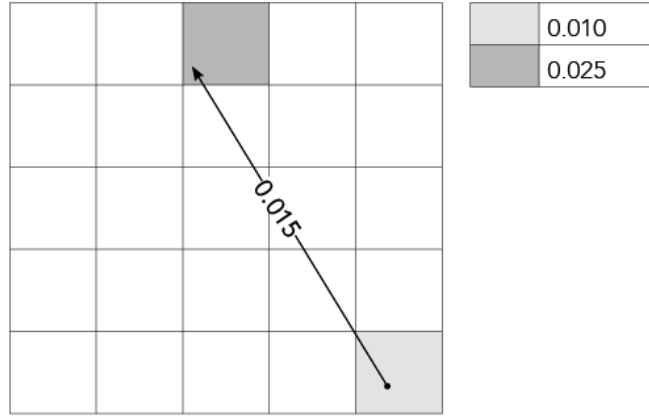


Figure 10: Illustration of the ZPV approach. The pass is given from a zone with value 0.010 to a zone with value 0.025 and thus the pass is assigned a value of 0.015

5.1.2 Step 1: Partitioning of the pitch

We evaluate the ZPV approach for two different partitions of the pitch. It is preferred to create an odd number of zones from top to bottom as well as an odd number of zones from left to right as this creates zones right in front of the goal and zones on the centerline of the pitch. For the ZPV approach it is chosen to partition the pitch either into zones of five meters long and four meters wide or into zones of one by one meters. For both the 15 seconds rule and the expected goals model based on the validation data set we decide which of these pitch partitionings works best.

5.1.3 Step 2: Valuing the zones using the training data

As said before, two ways of valuing the zones are observed for the ZPV approach: the 15 seconds rule and the expected goals model. We explain both ways in the following paragraphs.

5.1.3.1 ZPV - 15 seconds rule

The 15 seconds rule is defined as the probability that when possessing the ball in a certain zone of the pitch there will be a goal scored by the team possessing the ball within 15 seconds. Let G_{z_i} be the total number of goals scored within 15 seconds when the ball is possessed in zone z_i , for $i = 1, \dots, n$. Let T_{z_i} be the total number of times the ball is possessed in zone z_i , for $i = 1, \dots, n$. As only event data is given, the passes and goal attempts originating from the zones are taken into consideration. Both G_{z_i} and T_{z_i} are calculated for the complete training data set. Now, the zone value $V15(z_i)$ is defined as follows:

$$V15(z_i) = \frac{G_{z_i}}{T_{z_i}}$$

Figure 11 shows for each zone of five meters long and four meters wide the probability that a goal is scored within 15 seconds when possessing the ball in that zone. All data is adjusted such that the direction of attack is always from left to right. In Figure 11 it can be observed that the zone right in front of the opponent's goal has a

very high value when compared to the rest of the pitch. This means that possessing the ball in this zone leads in about 60% of the cases to a goal within 15 seconds. This result is expected as possessing the ball from such a small distance from goal is very valuable. However, the high value for the zone in front of the own goal is unexpected. This might be caused by the fact that when the goalkeeper catches the ball in this area he often starts a counter attack which may lead to a goal being scored in the next 15 seconds. However the high value for this zone might also be caused by inconsistency in the data.

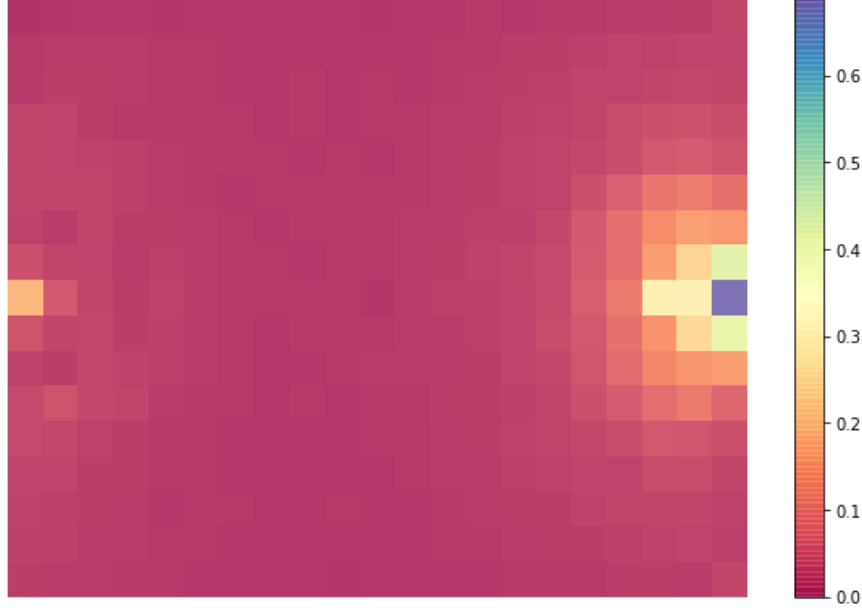


Figure 11: Heat map of the probabilities of a goal being scored within 15 seconds for zones of size five by four meters. The direction of attack is from left to right.

5.1.3.2 ZPV - expected goals model

The second way of valuing the zones relies on the expected goals model. For both the partitioning of five by four meters zones and the partitioning of one by one meters zones, the expected goals value for each of the zones is determined. In section 3.3 the expected goals model is explained. In short, the expected goals model determines for each zone the probability that a goal is scored when a goal attempt is made from this zone. Let $VXG(z_i)$ denote the expected goals value of zone z_i , for $i = 1, \dots, n$. In the next section it is explained how these values are used to value the passes.

5.1.4 Step 3: Determining the pass value

Pass p_j of the test data set, with origin zone z_{o_j} and destination zone z_{d_j} is valued as follows:

$$\begin{aligned} ZPV15_{p_j} &= V15(z_{d_j}) - V15(z_{o_j}) \\ ZPVXG_{p_j} &= VXG(z_{d_j}) - VXG(z_{o_j}) \end{aligned}$$

Where $ZPV15_{p_j}$ denotes the ZPV of pass p_j for the 15 seconds rule and $ZPVXG_{p_j}$ denotes the ZPV of pass p_j for the expected goals model. To summarize, the ZPV

measures the increase or decrease from the zone possession value from before the pass and after the pass.

5.2 Pass-oriented pass value (PPV)

The pass-oriented pass value relies on measuring the similarity of passes. We introduce a new distance measure for passes that makes sure that not only the geometric features of the pass are taken into account, but also takes as input the history of the possession sequence leading to the pass. Then, we value the individual passes on the average outcome of the possession sequences of the most similar passes. We explain the PPV approach in more detail in this section.

5.2.1 Algorithm

The pass-oriented pass value approach consists of five steps:

1. Define a distance measure to measure the similarity of passes
2. For all passes of the training set determine the outcome of the possession sequence it belongs to
3. Divide the passes of the training set over different pre-clusters by only taking into account their origin and destination
4. For each of the passes of the test set determine its pre-cluster and find the k nearest neighbors in the pre-cluster based on the introduced distance measure. The pass is assigned the average value of the outcomes of the neighbors' possession sequences

We explain each of these steps in further detail in the next paragraphs. Figure 12 illustrates the pass-oriented pass value (PPV).

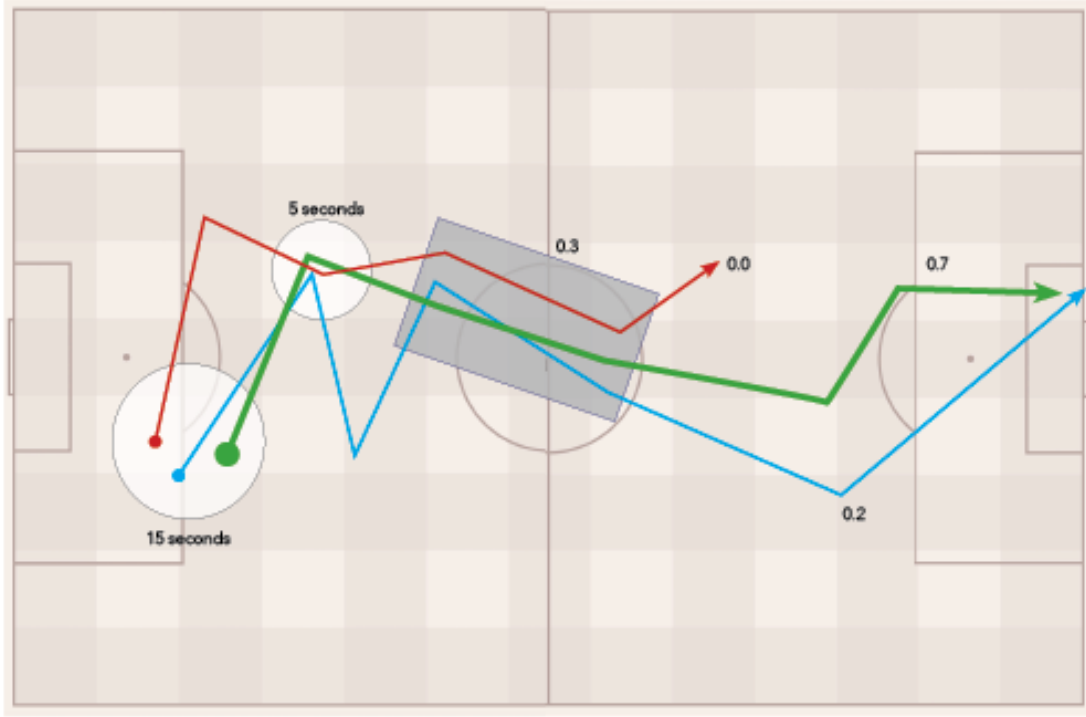


Figure 12: Each of the colored lines represent a possession sequence. The red possession sequence ends in a loss of possession and thus an outcome of zero. The green and blue sequences both end with a goal attempt, with respectively an expected goals value of 0.7 and 0.2. The passes in the grey box have a PPV of 0.3 (mean of 0.7, 0.2 and 0) as they have similar geometric features and the ball was in nearly the same location five and 15 seconds before the passes were performed.

5.2.2 Step 1: Defining and evaluating the distance measure

A distance measure is needed to be able to measure how similar passes are and to find for each pass the most similar passes. Measuring the similarity of passes using only event data is difficult because for each passing situation it is not known where all other players are on the pitch. Therefore, the current situation of play is not known in full detail and an inventive distance measure is needed. This is why the location of the ball preceding the pass is taken into account in order to be able to trace down the situation of play of each individual pass. When for example five seconds ago the ball was on the other side of the pitch, it may be assumed that the current attack is a counter attack. In a counter attack there usually is more space and it is more fair to compare passes in a counter attack with each other and to not compare them to passes in a situation where the defending team is situated well.

We introduce a new distance measure for passes. The following aspects are incorporated in the distance measure:

1. The difference in length of the passes
2. The Euclidean distance between the origins of the passes

3. The Euclidean distance between the destinations of the passes
4. The location of the ball five seconds before the passes are given
5. The location of the ball 15 seconds before the passes are given

A combination of these aspects, by using weights for each of these aspects, leads to a distance measure for the passes.

5.2.2.1 Determining the weights

By adjusting the weights multiple distance measures can be produced. First, for a subset of 1000 passes of the training set two different weight-sets are analyzed. Table 3 describes these weight-sets.

	w_{origin}	$w_{destination}$	w_{length}	w_{5secs}	w_{15secs}
W_1	1/4	1/4	1/4	1/6	1/12
W_2	1/5	1/5	1/5	1/5	1/5

Table 3: The different weight-sets that are analyzed

The similarity of passes obtained using these two weight-sets are analyzed with football experts. First, we analyze a subset of 1000 passes in order to find out which method works best. Figure 13 shows the red pass' 10 nearest neighbors in this data set using weight-set W_1 . We choose the weights of this weight-set in this way as we consider the geometric features of the passes to be most important when comparing passes. Furthermore, we consider the position of the ball five seconds ago to say more about the pass than the position of the ball 15 seconds ago, and therefore is weighted twice as much.

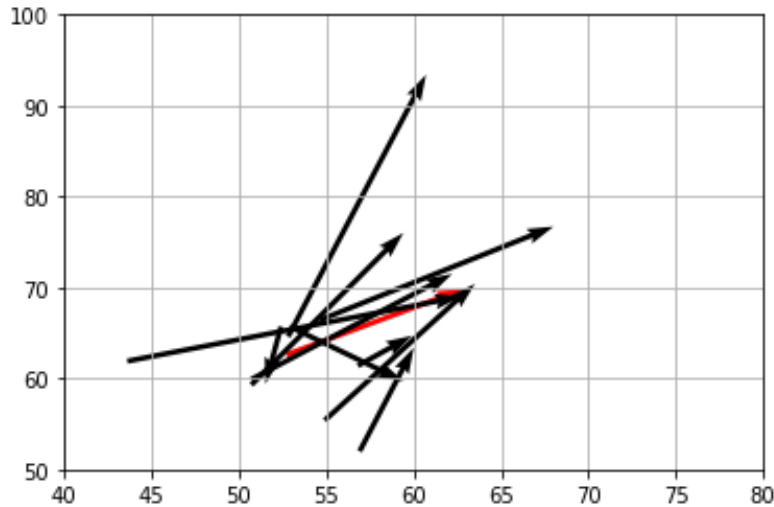


Figure 13: Close up of the red pass' 10 nearest neighbors according to weight-set W_1

Using weight-set 2 to measure the distances results in the 10 nearest neighbors of this data set as shown in Figure 14. The red arrow represents the same pass as in the previous figure. This yields completely different nearest neighbors. The weights given to the parameters have a huge influence on the final distance measured between the passes.

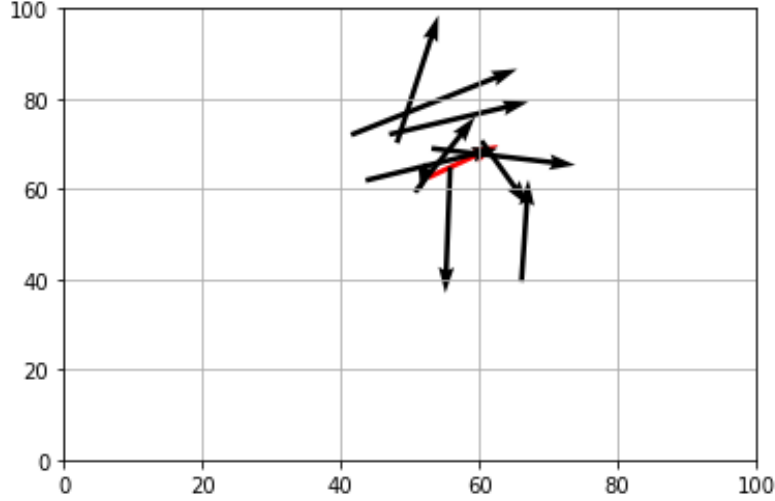


Figure 14: Close up of the 10 nearest neighbors of the red pass

Observing these two figures, and there are many more examples, we conclude that weight-set W_1 obtains the best results, yielding small distances between comparable passes in a similar situation of play. Therefore we use this weight-set to measure the distances between passes in the PPV approach.

5.2.3 Step 2: Determining the outcome of the possession sequences

For each of the passes in the training set we need to determine the outcome of its possession sequence. As explained before, this outcome depends on the fact whether a goal attempt was made on the end of the attack and if so, what expected goal value this attempt has. In this way, every pass in the training set has a value: the value of the outcome of its possession sequence. We use these values in step 4 when assigning values to the passes in the test set.

5.2.4 Step 3: Pre-clustering the passes

It is computationally not possible to calculate the distances between the millions of passes of the training and test set in a reasonable amount of time. Therefore, we distribute the passes of the training set over a number of different pre-clusters. This speeds up the search for the nearest neighbors of the passes of the test set in the next step.

The pre-clustering of the passes only takes into account the origin and the destination of the passes. For both the origin and the destination of a pass the zone

of the pitch it belongs to is determined. When two passes have similar origin and destination zones, the passes are assigned to the same pre-cluster. By doing this, also the length of the passes is indirectly taken into account as passes starting and ending in the same zone must have nearly equal lengths. We desire to end up with pre-clusters of about equal sizes and each of the pre-clusters should not consist of too many passes. A good trade-off between the number of pre-clusters and the size of the pre-clusters limits the computation time. On the other hand, we also desire not to make too many assumptions that cause similar passes to end up in different pre-clusters. Therefore, we use zones of five meter long and four meter wide to pre-cluster the passes.

A problem that occurs when using this method of pre-clustering is that the boundaries of the zones are strict which causes similar passes to end up in different pre-clusters. Figure 15 shows an example of two very similar passes, with different origin and destination zones. Applying this pre-clustering on the passes has as a consequence that those two passes do not end up in the same pre-cluster. Therefore, these passes are not seen as each others neighbors when the passes are clustered.

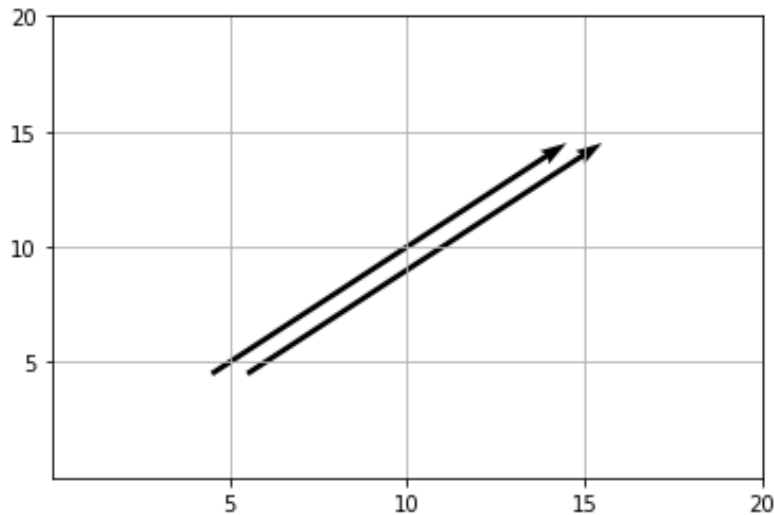


Figure 15: Figure showing two very similar passes that end up in different pre-clusters

The two passes in Figure 15 ending up into different pre-clusters is of course not desired. Therefore, we assign origins and destinations that are close to the boundaries of the zones to multiple zones. When the distance to a neighboring zone is lower than a certain threshold, we also assign the origin or destination to this neighboring zone. Thus, in the extreme case where the origin or destination is in the far corner of a zone we assign it to four zones. This may result in a pass ending up in at most sixteen (four times four) pre-clusters. We search for the nearest neighbors of a pass in each of its pre-clusters, taking into account that a pass is selected at most only once as nearest neighbor.

A football pitch of 105 by 68 meters consists of 357 five by four meters zones and thus $357 \cdot 357 = 127449$ different pre-clusters of passes are created. The zones are chosen to be five by four meters as this causes the number of zones on the x-axis and

on the y-axis to be odd, such that the center of the pitch is captured well. Next to that, zones of this size causes the sizes of the pre-clusters to be reasonable as well as the number of pre-clusters. When using a threshold value of one meter for assigning passes to multiple pre-clusters, the largest pre-cluster consists of 19844 passes. This pre-cluster is the one with passes originating in the zone $[50 : 55, 32 : 36]$, so from the half way line till just five meters passed the half way line and the center part of the pitch, and ending in the same zone. It is not unexpected that this pre-cluster contains the highest number of passes, as most passes in football occur on the midfield. The size of this pre-cluster is reasonable for clustering. Furthermore, there are some very small and 36914 empty pre-clusters. The pre-clusters of size at least one consist on average of 315 passes. The sizes of the pre-clusters are reasonable to search for the k-nearest neighbors. Therefore we choose to use these zones.

5.2.5 Step 4: Determining the pass value

When the pre-clusters of passes are created, the only task left is assigning values to the passes. We assign the passes a value equal to the average of the sequence value of similar passes. We search for the k nearest neighbors of each pass in the pass' pre-cluster(s). We use the newly introduced distance measure to measure the similarity of the passes. The PPV of the pass is the average value of the expected outcomes of the possession sequences of its k nearest neighbors. In this thesis we use $k = 100$.

5.3 Sequence-oriented pass value (SPV)

The third approach of valuing the passes is called the sequence-oriented pass value (SPV). This approach takes into account the possession sequence the pass is in. Intuitively, the pass is valued on the influence it has on the expected outcome of its possession sequence. It requires values for the possession sequences. Next to that, we value the subsequences of the sequences such that we can measure the influence of the pass on the sequence's outcome.

The (sub)sequences are compared to each other in order to find out how valuable a certain (sub)sequence is. To do this, we use three different distance measures: the DTW distance measure, the Fréchet distance measure and the LCSS distance measure. First, we evaluate these distance measures on the training data set to find out whether they are applicable for football possession sequences. The distance measures that seem applicable are then used to measure the distance between the possession sequences and to value the (sub)sequences.

When all (sub)sequences are compared and assigned a value of the expected outcome of the sequence, the values of the passes can be determined. We calculate the value of a pass as the difference in value between the subsequence that ends with this pass and the subsequence that ends just before this pass. In this way we determine the contribution of the pass to the outcome of the sequence.

Figure 16 illustrates the SPV approach.

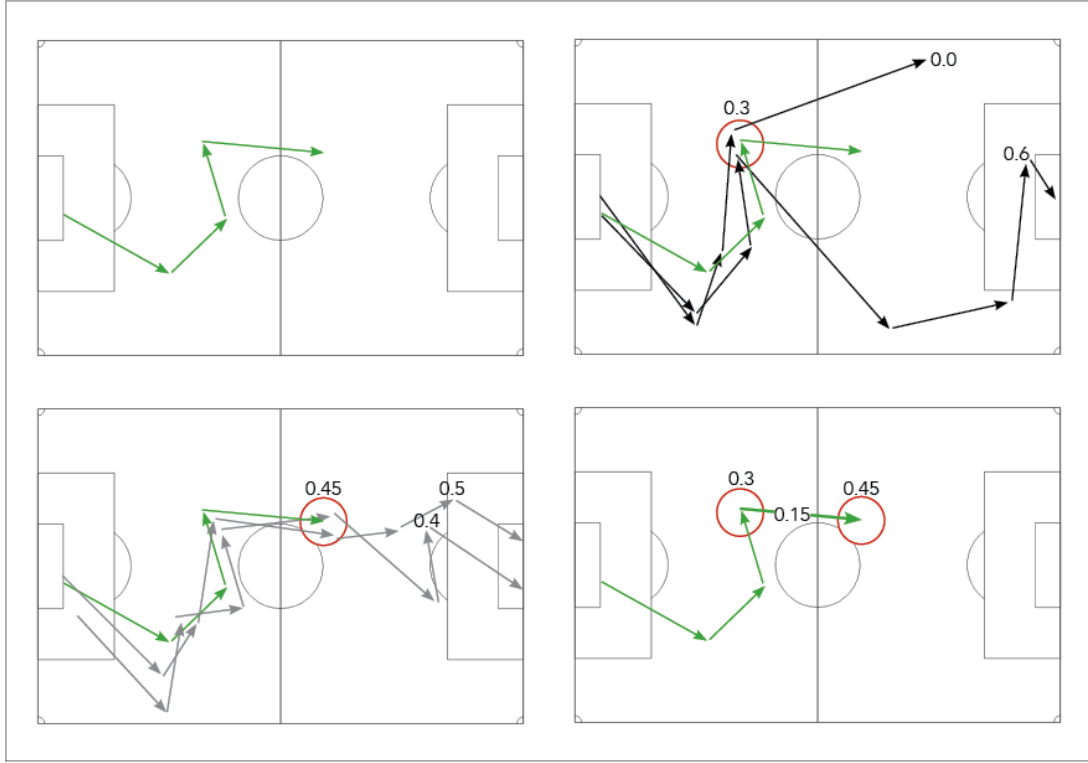


Figure 16: The last pass of the green sequence is being valued in this illustration. The two sequences with a similar possession sequence up until the valued pass end up into either loss of the ball (value 0.0) or a goal attempt with an expected goal value of 0.6. That is why the start of the green pass has a value of 0.3 (average of 0.0 and 0.6). The two sequences that are similar to the complete green sequence end in goal attempts with expected goals values of 0.5 and 0.4. Therefore, the end of the green pass has a value of 0.45 (average of 0.5 and 0.4). Now that the end of the green pass has value 0.45 and the beginning has a value of 0.3, the green pass gets an SPV of $0.45 - 0.3 = 0.15$.

5.3.1 Algorithm

The SPV approach consists of the following five steps:

1. Step 1: Determine which distance measures to use for comparing the (sub)sequences
2. Step 2: Break up the sequences in subsequences and value them
3. Step 3: Pre-cluster the subsequences of the training set and the test set on their origin and destination
4. Step 4: Calculate the distances between the subsequences and find the 100 nearest neighbors
5. Step 5: Value the passes

In the following paragraphs we explain each of these steps in more detail.

5.3.2 Step 1: Determine which distance measures to use

As the possession sequences differ in length and time, creating a distance measure for the passing sequences is not straightforward. The sequences of passes can be viewed as time series; each pass has an origin, a destination and a time frame. We try three distance measures to measure the similarity between passing sequences: Dynamic Time Warping (DTW), the Fréchet distance measure and the Longest Common Subsequence distance measure. We explained the methodology of these distance measures in the Machine Learning chapter.

To test whether the distance measures are good measures for football possession sequences, we select a subset of the training set containing 1000 possession sequences with a length of at least five passes on which to test the distance measures. First, we calculate the distances between these possession sequences for each of the distance measures. Then, we further analyze the two most similar possession sequences for each of those distance measures.

Figure 17 shows the two possession sequences of the subset with the lowest DTW distance. It can be observed that these sequences indeed are much alike.

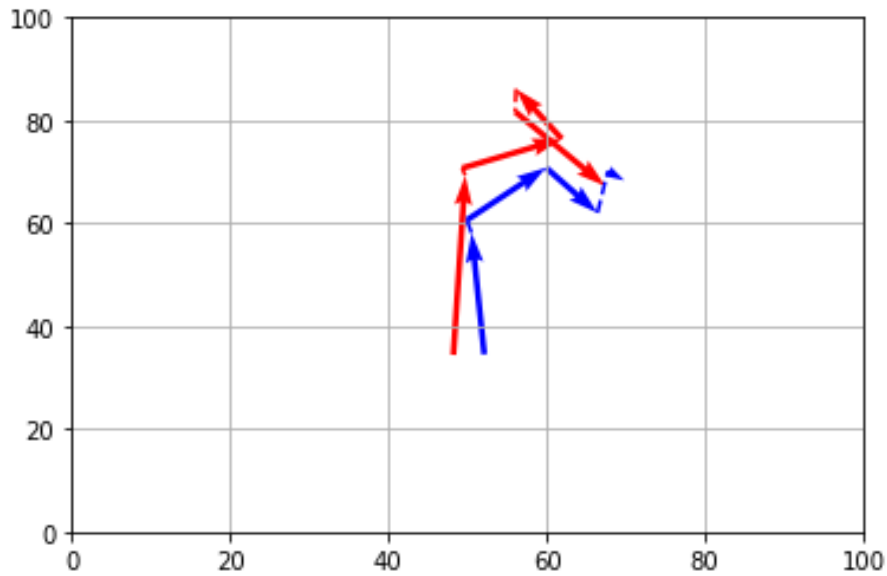


Figure 17: Two sequences with the lowest DTW distance value

Figure 18 shows the two possession sequences of the subset which have the lowest Fréchet distance. It can be observed that these sequences indeed are a lot alike.

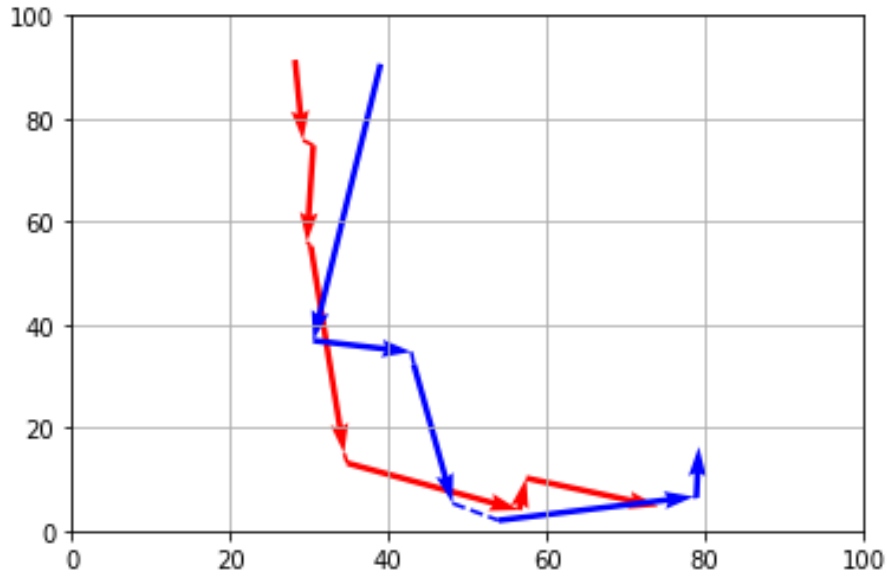


Figure 18: Two sequences with the lowest Fréchet distance value

Figure 19 shows the two possession sequences of the subset which have the lowest Longest Common Subsequence (LCSS) distance measure for a threshold of five meters. The beginnings of those sequences are very much alike, however the end points of the sequences are far apart. Intuitively, those two sequences should not be seen as very similar and should therefore not have such a low distance value.

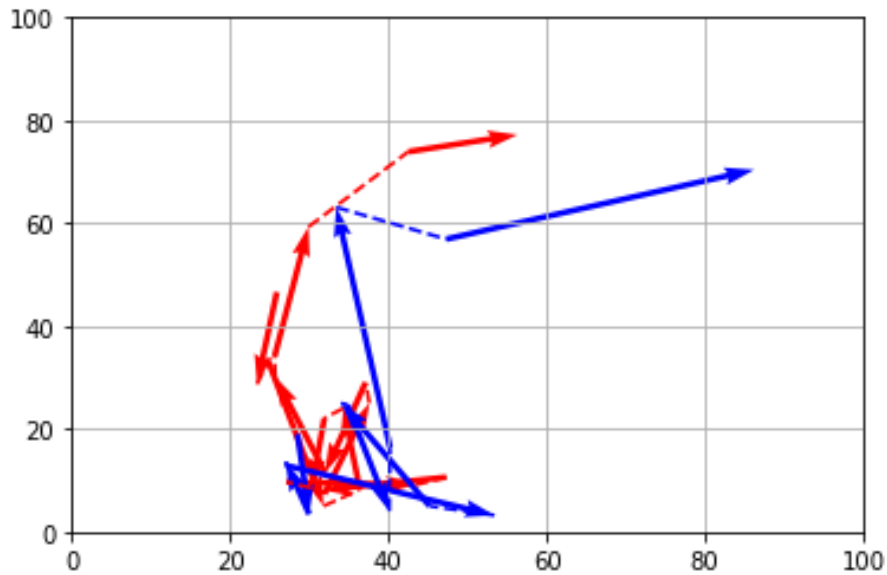


Figure 19: Two sequences with the lowest LCSS distance value for a threshold of five meters

The LCSS performs even worse when the threshold is put to 3 meters. Figure 20 shows the two sequences of the subset with the lowest LCSS distance. As the LCSS distance measure only looks at the similarities of the sequences and ignores the differences, it seems to be a distance measure that does not suit well for comparing possession sequences in football.

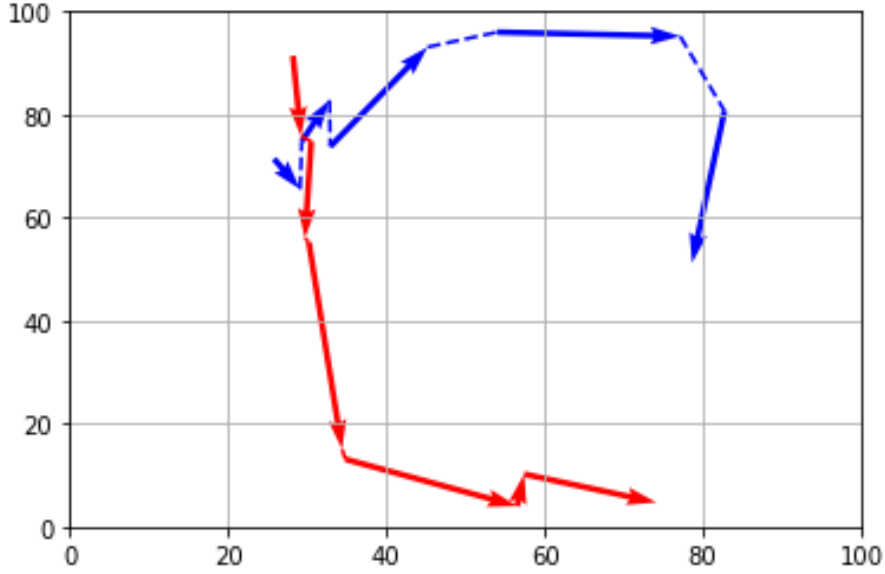


Figure 20: Two sequences with the lowest LCSS distance value for a threshold of 3 meters

From the examples in this section, and there are many more, we conclude that the DTW distance measure and the Fréchet distance measure both seem to work well on defining the similarity of the possession sequences in football. However, the LCSS distance measure seems not to work well and therefore we do not use this distance measure for comparing possession sequences.

5.3.3 Step 2: Break up the sequences into subsequences

The SPV approach does not only require the values of the possession sequences, but also requires the values of the subsequences of the possession sequences. These subsequences have the same first pass as the possession sequence and have a length of at least 2. Let possession sequence S_j be denoted by the sequence of actions $[p_1, \dots, p_{l_j}]$, where l_j is the number of actions in possession sequence S_j . The subsequences of each possession sequence are denoted by $S_j(r) = [p_1, \dots, p_r]$, $r = 2, \dots, l_j$. We assign a value to all subsequences by valuing their outcome in the same way as the complete possession sequences are valued.

5.3.4 Step 3: Pre-cluster the possession (sub)sequences

In a way similar way to the pre-clustering of the passes, we assign the possession sequences to different pre-clusters in order to speed up the process. We assign (sub)sequences to pre-clusters taking into account the coordinates of the starting point of the sequence and the coordinates of the end point of the sequence.

For the possession sequences it is very important what happens between the starting point and end point, therefore those two criteria are not made too strict in order to prevent very similar sequences to end up in different pre-clusters. That is why we use bigger zones for this pre-clustering than those of the pre-clustering of the passes in the pass-oriented pass value (PPV) approach. We use zones of ten by eight meters. Next to that, we use a threshold of 2 meters for the corners of the zones to

assign sequences that are within 2 meters of the boundaries to multiple pre-clusters. In the next step, we only compare the sequences of the test set to the sequences of the training set that are in the same pre-cluster. This speeds up the algorithm.

5.3.5 Step 4: Calculate the distances and find the k nearest neighbors

In this approach, we use two distance measures for the similarity of possession sequences: the DTW distance measure and the Fréchet distance measure. In step 3, we distributed the possession (sub)sequences of both the training set and the test set over the pre-clusters. Next, we calculate the distances to all (sub)sequences of the training set in the same pre-cluster for each (sub)sequence of the test set. When a (sub)sequence of the test set belongs to multiple pre-clusters, due to the fact that its begin or end point is close to the border of a zone, we calculate the distances to all (sub)sequences of all pre-clusters it belongs to. If the group of possible neighbors is smaller than k, due to a small pre-cluster, we choose all these (sub)sequences as the nearest neighbors. Also in the SPV approach we set k equal to 100.

5.3.6 Step 5: Assigning values to the passes

Now that for each (sub)sequence of the test set its nearest neighbors are known, we determine the value of each (sub)sequence. We assign a value to each subsequence as the average outcome of the nearest neighbor (sub)sequences. Let the value of subsequence $S_j(r) = [p_1, \dots, p_i, \dots, p_r]$ be denoted by $V(S_j(r))$. When all (sub)sequences are valued, it is possible to determine the value of the individual passes. We calculate the SPV value of pass p_i as follows.

$$SPV(p_i) = V(S_j(i)) - V(S_j(i-1))$$

In words, the SPV value of the pass is the value of the subsequence that ends with this pass minus the value of the subsequence that ends just before this pass. In this way we calculate the expected contribution of the pass to the sequence's outcome.

5.3.7 How does DTW compare to the Fréchet distance?

We determine the 100 nearest neighbors for each of the possession sequences in the SPV approach using the DTW distance and using the Fréchet distance in order to value the passes. Those 100 nearest neighbors are in the 2015/2016 season on average for 62.69% the same for both distance measures and in 2016/2017 season 62.35%. In particular, for sequences that travel more distance, there are more differences between both distance measures.

5.3.8 Application: identifying teams' tactics

The comparison of the possession sequences to obtain the SPV values can also be used to identify teams with similar possession sequences. This helps improving a team's playing style as well as being able to better analyze the attacking style of the opponents.

The identification of the teams' attacking styles is done as follows. For each possession sequences of the 2016/2017 season the 100 (if the pre-cluster size is greater than 100) nearest neighbors using the DTW distance measure are known. These 100 neighbors are possession sequences from the seasons 2012/2013, 2013/2014, 2014/2015 and 2015/2016 and were executed by a team in one of these seasons. Let $T_{old} = old_1, old_2, \dots, old_{98}$ denote the set of the 98 teams of these seasons. And let $T_{new} = new_1, new_2, \dots, new_{78}$ denote the set of the 2016/2017 teams. The $|T_{new}| \times |T_{old}|$ -matrix P is defined as follows:

$$P_{new_i, old_j} = \frac{\# \text{ sequences of team } old_j \text{ that are neighbor of sequences of team } new_i}{\# \text{ sequences of team } new_i}$$

for $i = 1, \dots, 78$ and $j = 1, \dots, 98$. So for each team of T_{new} a vector of length 98 is build that represents the percentage of comparison between the teams. The goal is to cluster the teams of T_{new} by using these vectors as input. However, the size of these vectors are too big to cluster on. This problem is also known as the curse of dimensionality and therefore dimension reduction is applied on the matrix P to create a matrix on which it is possible to cluster. The number of dimensions is brought back to three dimensions by using Principal Components Analysis. Figure 21 shows the elbow plot that indicates that three components is the best choice.

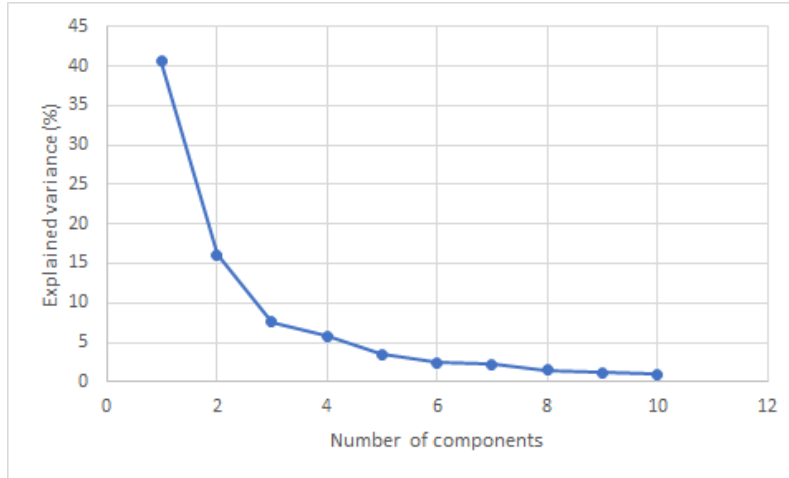


Figure 21: Elbow plot of the Principal Components Analysis showing the explained variance of the principal components on the y-axis and the number of principal components on the x-axis. It shows that 3 components is at the 'elbow'.

We apply k-means clustering on the new matrix by using the Euclidean distance as distance measure. We perform the clustering for values of k ranging from two to ten. Figure 22 shows the elbow plot where the y-as represents the average within-cluster sum of squares. This elbow plot indicates that four clusters best segment the data. However, when we analyze the results of the different clusterings, it turns out that six clusters gives the best clustering. The six clusters that result from the k-means clustering are best interpretable. Figure 23 shows the scatter plot of the six clusters.

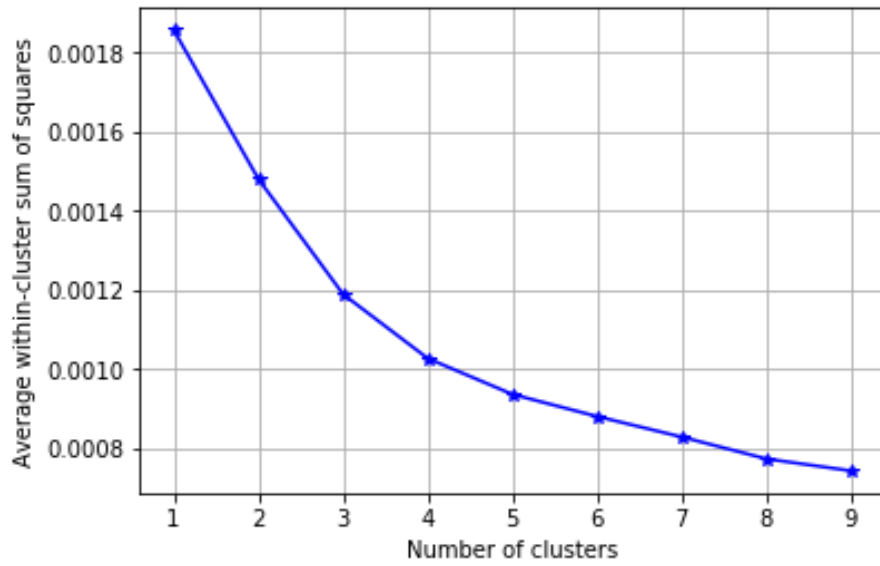


Figure 22: Elbow plot of the k-means clustering showing the average within-cluster sum of squares on the y-axis and the number of clusters on the x-axis.

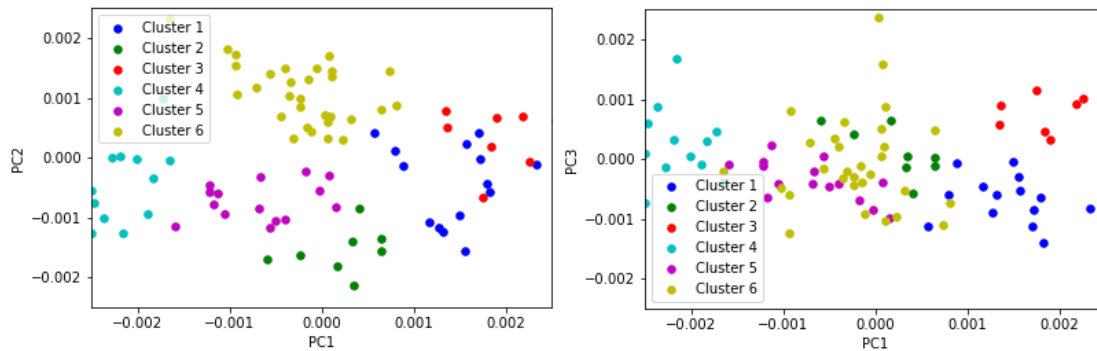


Figure 23: Scatter plot for the six clusters. The left figure shows the data points plotted on the first two principal components. The right figure shows the data points plotted on the first and third principal component. In the left figure all clusters, except clusters 1 and 3, are clearly formed. The right figure shows the distinction between cluster 1 and 3 with respect to the third principal component.

Table 4 shows the six clusters that result from the k-means clustering.

Cluster 1	Cluster 2
Barcelona, Napoli, Bayern München, Nice, PSG, Manchester City	Sunderland, FC Augsburg, West Bromwich Albion, FSV Mainz 05, FC Ingolstadt 04, Burnley, SV Darmstadt 98
Cluster 3	Cluster 4
Real Madrid, Arsenal, Chelsea, Manchester United, Liverpool, Juventus, Fiorentina, AS Roma, Borussia Dortmund, Sevilla, Lyon, Tottenham Hotspur, Las Palmas, Marseille	AC Milan, Internazionale, Atlético Madrid, Monaco, Torino, Atalanta, Bologna, Lazio, Sampdoria, Borussia Mönchengladbach, Real Sociedad, Celta de Vigo, Villareal, Lorient, Montpellier, Southampton, Everton, Empoli, TSG 1899 Hoffenheim, Lille, Rennes, Bournemouth
Cluster 5	Cluster 6
Udinese, Eintracht Frankfurt, SC Freiburg, Granada CF, Sporting de Gijón, Bastia, Toulouse, Watford, Leicester City, Stoke City, FC Köln, Werder Bremen, Eibar, Crystal Palace	Valencia, Chievo, Bordeaux, Swansea City, Sassuolo, Palermo, Bayer 04 Leverkusen, Schalke 04, Hertha BSC, Deportivo de La Coruna, Guingamp, St Etienne, Angers, Caen, Nantes, West Ham United, Hull City, Genoa, VFL Wolfsburg, Real Betis, Espanyol, Málaga, Athletic Club

Table 4: Clusters resulting from the comparison of possession sequences using the DTW distance measure.

The clustering of the teams using the results from the DTW distance measure leads to interesting results. The teams of cluster 1 are all well-known for their possession-based style of play and do all top their league. The teams in cluster 3 are also mostly teams that play in the top of their league. However, they are known for playing on the wings more than the teams in cluster 1. Cluster 4 mostly consists of teams that are sub-toppers in their league. Clusters 5 and 6 contain teams that are more in the bottom of the league. Cluster 2 is an interesting one as it contains 7 teams that often play against relegation and their type of play is very defensive. These teams mainly live on counter attacks which is probably why they are in the same cluster.

5.4 Comparison of the approaches

This thesis introduces three new ways of valuing passes in football, the zone-oriented pass value (ZPV), the pass-oriented pass value (PPV) and the sequence-oriented pass value (SPV). These different approaches have some similarities as well as some clear differences. In this section, we explain the similarities and differences of these approaches in more detail, to make the distinction of the approaches clear.

Table 5 shows the features of the approaches.

Feature	ZPV	PPV	SPV
Partitioning of the pitch needed	X	X	X
Expected goals model	X	X	X
15-seconds rule	X		
Similarity of passes		X	
Similarity of possession sequences			X
Value before and after pass taken into account	X		X
Valuing possession sequences required		X	X
Pre-clustering applied		X	X
k-nearest neighbors		X	X
k-medoids clustering			X

Table 5: An overview showing the features of the different approaches to value passes

The three approaches that we introduce in this thesis all have their weaknesses and strengths. The first approach, the ZPV, is the most simplistic and by far the fastest one to assign values to passes. When the values of the zones are determined it is simply a subtraction of the value of the zone of the pass’ destination and the value of the zone of the pass’ origin.

All approaches require an expected goals model. The ZPV approach uses this model to value the zones, whereas the PPV approach and the SPV approach use the expected goals model to assign values to every possession sequence that ends with a goal attempt. In the PPV approach, we use the values of the possession sequences’ outcomes (e.g. goal attempt or not) to value the nearest neighbors of a pass. We also use the values assigned to the outcomes of the possession sequences in the SPV approach. However, in this approach also the subsequences play a role.

One of the differences between the PPV approach and the SPV approach is that in the SPV the similarity of possession (sub)sequences is measured and passes are valued on the influence they have in their possession sequence, whereas in the PPV approach the similarity of individual passes is measured.

In the ZPV approach solely the origin and destination of the pass are taken into account and the expected possession values of possessing the ball before the pass and after the pass determine the value of the pass. We expect that this naive approach highly rates crosses that find a teammate in a crowded penalty box. In this approach we do not take into consideration whether similar passes led to goals being scored or not and in what situation of play the pass was given. That is where the PPV approach comes in. It takes into account whether similar passes led to goal opportunities. Passes are not only compared on their origin, destination and length, but by looking at the position of the ball five and 15 seconds ago the history of the game is captured in the distance measure. In the PPV approach, however, the value of possessing the ball before the pass was given is not taken into account, solely the expected outcome of the attack is used to give the pass a value. We expect that in the PPV approach the crosses into the penalty box are not valued extremely high anymore. We also expect that defenders who start an attack will get more credits for the expected contribution of their passes. The SPV approach is even more detailed as it does take into account the complete sequence up to the pass was given. We expect this approach to do better in finding the good passers, who do not really

stand out in assists, but whose passes are valuable.

6 Experimental evaluation

In this chapter we evaluate the results of the three approaches for the 2015/2016 season using several different evaluation criteria. We first introduce the evaluation criteria and then evaluate each of the approaches on these criteria.

6.1 Player ECOM metric

The three approaches generate values for all passes in the 2015/2016 season. This allows us to value players based on the values of their passes. We rank the players according to their passes' total Expected Contribution to the Outcome of the Match (ECOM). The player ECOM metric measures the expected influence a player's expected influence on the outcome of a match. We compute this metric as follows.

$$ECOM = \frac{\text{sum of the values of the player's passes}}{\text{number of minutes played}} \cdot 90$$

6.2 Evaluation criteria

We evaluate the approaches on their rating of the players' ECOM in the 2015/2016 season. The three evaluation criteria are:

- The correlation between the player ECOM value and the FIFA 16 passing and vision skills
- The importance of the player ECOM value for estimating player market values
- The suitability of the player ECOM value to predict the outcome of football matches

6.2.1 Correlation FIFA 16 passing and vision skills

The FIFA 16 videogame contains a large database of players and their skills. Two of these skills are interesting for this research: the passing skill and the vision skill. The passing skill represents the player's ability in giving accurate passes. The vision skill represents the player's ability in having a good overview of the pitch and therefore finding the best options to pass the ball to. The passing and vision skills of players are determined by a combination of data and the opinion of football experts (Saed (2016)). Therefore it is interesting to compare these FIFA 16 skills with the player ECOM values of the three approaches proposed in this thesis. We desire the player ECOM values to be strongly correlated with the passing and vision skills of FIFA 16. As the ECOM metric captures the contribution of a player's passes to the teams attack, the metric is best comparable to FIFA's vision skill. Therefore, we prefer a strong correlation between the player ECOM values and FIFA's vision skills to a strong correlation between the player ECOM values and FIFA's passing skills.

6.2.2 Estimating market values

Another way of evaluating the player ECOM values is by evaluating the ability of the player ECOM values to estimate player market values. Twice a year the transfer window opens and clubs can buy players from other clubs. The prices they pay for players strongly depend on the skills of the players, but are also influenced by their age, their position on the pitch, the number of matches they have played, their contract length and many more features. We expect the player ECOM values to be positively correlated with the player market values.

We learn a random forest model and a linear regression model to estimate the market values of the players. Random forest is a machine learning technique that uses multiple decision trees to predict the value or classification of the dependent variable (Breiman (2001)). In random forest prediction for a number of random selected (with replacement) training sets a decision tree is fit and by averaging the predictions of all those individual trees a prediction for the dependent variable is done. Furthermore, random forest can be used to measure the importance of each of the variables. This is interesting in this research as it can provide insights into which of the pass valuing models is most important in estimating player market values. Therefore, a random forest model is fit on various variables including the player ECOM values generated by the different pass value approaches.

The Gini importance of a variable can be used as a measure of feature relevance (Menze et al. (2009)). By evaluating the Gini importance of the player ECOM values it is evaluated to what extent the player ECOM values are capable of estimating player market values. The higher the Gini importance, the more important the variable is for the market value. We prefer the pass valuing approach whose player ECOM metric that has the highest Gini importance.

A total of six linear regression models is run to estimate the market values of the players. In these six models the set of independent variables consists of a number of fixed variables plus an extra variable that is different for each of the regressions. These six extra variables are the player's number of assists per 90 minutes, the player's ECOM generated by the ZPV - 15 seconds rule approach, the player's ECOM generated by the ZPV - expected goals approach, the player's ECOM generated by the PPV approach, the player's ECOM generated by the SPV - DTW approach and the player's ECOM generated by the SPV - Fréchet approach. We compare these six regression models on their R^2 , mean squared error (MSE) and mean average error (MAE) to find the one that best estimates the player market values. Next to that, we compare the coefficients and level of significance of the six extra variables. We prefer the pass valuing approach whose player ECOM metric that has the highest coefficient and thus also has the highest influence on the prediction.

6.2.3 Predicting football match outcomes

The last evaluation criteria is the ability of the player ECOM values to predict the outcome of football matches.

Maher (1982) introduced a model that uses two independent Poisson distributions to model the number of goals scored by the home and away team. Let us consider match k between team i and team j and let X_k be the number of goals scored by

team i and let Y_k be the number of goals scored by team j . Like Maher (1982) we assume that $X_k \sim \text{Poisson}(\mu_i)$ and $Y_k \sim \text{Poisson}(\mu_j)$. Maher (1982) determined the means of the Poisson distributions by taking into account the number of goals scored and conceded by the teams in previous matches. We use the player ECOM values of the 2015/2016 values to determine μ_i and μ_j for each match of the 2016/2017 season. The means of the Poisson distribution are determined by summing up the player ECOM values for the players for team i and j that start match k . The line-ups for the teams are known for each match and this makes it possible to determine μ_i and μ_j . The substitutions are not taken into account, as before the match starts only the line-up is known and we want to predict the outcome before the match starts. Some players did not play matches in our data set in the 2015/2016 season and therefore do not have an ECOM value. For these players the average ECOM value of their team is used. We do not make predictions for matches in which one of the teams did not play in the highest league of their country in the 2015/2016 season. For those teams we have too little data to determine the player ECOM values.

The goal is to determine, for each match of the 2016/2017 season, a probability distribution over the possible outcomes: home win (p_h), draw (p_d) and away win (p_a). The Skellam distribution (Skellam (1946)) is the probability distribution of the difference between two independent Poisson distributed random variables. We use the Skellam distribution to determine, for each match k , the probabilities $p_h(k)$, $p_d(k)$ and $p_a(k)$.

The predictions are evaluated on the logarithmic loss metric (Kaggle (2017)). The logarithmic loss is defined as follows:

$$\text{logloss} = -\frac{1}{N} \sum_{k=1}^N h(k) * p_h(k) + d(k) * p_d(k) + a(k) * p_a(k)$$

where $h(k) = 1$ if match k ends in a home win and zero otherwise, $d(k) = 1$ if match k ends in a draw and zero otherwise, and $a(k) = 1$ if match k ends in an away win and zero otherwise. N is the number of matches that are predicted, in this case $N = 1686$. The lower the logarithmic loss, the more accurate the predictions are. Therefore, we prefer the pass valuing approach that yields the best predictions, and thus has the lowest logarithmic loss value. The predictions generated from the player ECOM values are also compared to a naive baseline.

6.3 Evaluation results

The pass valuing approaches that were introduced in this thesis are evaluated on the above-mentioned evaluation criteria. The results are presented in the following sections.

6.3.1 Correlation FIFA 16 passing and vision skills

Figure 24 shows the correlation between the player ECOM values generated by the different approaches and the passing and vision skills of the FIFA 16 game.

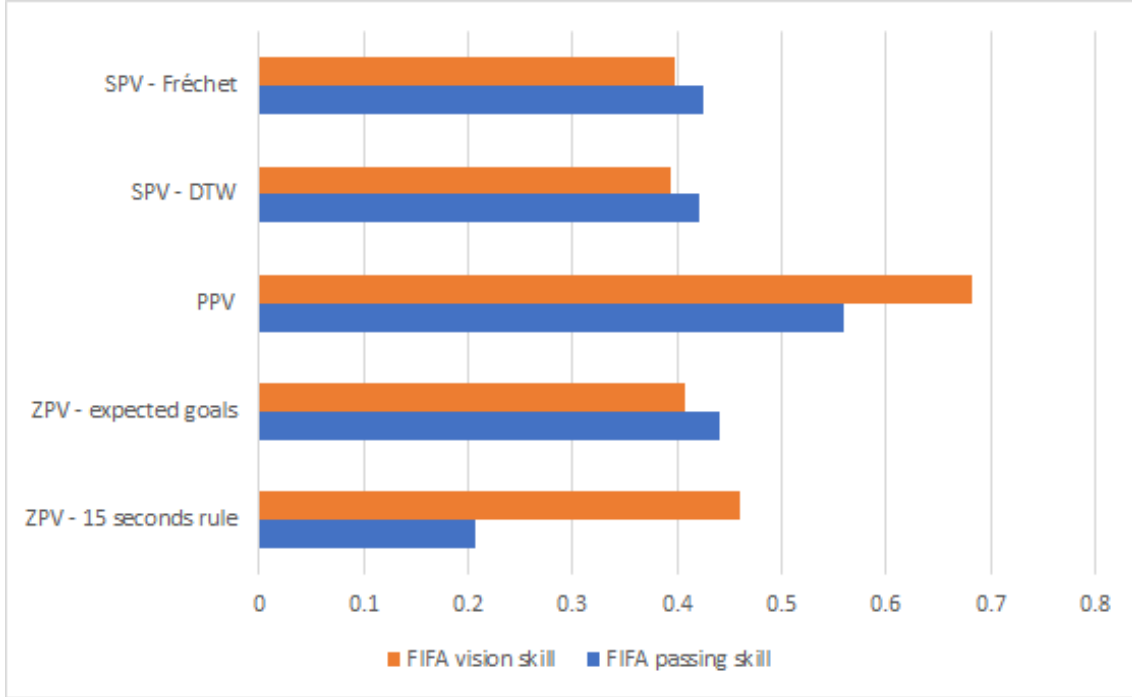


Figure 24: The correlation for both the FIFA passing and vision skills is the highest for PPV’s player ECOM values. The correlation between PPV’s player ECOM values and FIFA’s passing skills is 0.56 and the correlation between PPV’s player ECOM values and FIFA’s vision skills is 0.68.

Figure 24 shows that the PPV approach scores best on this evaluation criterion as it has the strongest correlation with both FIFA’s vision skill and FIFA’s passing skill.

6.3.2 Estimating market values

We compose a data set consisting of the following variables that we consider important for the market values of players:

- Number of goals per 90 minutes in the last 12 months
- Number of non-penalty goals per 90 minutes in the last 12 months
- Number of assists per 90 minutes in the last 12 months
- ELO score (ranking of the strength of a team) of the player’s team
- Number of minutes played in the last 12 months
- Age in years
- Total number of matches played for the national team (caps)
- Remaining length of contract in years

Both random forest and linear regression models are used to measure the importance of the player ECOM values generated by the various pass valuing approaches in estimating the market values.

6.3.2.1 Random forest

The variables and the player ECOM values of the different pass valuing approaches are put in a random forest model with 1000 trees. Figure 25 shows the average Gini importance of the variables in the model.

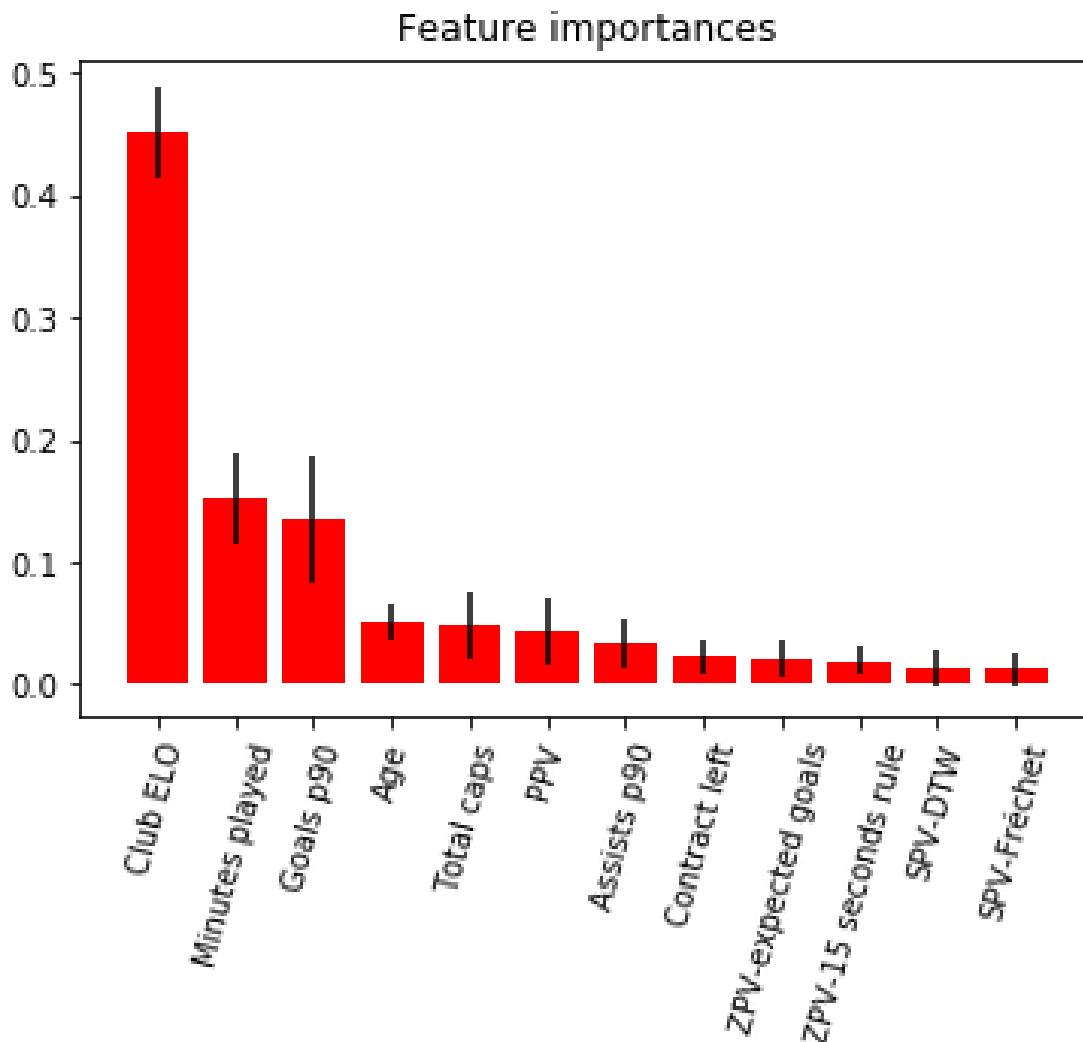


Figure 25: The average Gini importance for the features in the random forest model with 1000 trees. The number of features to consider when looking for the best split is the default setting: the square root of the number of features, so three in this case. Especially PPV's ECOM has a high Gini importance, when compared to the other ECOM values.

6.3.2.2 Linear regression

The dependent variable of the linear regression is the natural log of the player market value. To start with, the variables mentioned above are tested on their significant influence on the player market value. For each of the variables, we also test whether its logarithmic or squared counterpart has a significant influence on a player's market value. Using backward elimination, we select the following fixed variables: ELO score and squared ELO score, number of minutes played in the last 12 months, age

and squared age, remaining length of contract and its squared counterpart, number of goals per 90 minutes and the total number of caps for the national squad.

Next, the six linear regression models with the six different extra variables are run. For each of these six linear regression models it holds that all independent variables are significant, including the extra variables. The independent variables were normalized in order to fairly compare the coefficients of the variables. Table 6 shows for each regression model the coefficient of the extra variable, the model R^2 , the model MSE and the model MAE. The PPV approach generates the player ECOM values with the highest coefficient in the linear regression model when compared to the other five extra variables. As this regression model also generates the best predictions in terms of the highest R^2 and lowest MSE and MAE, the PPV approach scores best on this evaluation criterion.

	Coefficient extra variable	R^2	MSE	MAE
Number of assists per 90 minutes	0.6680	0.7309	0.3722	0.4840
ZPV - 15 seconds rule	0.9522	0.7326	0.3698	0.4807
ZPV - expected goals	0.4344	0.7287	0.3752	0.4859
PPV	1.2595	0.7437	0.3545	0.4727
SPV - DTW	0.4651	0.7281	0.3760	0.4869
SPV - Fréchet	0.4932	0.7282	0.3758	0.4868

Table 6: This table shows the results of the six linear regression models. The model including the PPV player ECOM values has the highest R^2 -value and the lowest mean squared error and mean average error. Furthermore, the coefficient of the PPV player ECOM is the highest of all extra variables.

6.3.3 Predicting football match outcomes

We predict the outcomes for 1686 matches of the 2016/2017 season are predicted using the player ECOM values as was explained before. As a baseline a naive prediction is used. The naive prediction method sets the probability of a home win, a draw and an away win to the proportion of home wins, draws and away wins observed during the 2012/2013 through 2015/2016 seasons. These probabilities are 45.38%, 25.25% and 29.37%, respectively.

Table 7 shows the logarithmic loss values for each of the predictions. The predictions based on the player ECOM values generated by the PPV approach are the most accurate.

Prediction method	Logarithmic loss
PPV	1.0377
Baseline	1.0498
ZPV - 15 seconds rule	1.0819
ZPV - expected goals	1.0868
SPV - Fréchet	1.1893
SPV - DTW	1.1931

Table 7: Logarithmic loss values of the different predictions. Only the predictions with the PPV player ECOM values beat the baseline.

6.4 Conclusion evaluation

We evaluated the three pass valuing approaches, the ZPV, PPV and SPV, on three evaluation criteria. The first criterion is the correlation with the FIFA passing and vision skills. It turns out that the PPV player ECOM values have the strongest correlation with the FIFA passing and vision skills. Also on the second criterion, the estimation of player market values, the PPV approach scores best. It turns out that the PPV player ECOM values are amongst the most important features of the random forest model to estimate market values. The third criterion is the capability of the player ECOM values to predict the outcomes of football matches. Again, the PPV player ECOM values score best. The model based on the PPV player ECOM values is the only one that beats the baseline with respect to the logarithmic loss. To conclude, the PPV approach scores best on all three evaluation criteria. Therefore, we consider the PPV approach the best approach and show more results for this approach in the next chapter.

7 Results PPV approach

The goal of this thesis was to build a model that assigns values to passes in football. Three approaches were introduced, executed and evaluated. The PPV approach turned out to score best on three pre-determined evaluation criteria. We present the results of this approach for the 2016/2017 season and answer the following questions in this chapter:

1. What do the high-valued passes look like?
2. Who are the top-ranked players based on their ECOM value in the 2016/2017 season?
3. What are the top-ranked teams based on their players' ECOM values in the 2016/2017 season?

7.1 Q1: What do the high-valued passes look like?

To get an idea of which passes are valued high by the PPV approach, we observe the three highest rated passes. The best pass in the PPV approach is the long ball by Lorient's goalkeeper Lecomte directly after catching a ball in the attack of the opponent. With this ball from his own box, he reaches a teammate just 30 metres away from the opponent's goal while all opponents are still in attacking position. Unfortunately his teammate loses the ball when he tries to dribble pass the only opponent's defender left. The second best pass is a free kick from around the centerline by Szymanowski (Légnés) giving his teammate a great opportunity to score (he fails). The third best pass of the PPV approach is a header by Weigl (Borussia Dortmund) that lengthens a free kick and gives his teammate the opportunity to score. For all of these passes it seems legit that they are valued high.

Figures 26 and 27 show heatmaps of the average value passes get when they originate or destinate in a certain zone of the pitch. The passes ending near the opponent's goal get on average the highest values. This is expected as on average these passes more often lead to goal opportunities.

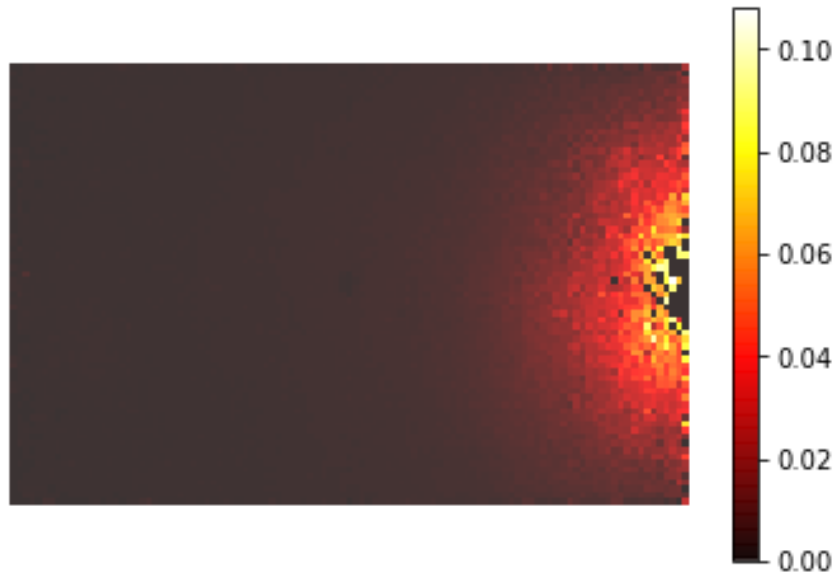


Figure 26: Heatmap of the average PPV for passes originating in the zones. The direction of attack is from left to right.

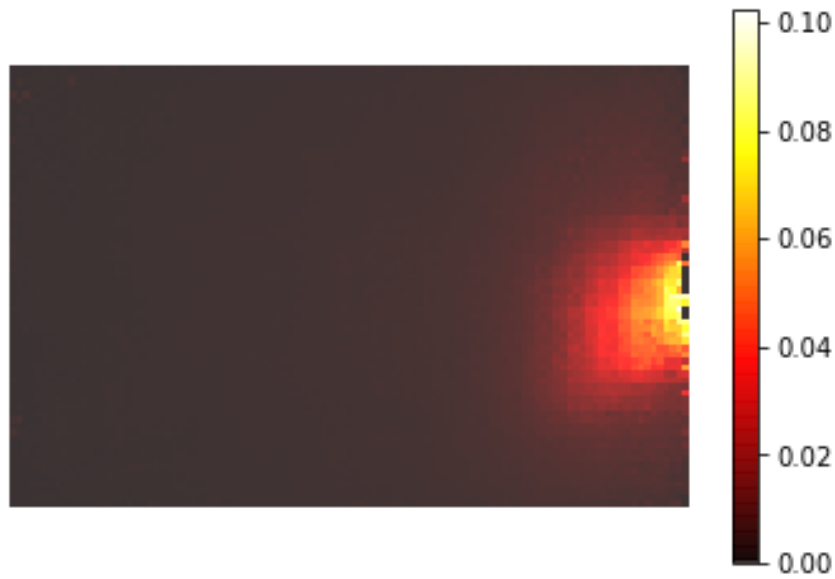


Figure 27: Heatmap of the average PPV for passes with destination in the zones. The direction of attack is from left to right.

7.2 Q2: Who are the top-ranked players based on their ECOM in the 2016/2017 season?

We analyze the player ECOM values of the 2016/2017 season and we only consider the players in the data set that played at least 450 minutes in the 2016/2017 season. Figure 28 shows the top 15 players with respect to their ECOM in the 2016/2017

season. This figure also shows the growth in ECOM the player experienced between the 2015/2016 and 2016/2017 seasons.




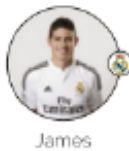



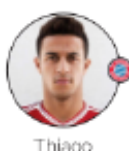







Rank	Player	Score	Rank	Player	Score
1	 Fabregas	0.407 ↑ 22.7% 0.81 assists 0.33 goals	9	 Banega	0.328 ↑ 27.9% 0.46 assists 0.35 goals
2	 Ribéry	0.385 ↑ 4.5% 0.80 assists 0.36 goals	10	 James	0.317 ↓ 3.1% 0.46 assists 0.61 goals
3	 Özil	0.384 ↓ 7.3% 0.32 assists 0.25 goals	11	 Coutinho	0.312 ↑ 31.3% 0.28 assists 0.52 goals
4	 Silva	0.372 ↑ 0.8% 0.26 assists 0.13 goals	12	 Thiago	0.305 ↑ 4.8% 0.20 assists 0.24 goals
5	 De Bruyne	0.352 ↑ 28.5% 0.66 assists 0.19 goals	13	 Hazard	0.301 ↑ 0.7% 0.15 assists 0.48 goals
6	 Cazorla	0.345 ↓ 1.5% 0.29 assists 0.29 goals	14	 Kroos	0.294 ↑ 11.9% 0.43 assists 0.11 goals
7	 Neymar	0.338 ↓ 4.3% 0.51 assists 0.44 goals	15	 Verratti	0.294 ↓ 9.4% 0.21 assists 0.13 goals
8	 Messi	0.338 ↓ 6.0% 0.38 assists 1.18 goals			

Figure 28: Top 15 players with the highest ECOM value in the 2016/2017 season. The player's ECOM value, the growth between the 2015/2016 and 2016/2017 season and the player's number of assists and goals per 90 minutes are presented.

All players in this top 15 are considered to be creative players and it looks like the

PPV approach is able to find players with a high level of creativity. The ECOM score of a player represents its expected contribution to the outcome of the match. Cesc Fàbregas tops the ranking with an ECOM value of 0.407 and therefore it can be said that Fàbregas' passes are expected to contribute to 0.407 goals per 90 minutes.

From the ranking of all players based on their ECOM we form a line-up of players with the highest ECOM per 90 minutes. We form this line-up by selecting the player with the highest ECOM for each position. Figure 29 shows this line-up.



Figure 29: A line-up of the players with the highest ECOM value per position. This line-up is expected to score on average 3.046 goals per 90 minutes.

In the line-up it can be observed that midfielders and strikers have a higher ECOM than defenders and goalkeepers. This can be supported by the fact that defenders and goalkeepers contribute less to the attack of the team as their primary task is to defend the own goal. Figure 30 shows how the ECOM values of defenders, midfielders and strikers are distributed.

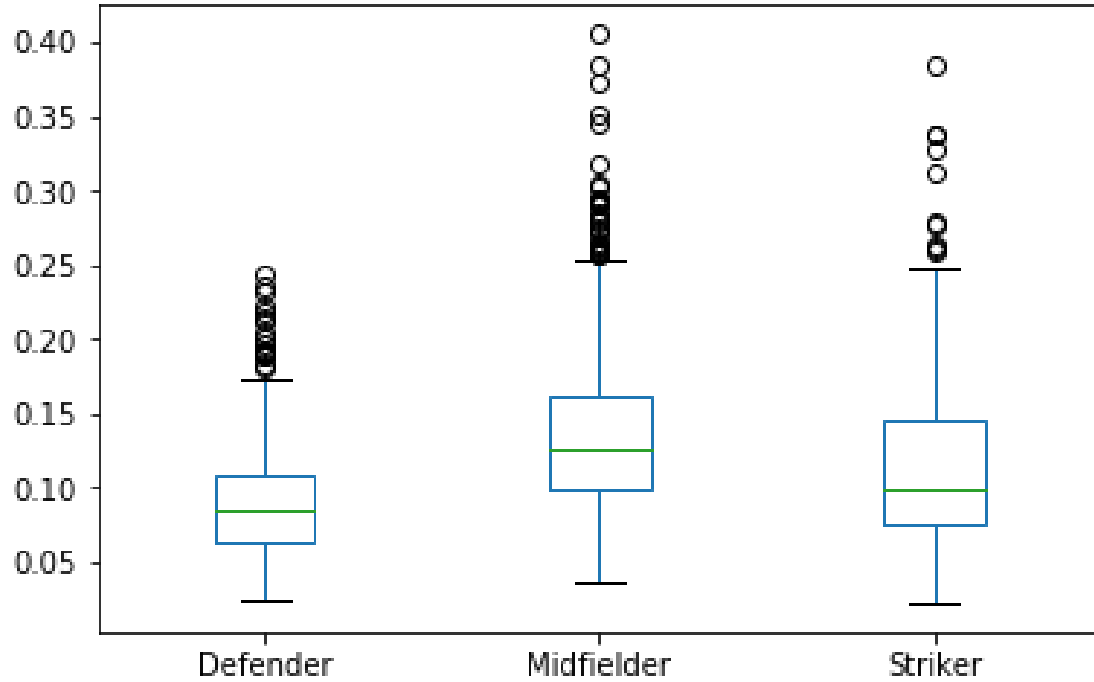


Figure 30: A boxplot of the ECOM values per position. The y-axis represents the ECOM player values. Defenders have the lowest median ECOM value and there is also less variation between defenders' ECOM values than between midfielders or strikers. Midfielders have the highest median ECOM value and there is a lot of variation between the ECOM values of midfielders.

7.3 Q3: What are the top-ranked teams based on their players' ECOM values in the 2016/2017 season?

As all players are assigned an ECOM value for each of the approaches, it is also possible to calculate the average ECOM value for all players in a team. Table 8 shows PPV's op 10 teams for the 2016/2017 season and the growth in average ECOM between the 2015/2016 and 2016/2017 seasons.

Rank	Team name	Average player ECOM 2016/2017 (Growth compared to 2015/2016 season)
1	Bayern München	0.2025 (+4.3%)
2	Manchester City	0.1943 (+28.2%)
3	Arsenal	0.1820 (+2.6%)
4	Barcelona	0.1770 (-8.9%)
5	Napoli	0.1753 (+10.3%)
6	Liverpool	0.1719 (+20.7%)
7	Chelsea	0.1686 (+14.1%)
8	Manchester United	0.1590 (+4.9%)
9	PSG	0.1579 (-8.6%)
10	Real Madrid	0.1552 (-7.3%)

Table 8: Teams with the highest average player ECOM for the PPV approach for the 2016/2017 season

The fact that Manchester City belongs to the biggest climbers could be explained by the manager switch from the more defensive Pellegrini to Josep Guardiola who is known for letting his teams play on ball possession. Internazionale changed from a defensive style to a more attacking style and had multiple changes of coaches. Also the influence of Jürgen Klopp on Liverpool can be observed in this table. The PSG switch from Laurent Blanc to Unai Emery did not have a positive influence on the average player ECOM of PSG. This is also seen in their ranking in the French Ligue 1 as they did not top the league in the 2016/2017 season, while they did the previous four seasons. In the 2015/2016 season Arsenal was in the race for the title for a long time and ended on the second place in the Premier League, whereas in the previous season they ended on the fifth place.

8 Conclusion

This chapter contains a summary of the research, answers to the research questions, a discussion on the research and possible extensions and future work.

8.1 Summary

This thesis presents three approaches of valuing passes and computes them on event data for five seasons of the top five leagues of Europe.

The first approach, the zone-oriented pass value (ZPV) approach, relies on a possession value for all zones of the pitch. We tested two different ways of valuing the zones, one taking into account the probability that a goal is scored within 15 seconds, and one that represents the probability that a goal is scored when shooting from that range. The ZPV approach is already an improvement to valuing the passes as 1 when they find a teammate and as 0 when they do not find a teammate.

The second approach, the pass-oriented pass value (PPV) approach, relies on a newly introduced distance measure for passes that is a combination of the difference in the origin and destination location of the passes, the difference in length of the passes and the difference in location of the ball five and 15 seconds preceding the passes. We use the k-nearest neighbors algorithm to value the passes as the average outcome of the possession sequences of their neighbor passes. We show that the PPV approach gives a better valuation of the passes, the passing skills of the players and the passing skills of the teams than the other approaches.

The third approach, the sequence-oriented pass value (SPV) approach, relies on the comparison of the possession sequences of the passes. The similarity of the possession sequences is calculated using two different distance measures: the dynamic time warping distance measure and the Fréchet distance measure. It was also tested whether the longest common subsequence distance measure was applicable to football possession sequence, but it was concluded that this was not the case. The results of this approach were promising, however it seemed to not be able to rank the players in the way the PPV approach was able to. This may be caused by the fact that when looking at the complete sequence, the more relevant parts of the sequence (closer to the present) are undervalued and the less relevant parts of the sequence (the beginning of the sequence) are overvalued when looking at a specific pass. The comparison of the sequences also yields the opportunity to compare teams on their possession sequences. The teams of the 2016/2017 were clustered and this resulted in some interesting clusters.

8.2 Answering the research questions

The main research question was to investigate to what extent machine learning techniques are capable of determining which players in Europe's top five leagues perform the passes that contribute most to the outcomes of their teams' matches. We showed that the PPV approach scores best on the three pre-determined evaluation criteria

and therefore we analyzed the results of the PPV approach in the Results chapter. Figure 28 on page 49 shows the top 15 players of the 2016/2017 season with respect to their ECOM value resulting from the PPV approach. Cesc Fàbregas is the player with the highest expected contribution to the outcome of the match in the 2016/2017 season. The top five is completed by Ribéry, Özil, Silva and De Bruyne. These players all play for big clubs in Europe and are considered good passers and as one of the most influential players of their teams. These players strike out with their number of assists and are also known for the risk in their play. It can thus be concluded that machine learning approaches can be used to value passes and to rank players on their passing skills. However, there is still room for improvement.

The second research question was to investigate how we can best measure the similarity of possession sequences in terms of their spatial similarities. In this thesis a new distance measure for passes was introduced and the results of this approach were promising. When tracking data is added to the model, this makes it possible to improve the distance measure even more by looking at the positions of the rest of the players.

The third research question was to explore whether we could use our pass valuation model to detect playing styles. This research question was answered in the Applications chapter where a clustering was performed on the teams, by taking the similarities of their possession sequences into account. Table 4 on page 39 shows the six clusters resulting from the k-means clustering. It is notable that teams from the same leagues are often in the same cluster, except for the top teams of the leagues which were all together in a cluster.

The fourth research question as stated in the introduction was to investigate to what extent the valuing of players' passing skills could help to estimate their market values. Section 6.3.2 shows that the players' ECOM values generated by the PPV approach were more important in the random forest model than the average number of assists per match. The valuing of players' passing skills can thus help to estimate their market values.

8.3 Discussion

Although the proposed approaches generate promising results, they all have their limitations and there is definitely room for improvement. The ZPV approach only looks at the location of the pass and ignores all other characteristics of the pass. When the 15 seconds rule is used for the valuation of the zones, the own box is valued too high, whereas the expected goals model values areas near the goal but with a small angle to the goal too low. The PPV approach does not take into account the value of the location before the pass was given, so when a player is already in a very promising position a relatively easy pass may be valued very high. Lastly, the SPV approach looks at the complete possession sequence up until the pass, whereas the beginning of the sequence might not be worth it to look at anymore.

Furthermore, all passes are valued in the same way in this thesis. There is no distinction between passes by head or by foot, or between low passes and crosses. The

model can be improved by making a distinction between these different types of passes. All approaches rely on the simple expected goals model that was explained in the Football Analytics chapter. This expected goals model is very simple and ignores a lot of features that may influence the probability of a goal being scored from a goal attempt. By improving this model, the pass values generated by the approaches can be improved as well.

The lack of data on the positions of all players on the pitch when a pass is executed makes it impossible to know whether a certain pass was hard to execute or not. This is thus not taken as an input in the model, while this is a good indicator of the passing skill of a player. This is another limitation of the models. Another limitation is that the vision of the coach and the style of play influence the decisions players make. Some coaches might tell their players to shoot from range, where others might tell them to continue passing until they are in the penalty area and have created a big possibility of scoring. The latter might cause players to try to give risky passes, which causes them to lose the ball more often. The models ignore the tactics of the teams. The model also ignores the accuracy of the passes, as it is not known whether the pass was perfect in the feet of the receiver or whether the receiver had a hard job controlling the ball.

The PPV and SPV approaches both use 100 nearest neighbors to value the passes. Although some neighbors are closer than others they all are weighted equally when valuing the passes. It might be better to use weights for the neighbors, such that the nearest neighbors are weighted more than the neighbors that are further away. In addition, we could learn the optimal number of neighbors from the data.

The passes of all leagues of all seasons are lumped all together. However, the playing styles of the leagues differ. Also some leagues are stronger than others and therefore it might be unfair to compare players across the leagues. The playing styles and strengths of the leagues differ, but also the manual annotation of the data might be done in a different way. This may influence the results as well.

8.4 Future work

The proposed pass valuing approaches presented in this thesis give promising results, but they could be improved in various ways. As explained in the previous section, the lack of data on the position of all players is a major limitation of this research. In future work, the positional tracking data could be added to the model such that the passing accuracy, passing precision and decision-making of the players can be determined. The pressure the player undertakes when possessing the ball, the positions of his teammates and the speed of all players on the pitch can now be taken into account when valuing the passes of players. A similar approach as Cervone et al. (2014), who describe an expected possession value for basketball that could be performed when positional data is added to the model.

Furthermore, the game of football can be viewed as a Markov game, with Markov states and their transitional probabilities, such as was done for ice hockey by Schulte et al. (2015). In contrast with this article shots and throw-ins could be added to

the absorbing states of the Markov game. For each state of the game, the expected reward can be calculated and thus the contribution of each action can be given a value. Other possible extensions include focusing on the teams' tactics. The clustering of the passing sequences makes it possible to extract the tactics of certain teams and to compare the passing tactics of different teams. When positional tracking data is added to this, further clustering of states of the game are possible and this results in better analyses of the opponent.

Another extension is predicting the choice a player makes in a certain state of the game. Le et al. (2017) introduce a ghosting method for football, which simulates match situations using historical tracking data. In the future this could lead to teams being able to test their tactics on the opponent's tactics by simulation of the match. However, this goes way beyond the scope of my thesis, so I do not focus on extracting the tactics of teams using the positional tracking data.

To improve the model the strength of the opposing team or even opposing player can be imposed in the model to better predict how 'hard' it was to make the pass. Furthermore, it is shown that the time of the match at which the pass was made is correlated to the value of the pass, which can be explained by the fact that the spaces become bigger during the end of the match. The model thus could be improved by incorporating this time-effect in the model. The results can also be used to see the performance of a player during the course of the game. Does he perform better or worse in the second half compared to the first half? How does the player perform when his team is on the winning side? And is it different to when his team is on the losing side?

A Results ZPV - 15 seconds rule approach

Robbie Brady ^{2.479}		Marco Reus ^{1.371}		
	Thomas Mangani ^{2.672}	Alejandro Gómez ^{2.962}	Cesc Fàbregas ^{2.773}	Ángel Di Maria ^{3.204}
Vincent Bessat ^{2.332}				Alessandro Florenzi ^{1.1}
	Ricardo Rodríguez ^{1.553}		Florentin Pogba ^{1.190}	
		Tom Heaton ^{0.1680}		

Table 9: Line-up for the ZPV - 15 seconds rule approach for the 2016/2017 season including the players' ECOM value per 90 minutes

Rank	Player name	Team name	Number of 'excellent' passes per 90 minutes
1	Cesc Fàbregas	Chelsea	32.09
2	Mesut Özil	Arsenal	26.61
3	Ángel Di Maria	PSG	26.12
4	Santi Cazorla	Arsenal	25.30
5	Kevin De Bruyne	Manchester City	24.93

Table 10: Players with the highest number of 'excellent' passes per 90 minutes for the 2016/2017 season for the ZPV - 15 seconds rule approach

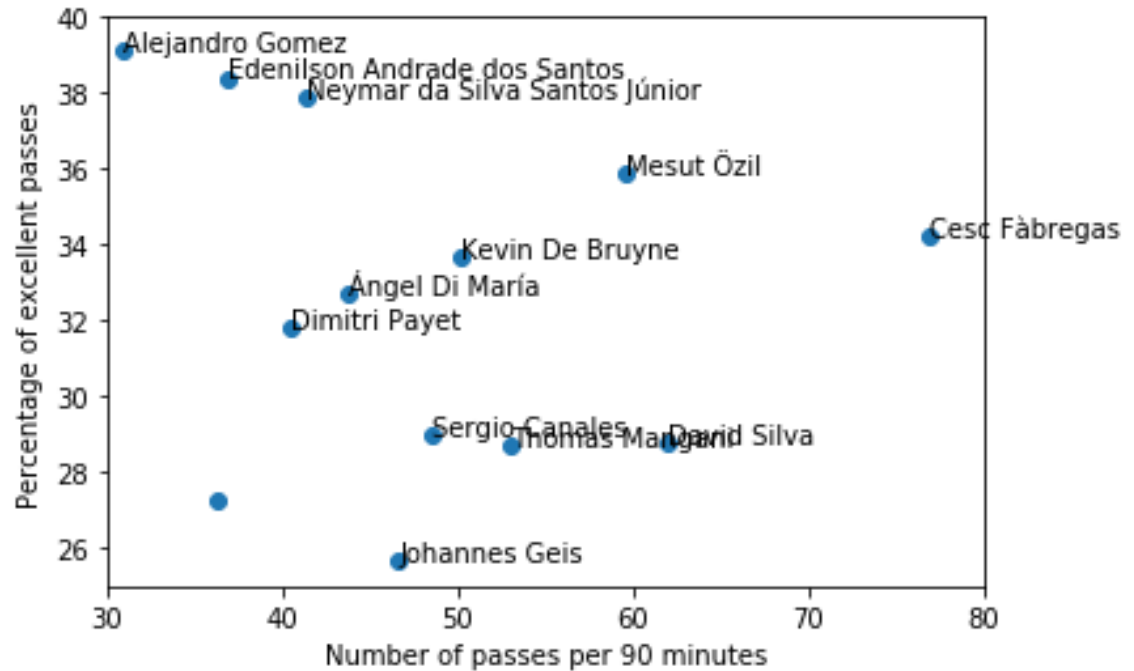


Figure 31: Scatter plot of the number of passes per 90 minutes and the percentage of excellent passes for the ZPV - 15 seconds rule approach

Rank	Player name	Age	Switch of club?	Growth
1	Oriol Riera	31	Yes	1365.35%
2	Fernando	30	No	65.82%
3	Carlos Castro	22	No	33.66%
4	Divock Origi	22	No	32.82%
5	Stefan Kießling	33	No	31.43%

Table 11: Players with the biggest growth in ECOM for the ZPV - 15 seconds rule approach

Rank	Team name	Average player ECOM
1	Internazionale	0.8788
2	Manchester City	0.8734
3	Liverpool	0.7895
4	Southampton	0.7562
5	Napoli	0.7533
6	Caen	0.7530
7	Bayern Munchen	0.7498
8	AS Roma	0.7472
9	Crystal Palace	0.7382
10	FC Ingolstadt 04	0.7343

Table 12: Best teams for ZPV - 15 seconds rule for the 2016/2017 season

B Results ZPV - expected goals approach

Robbie Brady ^{1.987}	Alexis Sánchez ^{1.098}	
	Kevin De Bruyne ^{2.327}	Ángel Di María
Faouzi Ghoulam ^{1.957}	Thomas Mangani ^{1.939}	Cesc Fàbregas ^{1.936}
	Ricardo Rodríguez ^{0.8404}	Alessandro Florenzi
	Tom Heaton ^{0.4926}	Shkodran Mustafi ^{0.5485}

Table 13: Line-up for the ZPV - expected goals approach for the 2016/2017 season including the players' ECOM value per 90 minutes

Rank	Player name	Team name	Number of 'excellent' passes per 90 minutes
1	Cesc Fàbregas	Chelsea	23.66
2	Neymar	Barcelona	20.00
3	Mesut Özil	Arsenal	19.31
4	Angel Di Maria	PSG	18.42
5	Kevin De Bruyne	Manchester City	17.85

Table 14: Players with the highest number of 'excellent' passes per 90 minutes for the 2016/2017 season for the ZPV - expected goals approach

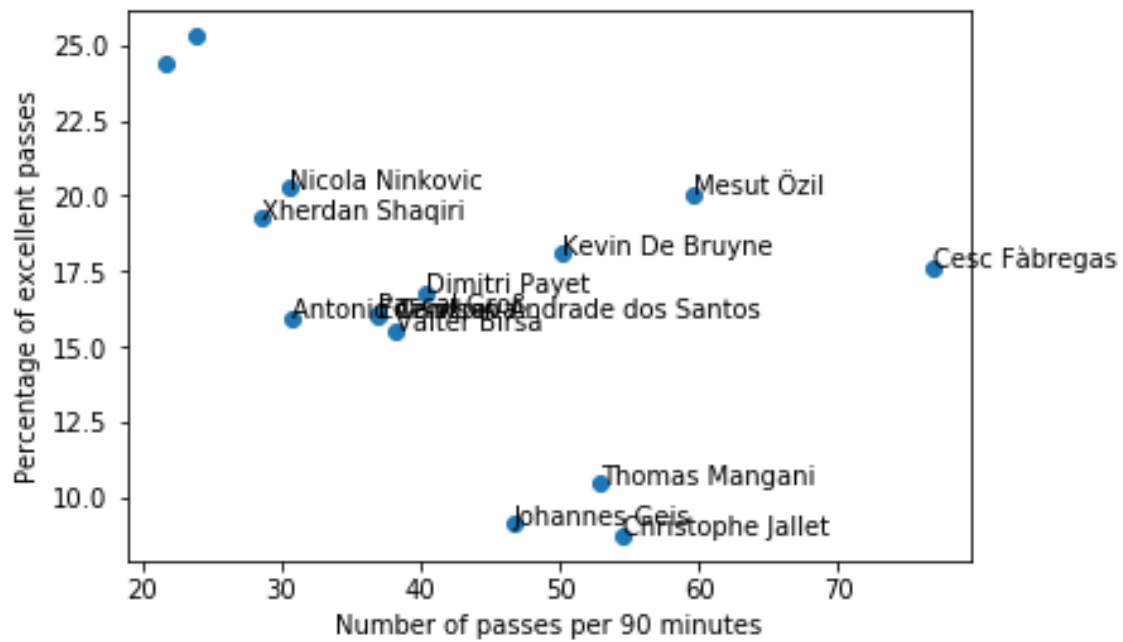


Figure 32: Scatter plot of the number of passes per 90 minutes and the percentage of excellent passes for the ZPV - expected goals approach

Rank	Player name	Age	Switch of club?	Growth
1	Andrea Belotti	23	No	340.38%
2	Juan Iturbe	24	Yes	188.21%
3	Mark Uth	26	No	146.15%
4	Jordi Amat	25	No	56.33%
5	Gonzalo Higuaín	29	Yes	48.49%

Table 15: Players with the biggest growth in ECOM for the ZPV - expected goals approach

Rank	Team name	Average player ECOM
1	Internazionale	0.6582
2	Manchester City	0.6111
3	AS Roma	0.5885
4	Liverpool	0.5666
5	Napoli	0.5476
6	Bayern Munchen	0.5463
7	Athletic Club	0.5433
8	Crystal Palace	0.5316
9	PSG	0.5292
10	Stoke City	0.5253

Table 16: Best teams for ZPV - expected goals for the 2016/2017 season

C Results PPV approach

Franck Ribéry ^{0.3846}		Alexis Sánchez ^{0.2817}		
		Mesut Özil ^{0.3861}		Lionel Messi ^{0.342}
James Milner ^{0.2344}	David Silva ^{0.3720}		Cesc Fàbregas ^{0.4085}	Dani Alves ^{0.2601}
	Javier Mascherano ^{0.1846}		Mats Hummels ^{0.1726}	
		Benjamin Lecomte ^{0.04064}		

Table 17: Line-up for the PPV approach for the 2016/2017 season including the players' ECOM value per 90 minutes

Rank	Player name	Team name	Number of 'excellent' passes per 90 minutes
1	Cesc Fàbregas	Chelsea	30.34
2	Santi Cazorla	Arsenal	29.52
3	Mesut Özil	Arsenal	27.84
4	Jorginho	Napoli	27.11
5	Thiago Alcántara	Bayern Munchen	27.07

Table 18: Players with the highest number of 'excellent' passes per 90 minutes for the 2016/2017 season for the PPV approach

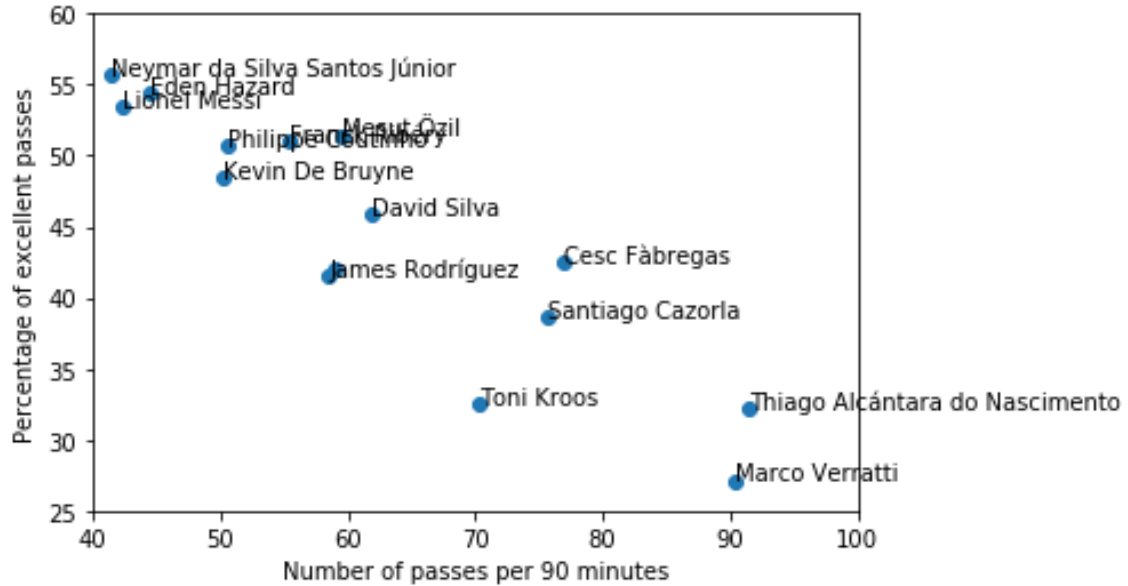


Figure 33: Scatter plot of the number of passes per 90 minutes and the percentage of excellent passes for the PPV approach

Rank	Player name	Age	Switch of club?	Growth
1	Matheus Doria Macedo	22	Yes	2.04%
2	Josef Martínez	24	No	1.84%
3	Benjamin Hübner	28	Yes	1.66%
4	Juan Iturbe	24	Yes	1.57%
5	Leroy Sané	21	Yes	1.24%

Table 19: Players with the biggest growth in ECOM for the PPV approach

Rank	Team name	Average player ECOM 2016/2017
1	Bayern München	0.2025
2	Manchester City	0.1943
3	Arsenal	0.1820
4	Barcelona	0.1770
5	Napoli	0.1753
6	Liverpool	0.1719
7	Chelsea	0.1686
8	Manchester United	0.1590
9	PSG	0.1579
10	Real Madrid	0.1552

Table 20: Best teams for the PPV approach for the 2016/2017 season

D Results SPV - DTW approach

Neymar ^{0.2657}	Alexis Sánchez ^{0.2105}
	Kevin De Bruyne ^{0.3047}
Cesc Fàbregas ^{0.3483}	Javier Pastore ^{0.2566}
Faouzi Ghoulam ^{0.2272}	Aleksandar Kolarov ^{0.1085}
Javier Mascherano ^{0.1336}	Örjan Nyland ^{0.01745}

Table 21: Line-up for the SPV - DTW approach for the 2016/2017 season

Rank	Player name	Team name	Number of 'excellent' passes per 90 minutes
1	Cesc Fàbregas	Chelsea	19.62
2	Mesut Özil	Arsenal	18.23
3	Santi Cazorla	Arsenal	18.17
4	Andrés Iniesta	Barcelona	14.67
5	Franck Ribéry	Bayern München	14.65

Table 22: Players with the highest number of 'excellent' passes per 90 minutes for the 2016/2017 season for the SPV - DTW approach

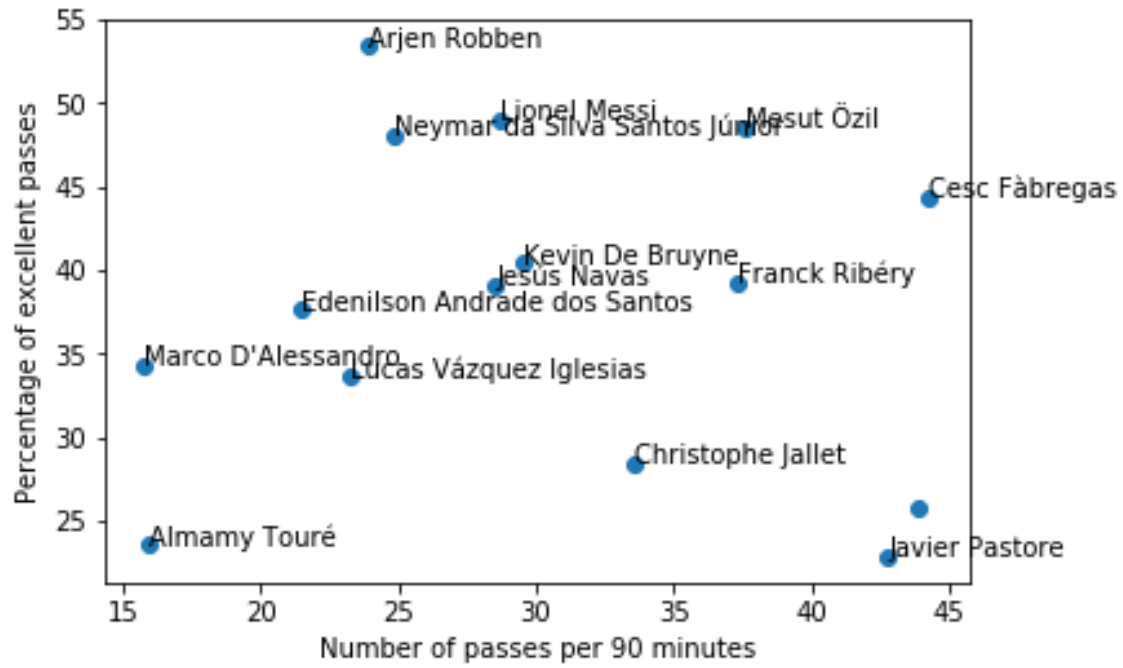


Figure 34: Scatter plot of the number of passes per 90 minutes and the percentage of excellent passes for the SPV - DTW approach

Rank	Player name	Age	Switch of club?	Growth
1	Marcus Rashford	19	No	41.86%
2	Iago Aspas	30	No	38.50%
3	Carlos Castro	22	No	32.42%
4	Gonzalo Higuaín	29	Yes	26.23%
5	Lukas Hinterseer	26	No	25.33%

Table 23: Players with the biggest growth in ECOM for the SPV - DTW approach

Rank	Team name	Mean of players' ECOM
1	Manchester City	0.106
2	Bayern München	0.105
3	Arsenal	0.105
4	Barcelona	0.104
5	Napoli	0.095
6	Real Madrid	0.092
7	Liverpool	0.089
8	PSG	0.089
9	Internazionale	0.086
10	AS Roma	0.085

Table 24: Teams with highest average player ECOM for the 2016/2017 season in the SPV - DTW approach

E Results SPV - Fréchet approach

Neymar ^{0.2668}		Alexis Sánchez ^{0.2113}		
		Kevin De Bruyne ^{0.3014}		Lucas Vázquez ^{0.294}
Faouzi Ghoulam ^{0.2322}	Cesc Fàbregas ^{0.3438}		Javier Pastore ^{0.2573}	
	Javier Mascherano ^{0.1374}		Jérôme Boateng ^{0.1122}	Dani Alves ^{0.294}
		Örjan Nyland ^{0.01669}		

Table 25: Line-up for the SPV - Fréchet approach for the 2016/2017 season

Rank	Player name	Team name	Number of 'excellent' passes per 90 minutes
1	Santi Cazorla	Arsenal	22.54
2	Mesut Özil	Arsenal	21.99
3	Cesc Fàbregas	Chelsea	21.84
4	Andrés Iniesta	Barcelona	19.47
5	Marek Hamsik	Napoli	18.80

Table 26: Players with the highest number of 'excellent' passes per 90 minutes for the 2016/2017 season for the SPV - Fréchet approach

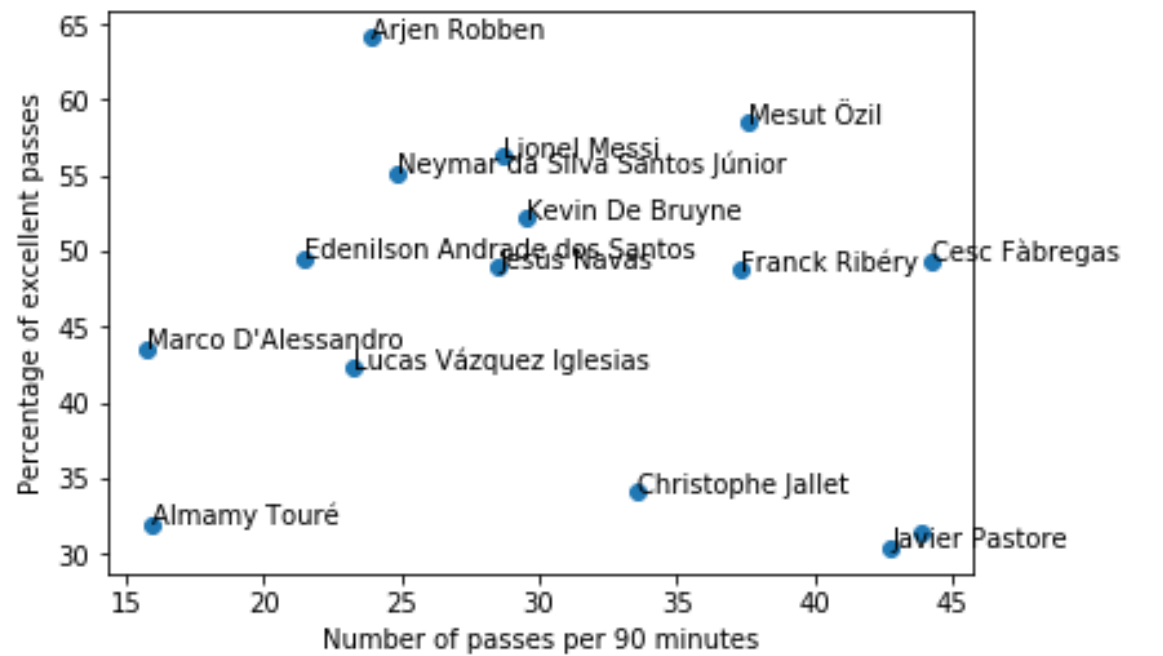


Figure 35: Scatter plot of the number of passes per 90 minutes and the percentage of excellent passes for the SPV - Fréchet approach

Rank	Player name	Age	Switch of club?	Growth
1	Andrea Belotti	23	No	378.44%
2	Carlos Castro	22	No	301.29%
3	Mario Mandzukic	31	No	78.81%
4	Iago Aspas	30	No	37.99%
5	Gonzalo Higuaín	29	Yes	34.80%

Table 27: Players with the biggest growth in ECOM for the SPV - Fréchet approach

Rank	Team name	Mean of players' ECOM
1	Manchester City	0.107
2	Bayern München	0.107
3	Arsenal	0.106
4	Barcelona	0.105
5	Napoli	0.097
6	Real Madrid	0.093
7	Liverpool	0.091
8	PSG	0.091
9	Internazionale	0.087
10	AS Roma	0.086

Table 28: Teams with highest average player ECOM for the 2016/2017 season in the SPV - Fréchet approach

F Top 5 teams

Rank	ZPV - 15 seconds rule	ZPV - expected goals	PPV	SPV - DTW	SPV - Fréchet
1	Southampton	Sevilla	Barcelona	Barcelona	Barcelona
2	Sevilla	Southampton	Bayern München	Arsenal	Arsenal
3	VFB Stuttgart	Real Madrid	Arsenal	Real Madrid	Real Madrid
4	Real Sociedad	VFB Stuttgart	PSG	PSG	PSG
5	Liverpool	Lazio	Borussia Dortmund	Napoli	Napoli

Table 29: The Top 5 teams with the highest average player ECOM for each of the approaches for the 2015/2016 season

Rank	ZPV - 15 seconds rule	ZPV - expected goals	PPV	SPV - DTW	SPV - Fréchet
1	Internazionale	Internazionale	Bayern München	Manchester City	Manchester City
2	Manchester City	Manchester City	Manchester City	Bayern München	Bayern München
3	Liverpool	AS Roma	Arsenal	Arsenal	Arsenal
4	Southampton	Liverpool	Barcelona	Barcelona	Barcelona
5	Napoli	Napoli	Napoli	Napoli	Napoli

Table 30: The Top 5 teams with the highest average player ECOM for each of the approaches for the 2016/2017 season

G High-valued passes

Rank	Who?	What?	Where?
1	Centre-back Andjelkovic (Palermo)	Pass back to the goalkeeper	From just outside own box to goalkeeper in own box
2	Centre-back Mustafi (Arsenal)	Pass back to the goalkeeper	From just outside own box to goalkeeper in own box
3	Midfielder Carroll (Tottenham Hotspurs)	Through ball	From near center line to 30 meters to the opponent's goal

Table 31: 3 highest-valued passes of the ZPV - 15 seconds rule approach

Rank	Who?	What?	Where?
1	Winger Deulofeu (AC Milan)	Pass from the back line in the opponent's box resulting in a goal	From the back line to straight in front of goal
2	Attacker Palacio (Internazionale)	Header from the back line in the opponent's box resulting in a goal	From the back line to straight in front of goal
3	Midfielder Barrera (Torino)	Pass from the back line in the opponent's box almost resulting in a goal	From the back line to straight in front of goal

Table 32: 3 highest-valued passes of the ZPV - expected goals

Rank	Who?	What?	Where?
1	Goalkeeper Lecomte (Lorient)	Long ball directly after catching the ball	From own box to 30 metres away from the opponent's goal
2	Winger Szymanowski (Léganes)	Free kick giving teammate great opportunity to score	From around the centerline near the sideline to 5 meters from the opponent's goal
3	Defender Weigl (Borussia Dortmund)	Heading ball to lengthen a free kick giving teammate opportunity to score	From corner of the opponent's box to 5 meters straight in front of goal

Table 33: 3 highest-valued passes of the PPV approach

Bibliography

Breiman, L. (2001), ‘Random forests’, *Machine learning* **45**(1), 5–32.

Caley, M. (2013), ‘Shot Matrix I: Shot Location and Expected Goals’. [Online; accessed 2-May-2017].

URL: <http://cartilagefreecaptain.sbnation.com/2013/11/13/5098186/shot-matrix-i-shot-location-and-expected-goals>

Cervone, D., DAmour, A., Bornn, L. & Goldsberry, K. (2014), POINTWISE: Predicting points and valuing decisions in real time with NBA optical tracking data, in ‘8th Annual MIT Sloan Sports Analytics Conference, February’, Vol. 28.

Cintia, P., Giannotti, F., Pappalardo, L., Pedreschi, D. & Malvaldi, M. (2015), The harsh rule of the goals: data-driven performance indicators for football teams, in ‘Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on’, IEEE, pp. 1–10.

Cintia, P., Rinzivillo, S. & Pappalardo, L. (2015), A network-based approach to evaluate the performance of football teams, in ‘Machine Learning and Data Mining for Sports Analytics Workshop, Porto, Portugal’.

Commons, W. (2016), ‘File:euclidean vs dtw.jpg — wikimedia commons, the free media repository’. [Online; accessed 21-September-2017].

URL: https://commons.wikimedia.org/w/index.php?title=File:Euclidean_vs_DTW.jpg&oldid=19

Decroos, T., Van Haaren, J., Dzyuba, V. & Davis, J. (2017), ‘STARSS: A spatio-temporal action rating system for soccer.’, *MIT Sloan sports analytics conference*.

Driemel, A. (2016), ‘Two Decades of Algorithms for the Frechet distance [Powerpoint slides]’.

URL: <http://www.win.tue.nl/~adriemel/shonan2016.pdf>

Gudmundsson, J. & Wolle, T. (2014), ‘Football analysis using spatio-temporal tools’, *Computers, Environment and Urban Systems* **47**, 16–27.

Gyarmati, L. & Anguera, X. (2015), ‘Automatic extraction of the passing strategies of soccer teams’, *arXiv preprint arXiv:1508.02171*.

- Gyarmati, L. & Stanojevic, R. (2016), ‘QPass: a Merit-based Evaluation of Soccer Passes’, *arXiv preprint arXiv:1608.03532*.
- Horton, M., Gudmundsson, J., Chawla, S. & Estephan, J. (2014), ‘Classification of passes in football matches using spatiotemporal data’, *arXiv preprint arXiv:1407.5093*.
- Kaggle (2017), ‘Multi class log loss’. [Online; accessed 28-September-2017].
URL: <https://www.kaggle.com/wiki/MultiClassLogLoss>
- Le, H. M., Carr, P., Yue, Y. & Lucey, P. (2017), Data-driven ghosting using deep imitation learning, in ‘MIT Sloan Sports Analytics Conference (SSAC)’.
- Mackay, N. (2016), ‘Measuring passing skill’. [Online; accessed 2-May-2017].
URL: <http://mackayanalytics.nl/2016/05/02/measuring-passing-skill/>
- Maher, M. J. (1982), ‘Modelling association football scores’, *Statistica Neerlandica* **36**(3), 109–118.
- Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W. & Hamprecht, F. A. (2009), ‘A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data’, *BMC bioinformatics* **10**(1), 213.
- Müller, M. (2007), *Information retrieval for music and motion*, Vol. 2, Springer.
- pena.lt/y (2015), ‘Expected Goals And Support Vector Machines’. [Online; accessed 2-May-2017].
URL: <http://pena.lt/y/2015/07/13/expected-goals-svm/>
- Power, P., Ruiz, H., Wei, X. & Lucey, P. (2017), Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data, in ‘Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining’, ACM, pp. 1605–1613.
- Rahnamai Barghi, A. (2015), Analyzing Dynamic Football Passing Network, PhD thesis, Université d’Ottawa/University of Ottawa.
- Saed, S. (2016), ‘EA explains how FIFA player ratings are calculated’. [Online; accessed 28-September-2017].
URL: <https://www.vg247.com/2016/09/27/how-ea-calculates-fifa-17-player-ratings/>
- Schulte, O., Zhao, Z. & Routley, K. (2015), ‘What is the value of an Action in Ice Hockey? Learning a Q-function for the NHL.’.
- SciSports (2016), ‘Expected Goals Model 2.0’. [Online; accessed 2-May-2017].
URL: <http://www.scisports.com/news/2016/expected-goals-model-2-0>
- Skellam, J. G. (1946), ‘The frequency distribution of the difference between two poisson variates belonging to different populations.’, *Journal of the Royal Statistical Society. Series A (General)* **109**(Pt 3), 296.

Soccermetrics (2017), ‘Everybody else is doing it, so why can’t we? Soccermetrics’ foray into expected goals’. [Online; accessed 2-May-2017].

URL: <http://www.soccermetrics.net/goalscoring-models/soccermetrics-foray-into-expected-goals>

Taki, T. & Hasegawa, J.-i. (2000), Visualization of dominant region in team games and its application to teamwork analysis, *in* ‘Computer Graphics International, 2000. Proceedings’, IEEE, pp. 227–235.

Vlachos, M., Kollios, G. & Gunopulos, D. (2002), Discovering similar multidimensional trajectories, *in* ‘Data Engineering, 2002. Proceedings. 18th International Conference on’, IEEE, pp. 673–684.

Zeinalipour, D. (2007), ‘Distributed Spatio-Temporal Similarity Search[Powerpoint slides]’.