Universiteit Leiden
The Netherlands

Opleiding Informatica & Economie

Studying player interactions and correlations during ball possession in soccer.

Martijn Vlak

Supervisors:
Mitra Baratchi & Arie-Willem de Leeuw

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

29/08/2018

Abstract

Recent technical innovations made it possible to obtain and analyze spatio-temporal data in soccer. In this thesis, we analyze position data of the WEURO 2017 tournament, by using the tool `MoveMine2.0`. Using this tool, we look at two different interaction types between players: the attraction/avoidance relationship and the following relationship. First, we explain how we applied the algorithms of `MoveMine2.0` and why we chose certain parameters. Then, we explain and describe both relationship types and give an overview of all the relationships found in the data. Finally, we look at how the two relationship types correlate between ball possessions. To do this, we compare the relationships found between players of whom their team has ball possession, with the relationships found between players of whom their team does not have ball possession. Furthermore, we compare average scores by performing a t-test. We found that there were more significant attraction- and avoidance relationships when teams do not have ball possession. For periods of 20-30 seconds, there were significantly more attraction- than avoidance relationships in ball possession, compared to not in ball possession ($p \leq 0.05$). For periods of 30+ seconds we could not find such results. Considering the following relationships, we found that there were more following pairs, when teams did not have ball possession. Furthermore, we saw that players in such following pairs, follow each other for a longer time during periods, when they do not have ball possession. However, the results were not significant: for periods of 20-30 seconds $p = 0.106$ and for 30+ seconds $p = 0.069$.

Acknowledgements

I would like to thank my supervisors Mitra Baratchi and Arie-Willem de Leeuw from the Leiden Institute of Advanced Computer Science(LIACS).

Contents

Chapter 1

Introduction

The phenomenon of using data analysis in the sport became known to many people, thanks to the movie Moneyball, which is based on an earlier released book [1]. In this particular book, which is based on a true story, a baseball club with a very low budget becomes successful thanks to data analysis. This was one of the first examples of a successful implementation of data mining in sports. Nowadays, it is more common; a recent development in (team)sports in general, is that data mining plays an increasingly important role when it comes to the field of analyzing [2]. Examples of data mining in sports are given in the articles [3–5]. In soccer this is also the case. Where there are many researches to individual player skills, not much attention has been paid into analyzing the tactical aspect of soccer, until a couple of years ago [6].

Measuring the tactical performance of teams during matches is done by game observations. There are two different types of game observations: qualitative and quantitative observations. Qualitative observations are usually performed by experts, who (subjectively) analyze a game, and use their knowledge to determine the tactical performance of teams. Quantitative observations are performed by analyzing data that is collected during matches. Purely based on findings in the data, the tactical performance of teams is assessed. Because the technology was never advanced enough to collect all this data about player's movements, qualitative game observations were more popular. However, recent developments in the computer science made it possible to collect this kind of data; with the help of position tracking systems, the positions of 22 players, referees, and the ball can be stored. Positions consist of an x-coordinate, which is parallel to the sidelines, and y-coordinates. [6]

In this thesis we will research positional data obtained from six matches of the WEURO 2017. We will analyze trajectories of players and see if we can find certain interactions between players. Furthermore we will look at whether the fact of having ball possession or not influences those interactions. It is interesting to look at ball possession, because ball possession is often regarded as one of the key performance indicators in soccer [7]. Hence, we would like to find interactions that keep repeating themselves when teams have ball possession and compare this to the interactions when teams do not have ball possession. In this way we hope to learn more about the behavior of teams when they are in ball possession. Our research question is as follows:

What are different interactions between players and how do these correlate with ball possessions?

The goal is to see if we can find certain "standard behavior" of soccer players, by finding patterns in interactions between players. We will primarily **describe** interactions between players and give a logical **explanation** why those interactions take place. By doing this, we hope that soccer coaches can evaluate this standard behavior of players; certain interactions might be desirable, others maybe not. Furthermore, by analyzing standard behavior in ball possession, coaches can learn what their team should do in order to maintain possession.

For finding interactions in the data, we will use the tool `MoveMine2.0` [8]. One of the reasons that we use this tool, is because it is specially designed for analyzing moving object data. Furthermore, the tool supports two well-known pattern mining functionalities: *attraction and avoidance relationship mining* [9] and the *following pattern mining* [10]. It will be very interesting to see if these two relationship types can be found in the player's trajectories and if we can recognize certain patterns. Considering the attraction/avoidance relationships for example, we are interested to see whether players are more tempted to avoid each other or move towards each other in ball possession. This tells us more about the player's behavior in ball possession. The same can be done for the following relationship.

Because we will be analyzing behavior in ball possession, we should only look at the trajectories of players when one of the teams has ball possession. In this thesis we will mention how we determine when a team has ball possession. After having defined this, we will only regard the trajectories in those so-called ball possession periods. Then we will look for interactions between players in those trajectories, by using the functionalities in `MoveMine2.0`. First we will look at a distribution of all the interactions found between players, and see if we already can draw conclusions based on these findings. Then, to compare interactions in ball possession with the interactions when teams do not have ball possession, we will look at all the interactions between players belonging to the same team (*interactions within teams*).

Thesis Overview

In Chapter 2 we discuss the data and the preprocessing. Furthermore, we will explain the different functions in `MoveMine2.0`. We will mention the related work of this research in Chapter 3. In Chapter 4, we show the results and discoveries of our research. We will first look at what kind of avoidance/attraction relationships `MoveMine2.0` finds in the data. Furthermore, we will look at the influence of the input parameters the algorithm requires. In the case of the attraction/avoidance relationship, the algorithm only requires one input parameter: a distance threshold, specifying when individuals are close to each other. This leads to the following sub-questions:

1. *Which attraction/avoidance relationships do we find and how does the distance threshold influence these results?*
2. *How do the attraction/avoidance relationships correlate between ball possessions?*

With the following relationship, we will do the same as we did with the attraction/avoidance relationship; we will first give an overview of what kind of following relationships we find in the data, and we will look at the influence of some parameters. Some of those parameters are required as input for the algorithm, others

for preprocessing. Finally, we will look at how the following relationships correlate between ball possessions. Hence, our sub-questions for the following relationship are:

1. *Which following relationships do we find and how do the different parameters influence these results?*
2. *How do the following relationships correlate between ball possessions?*

Finally, we draw conclusions and mention possible future work in Chapter 5.

Chapter 2

Data & Methodology

The data we used in this research originates from the WEURO 2017 tournament (UEFA European Women's Championship). The data consists of the six matches the Dutch team played during this tournament. The data was collected with position tracking systems, which use multiple cameras to capture the player's movements. For each match, the data consists of three major categories:

- **Match sheet**: Here, general information about the match is given, like the stadium it was played in, the teams that played against each other, the date it was played at and even information like the pitch width and height. Furthermore, the players of both teams are mentioned, including certain features, like player-id, name, number, etc.
- **Events**: The event data describes all events that occurred during a match, including the time, position and player(s) involved. Events can be divided in "match events" and "ball events". Examples of match events are: goals, fouls, ball out of play, end of half, etc. Examples of ball events are: passes, receptions, shots, dribbles, crosses, etc.
- **Trajectories**: Here, position data of the ball, players, and even referees is given. The position data consists of a timestamp and an x- and y-value, that correspond to the position on the pitch. Thus, there is no information about the height of the "data-object". The positional data is gathered every tenth of a second.

Now, we will give you some properties of the data. In total over the six matches, **8840086** positions are recorded. This comes down to an average of **1473347** positions per match. If you exclude the trajectories of the referees and the ball, the total is **7575854** and the average per match is **1262642**. The total amount of ball events over all the six matches is **11748** and the average per match is **1958**. From those ball events **39.8%** are passes, **24.4%** dribbles and **1.3%** shots. The total amount of match events over all the matches is **832**. On average per match this is $\pm$ **139**. From those match events, **65.3%** of them is a ball going out of play, **17.8%** are fouls and **1.8%** are goals.

2.1 Preprocessing

During this research, we have been using the tool `MoveMine2.0` to analyze the data and find interactions between players [8]. Before we can use this tool, the data first needs to be put into the right format, so we will be able to process. `MoveMine2.0` accepts datafiles with the following columns: "individual-local-identifier", "timestamp", "location-long" and "location-lat". So, in order to make the actual dataset accepted by `MoveMine2.0`, two steps are necessary to take: the time-value in the actual dataset needs to be converted to a "datetime-object" and the x- and y-values need to be converted to GPS-coordinates. For converting the x- and y-values to coordinates, we used the code of a utility provided by the `Woods Hole Oceanographic Institution` [11]. This tool calculates GPS-coordinates based on a reference point, and x- and y-values. The reference point is the center spot of a certain football field. The x- and y-values represent how far away (in meters) one is from the reference point, both horizontally and vertically. However, the problem here is that the positions in the data represent an actual football field which will probably not be laying perfectly parallel to the horizon. Hence, if you input 30 meters as y-value, you do not want the GPS-coordinate perfectly to the north of the center spot, but you want to be able to rotate, so that you take the fact that a pitch is skewed compared to the horizon into account. This is why the utility allows a rotation value (in degrees) as input.

During this research, we used the `Arke Stadium` as a basis for converting x- and y-values to GPS-coordinates (see Figure 2.1). In this stadium the final game of the tournament was played. To make sure that there were no other stadiums with a larger pitch in the dataset, we compared the sizes of all football fields with each other: all pitches had the same width and length (105 by 68 meters). The GPS-coordinates of the center spot of the `Arke Stadium` are (52.236534,6.837838) and the rotation is 45.74°.

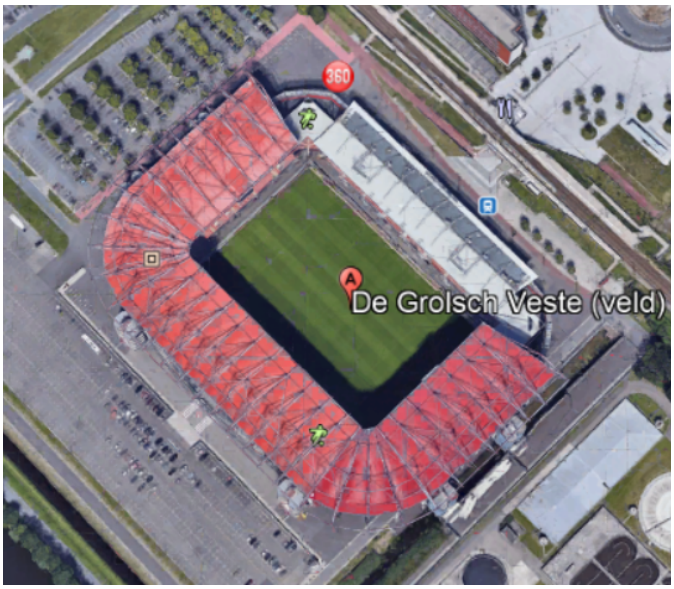

Figure 2.1: The Grolsch Veste (Arke Stadium) was used as a reference point for computing GPS-coordinates from the data.

Finally, there are two problems that arise when analyzing the football data:

1. How should we deal with teams switching sides each period?
2. How should we handle substitutions?

First, we will look at the problem that arises because teams switch sides each period. This is an issue because we want to analyze complete matches or maybe even make aggregates of multiple matches. We can not make comparisons between teams if players play at opposite positions and walk in different directions in the second half of a match. That is why we came up with the following solution: teams that play at home are always playing from left to right and away teams are always playing from right to left[1]. We accomplished this by inversing all position values of the players in the second half, by multiplying them by minus one.

Secondly, we have to deal with substitutions. When we want to compare trajectories of players, ideally, they should have the same amount of positions. If a certain player comes in two minutes before the end of a match, his/her trajectory length will be much shorter than that of players playing a whole match and it is no use comparing those trajectories with each other. We will solve this by letting the trajectory of a player that is substituted continue with the trajectory of a player that is brought in. In this way, we will have the same amount of trajectories, with the same amount of positions, per match. The downside to this is that a player that is brought in, could play on a completely different position than the player that was substituted. This leads to the fact that we can not compare individual players with each other, meaning we can only compare entire teams with each other.

[1]When we say left to right or vice versa, we mean that players move in the **latitudinal** direction; from one goal to the other.

2.2 MoveMine

Now that we know how to preprocess the data, so we can use it in `MoveMine2.0`, we will look at the different functions `MoveMine2.0` has to offer. `MoveMine2.0` incorporates four different functionalities:

1. Distance Calculation
2. Attraction and avoidance relationship mining
3. Following pattern mining
4. Plotting

Distance Calculation

The first function `MoveMine2.0` has to offer, is calculating the distance between moving objects. The function computes the pairwise distance between the trajectories of selected individuals. The function compares all pairwise points in two trajectories and computes the Euclidean distance in meters for all these points. It takes the sum of all those distance values and normalizes it by trajectory length. Next to a text file with all the distance values between players, the function also outputs a matrix. An example of this is given in Figure 2.2. This is an example of distance between players from the home playing team in a certain period of ball possession in the final. In this figure, colors in the (dark) blue spectrum indicate that players are moving close to each other, and colors more towards white indicate that players are more far away from each other. The numbers on the sides are the players of the home playing team. We can for example see that, number one (the goalkeeper) is much further away from all the players, compared to the other players.

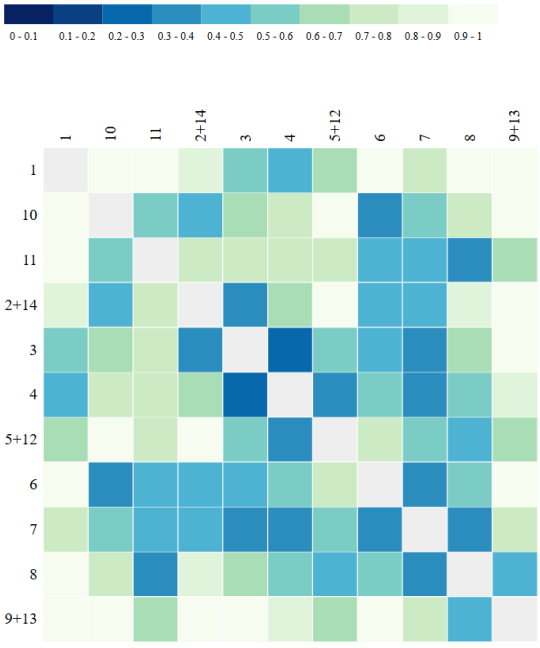

Figure 2.2: Example of a visualization of the Euclidean distance between players

The distance calculation function is useful for our research, because it can help us to understand more about the interactions between players. For example, we can look at the average distance between players who have an attraction-relationship.

Attraction and avoidance relationship mining

Secondly, we will look at the attraction/avoidance relationship function. An *attraction* relationship means that the presence of one individual causes the other one to approach. This means the distance between them will be reduced. An *avoidance* relationship is the opposite of this. When two individuals have a *neutral* relationship, they do not alter their movement based on the presence of one another. Computing the attraction/avoidance relationship works by comparing the expected meeting frequency with the actual meeting frequency, based on a distance threshold. This distance threshold specifies when two individuals are spatially close (or meeting each other). The expected meeting frequency is determined by random permutating two movement sequences.

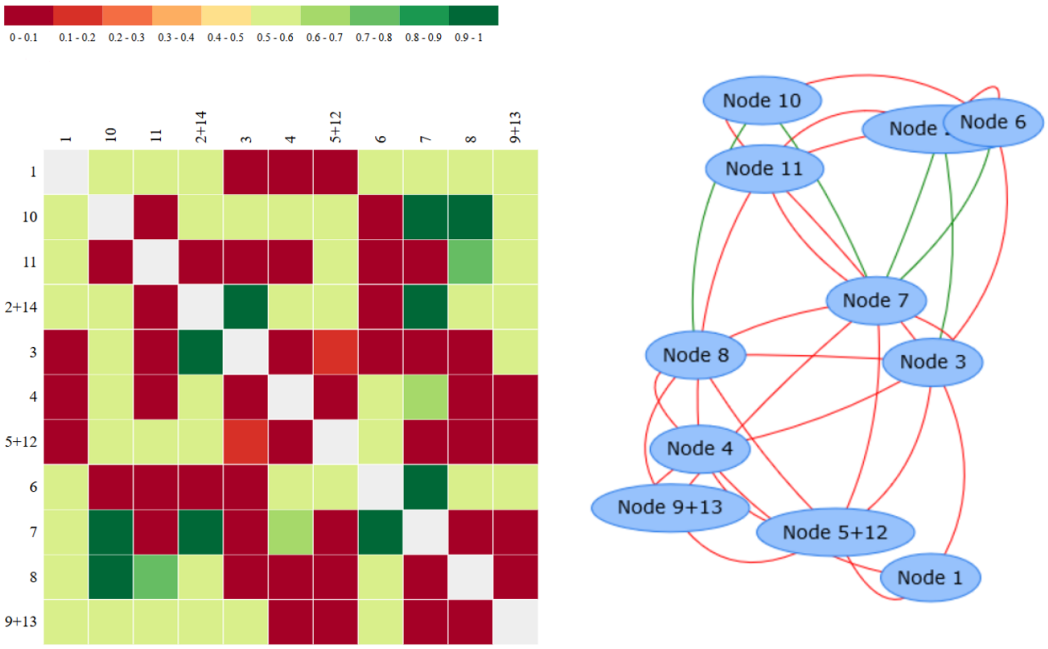

Figure 2.3: Example of a visualization of the attraction/avoidance relationship. On the left we see the relationships visualized by a matrix, and on the right the same relationships are visualized with a network graph

The idea behind the algorithm is, that if two movement sequences are independent, the meeting frequencies between two random permuted movement sequences should be similar to the actual meeting frequency. If this is not the case, the two meeting frequencies are not independent. To compute the attraction/avoidance relationship between two objects, the significance of the actual meeting frequency is tested under the independent hypothesis. By using a permutation test, the test statistic is computed. The resulting value specifies whether the actual meeting frequency is *significantly* different from the expected meeting frequency. A value of 0.5 means two individuals have a neutral relationship. Values more towards one indicate an attraction relationship, and values towards zero an avoidance relationship. By using a threshold α, one can determine from which value a relationship can be called significant. Usually $\alpha = 0.05$, thus when a relationship between two players has a value ≤ 0.025 or ≥ 0.975, then a relationship between two individuals can be called significant. [9]

Next to a text file with the values specifying the relationships between individuals, the algorithm also provides a matrix and a network as visualization. In Figure 2.3 we give an example of such a visualization. This example shows the attraction/avoidance relationships between players from the home playing team in the final match, in a certain period of ball possession. The color green indicates an attraction relationship, and the color red indicates and avoidance relationship. The network graph on the right shows another representation than the matrix of the different attraction/avoidance relationship. In this network graph, all nodes are the players and the color of the line indicates an attraction or avoidance relationship. In this example, we see that most players have a neutral relationship with each other. When only looking at significant relationships, we see that there are way more avoidance than attraction relationships.

Following pattern mining

Next, we will explain the following pattern mining algorithm in `MoveMine2.0`. This algorithm computes whether two individuals have a following relationship or not. A following relationship is two-sided; an individual can be a follower and/or a leader. The algorithm works by comparing the trajectories of two individuals. The algorithm compares the trajectories of two players by looking if they are both temporarily and spatially close at some point(s) in the trajectories. Whether an individual is temporarily or spatially close is determined by two parameters, which the user has to specify. Suppose we have two trajectories, where one individual follows somebody else. Then this means that the follower should have at least one point in the trajectory close (temporarily and spatially) to some point of the trajectory of the other individual. This point in the leader's trajectory is called a *Local Minimizer* to the point in the follower's trajectory.

It is logical that the follower should be close to the leader with some time lag. Hence, what the algorithm does, is look whether the local minimizer has a later timestamp than the point in the follower's trajectory. If this is the case, then we can call those points a following pair. The algorithm looks at intervals where such following pairs occur. Finally, the score of a following relationship is based on comparing the amount of following pairs during an interval with the expected amount of following pairs during this interval. A relationship can be called significant if the following time interval has more following pairs compared to the expectation. The expectation is, that if two individuals are moving together, there should be 50% chance of a following pair to occur at a timestamp. At the end a score is given which goes from 0 to 1. The higher the score, the more following pairs occurred compared to the expectation, and thus the clearer the following relationship. [10]

It is possible to look at following relationships between players in `Google Earth`. An example of a following relationship lasting five seconds is given in Figure 2.4. This relationship has a score of 0.962. The red dots are points of the follower when he/she is temporarily and spatially close to some point in the leader's trajectory, which is shown with the blue dots. The most left image is the start of the following interval and the most right is the end of the following interval.

Figure 2.4: Example of a following relationship of five seconds with a high score

Plotting

Finally, `MoveMine2.0` offers some functionalities which allow you to plot trajectories and information about them. These functionalities include: plotting the trajectory of an individual, plot the heatmap of an individual and plot the meeting points between two individuals. In Figure 2.5, 2.6 and 2.7 some visualizations are shown. These examples show that the conversion of the x- and y-coordinates to GPS-coordinates went well, as it is very accurate, e.g. in Figure 2.5 we can see that the goalkeeper indeed is always moving near her own goal. In Figure 2.6 we see that most passes are given in the midfield, which also makes sense. The plotting functions are useful in our research, because we can show the behavior of players graphically.

Figure 2.5: Trajectory of goalkeeper during final

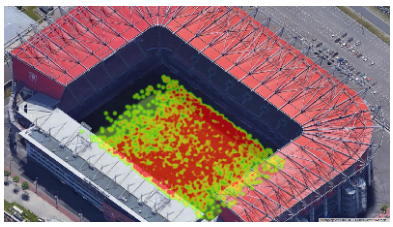

Figure 2.6: Heatmap of the positions of all passes in the six matches

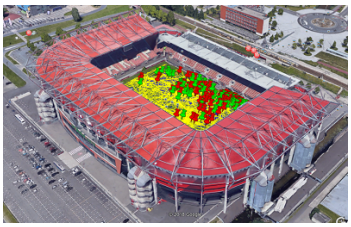

Figure 2.7: Meeting points between two players' trajectories during final

2.3 Defining ball possession

To answer the research question, we first have to define ball possession. In other words: how do you know when a team has ball possession or not, by looking at the data? This is done by looking at the event data; there, one can see for which period of time a team is in ball possession. If for example three consecutive "ballevents" occur, like a pass or a dribble, and those three events were executed by a player from the same team, then this means that the team has had ball possession for at least this period. Now, in order to define a ball possession period, we need to know when a ball possession ends. This can happen in two ways:

1. A "matchevent" occurs, e. g. the ball goes out of play, a goal is scored or the first half ends.
2. A "ballevent" occurs with a player from the **other** team.

The start time of a new ball possession period is the time when the first ballevent, linked to a player from one team occurred. Then, more ballevents could take place, linked to players from the same team, or the ball possession period could end, like explained above. The end time of a ball possession period is the time corresponding to the last ballevent in the sequence of consecutive ballevents from one team. The length of a period is computed by subtracting the end time with the start time. When examining ball possession periods, the length of a period is also something we need to take into account; a period of five or ten seconds where only two passes were played, is very short and teams will not be able to organize themselves. This means we will not see the typical interactions that occur when teams have ball possession. Hence, we are only interested in longer periods of ball possession.

This means that we now have to look at what value we should choose for the minimum length of the ball possession periods, which we later will examine. We need to choose the length of a period in such way, that (i) there are enough periods to research and (ii) that those periods are representative enough. Hence, if the minimum length of a period is too small, some periods will not be representative for "ball possession behavior" and if it becomes too big there are not enough periods to examine. At first, we look at how the minimum length of a period (in seconds) influences the number of periods that can be found in the data. In Figure 2.8 this is shown. In total, there are 1757 ball possession periods to be found in the event data. However, most of those are periods between zero and ten seconds. Furthermore, there are nearly no periods that are longer than 40 seconds.

In Table 2.1 some properties are shown for certain length-intervals. These properties are: the average amount of trajectories in a period for each player, the average amount of ballevents that occurred during the period and the amount of periods that fit in this time-interval. The longer a period is, the more trajectories can be examined, and this means that more interactions between players will be found. This also applies to the amount of events per period.

Considering the information above, we choose to look at the periods with a duration of at least 20 seconds. This is because they have more trajectories to examine and more events occur, and also because there are enough periods to consider, like we can see in Table 2.1: $102 + 30 = 132$ periods. During this research, we

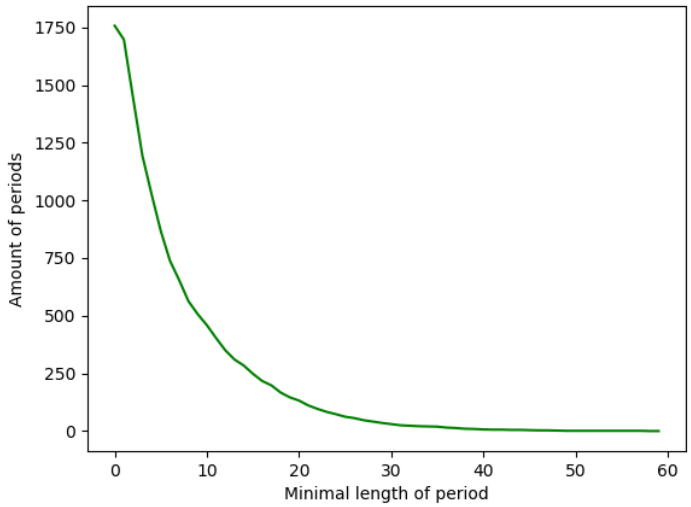

Figure 2.8

Length of period	avg. nr. of trajectories	avg. nr. of events	nr. of periods
0-10 sec.	39.81	3.82	1298
10-20 sec.	138.97	9.81	327
20-30 sec.	238.14	17.11	102
30+ sec.	372.30	28.33	30

Table 2.1: Properties in data for certain length-intervals

will constantly look at what difference the duration of a period makes, regarding attraction/avoidance- and following-scores. That is why we will be looking at periods lasting 20-30 seconds and periods with a duration of 30 seconds or more, separately. One final note regarding the ball possession periods: in the periods where one team has ball possession, the other team of course has no ball possession. We will use this fact to compare between ball possession and no ball possession.

With the information of when a team has a period of (no) ball possession, we can examine the trajectories of all players per individual period, and can compute the attraction/avoidance- and following scores from those trajectories.

Chapter 3

Related Work

Until now, we have only mentioned spatio-temporal data analysis in sports. However, spatio-temporal data analysis has many more applications. We have found some articles where research is done to interactions between moving objects. An example of this ,is article [12]. In this article, trajectories of traffic are analyzed. This is done by constructing a network graph, based on certain relationships/interactions in traffic. In the paper the authors introduce methods to detect anomalies in traffic with the help of such network graphs. What our thesis and this article have in common, is that both analyze trajectories and try to discover interactions and/or patterns in those interactions. However, we focus on soccer player's trajectories and the article researches trajectories in traffic. In our case we try finding patterns in attraction/avoidance interactions or following relationships, and in the article the authors try to detect patterns/anomalies in the routes that cars take. The main difference between the article and our thesis, is that the article is primarily focused on finding a novel method for analyzing traffic data and in our thesis, we are more interested in the results we find, rather than finding a new method for discovering interactions. Hence, we use the tool `MoveMine2.0`.

Another example of an article unrelated to soccer, that resembles our research, is [13]. This article researches wildlife tracking data, and tries to find spatio-temporal interactions in movements of animals. Furthermore, the authors try to capture different patterns in the interaction types found. Like in our thesis, the article also analyzes attraction/avoidance- and following relationships. Furthermore, interactions as encounter, coordinations and grouping are researched. The authors compose different "interaction" scenarios, like territoriality and mating, and try to detect what the dominant interaction type in those scenarios is. For finding interaction types, the authors use and compare different methods. The research in this article is very similar to our thesis, as such that the authors look at the same sort of interactions in moving object data, and also try to find patterns in interactions based on certain scenarios. In our case, the scenarios are whether a team has ball possession or not. The biggest difference between the article and our thesis, is that the article uses and compares different methods for finding interaction types. Furthermore, the article focuses more on individual interactions (between only two animals), whilst we are more interested in aggregated behavior in a *group* of individuals.

In the article [6], current approaches of quantitative data analysis for soccer are given. In this paper the authors mention that besides looking at the individual qualities of players, one can also consider the tactical aspect of a soccer game, by analyzing spatio-temporal data of players. With this information one can look at inter-player, inter-team, inter-line interactions, before critical events. Furthermore, the authors mention examples of researches to team-team interactions, meaning they compare the behavior of teams with each other, like constellations of players and compactness. This article is very general and mentions all kinds of ways to analyze position data in soccer.

In [6] there are some examples of articles that show more resemblance with our thesis. In [14] and [15] the authors try to discover patterns in tactical constellations by using more sophisticated techniques such as self-organizing maps and neural networks. The goal of both articles is to use pattern-based tactics analysis to see when certain patterns or interactions lead to success, measured in ball possession or goals. Both articles and this thesis follow the same principle; looking at interactions within teams, try to find patterns and compare the **teams** with each other. However, the articles mentioned propose their own methods to find interactions and also test this method. Furthermore, the articles do not look at predefined interaction types, such as the attraction/avoidance relationship.

Finally, there are many examples of researches to ball possession in soccer. However, most of them, like [16] and [17], are more interested in how to *detect* ball possession, by analyzing spatio-temporal data. In article [18], the authors look at determinants of ball possession, such as whether the team is winning or losing, or whether a team is playing at home or away. This research is quite different from our thesis, given that it does not analyze trajectories of players. However, it is an example of a research that analyzes the behavior and/or characteristics of teams when they have ball possession.

Chapter 4

Results

4.1 Attraction and avoidance relationship

The first relationship type we will be looking at is the attraction/avoidance relationship. For a detailed explanation of this relationship, we refer to Chapter 2.

What we want to accomplish in this section, is to discover patterns in attraction- or avoidance relationships. Furthermore, we will look at what difference it makes whether teams have ball possession or not, regarding the relationships found. We constantly make a distinction between the periods of ball possession of 20-30 seconds and the periods of 30+ seconds. However, the results we find in periods of 20-30 seconds and 30+ seconds are often quite similar. That is why we sometimes only show the results we found in periods of 20-30 seconds. In this case, the results of the periods of 30+ seconds will be mentioned in the appendix. This will also be the case for the following relationship.

4.1.1 Determining the distance threshold

Before we can take a look at the results, we first have to decide what distance threshold we will be using; the algorithm requires an input value defining when objects are spatially close to each other. Thus, we need to know when one can say that players are spatially close to each other. First, we explain how we used the Euclidean distance between players, to determine "the initial value" around which the distance threshold should lie. Secondly, we compare some results of attraction and avoidance scores with each other and look at what influence the threshold has on those results.

Looking at the Euclidean distance between players is a good way to get insight in how close players are to each other, on average. However, this does not answer the question from which distance one can say that players are *near* to each other. A better way would be to look at the lowest distance value that is found between any of the player pairs. This only gives an example of two individual cases, though, meaning it will not be

representative for all the players on the pitch. That is why we looked at what the nearest player to *every* player on the field was. This gives a certain value for every player and from that you can compute the average and the corresponding standard deviation.

Because this thesis is about researching ball possessions, we looked at the Euclidean distances between players when one of the teams had ball possession. Hence, we looked at the *individual ball possession periods*, and computed an average value and corresponding standard deviation from all those periods. One note that we should add, is that the goalkeepers are considered as special cases, because goalkeepers fulfill a different role in a match, when one of the teams has ball possession. Also, they have a small "movement space" and they are on average much further away from any other player on the field. To prove this, we show the influence of goalkeepers when computing the average distance between all players. Also we look, for all players, at the Euclidean distance to their nearest player on the field (with and without the goalkeepers). These results are shown in Table 4.1.

Periods 20-30 seconds

	Avg. distance between all players	Avg. distance to nearest player
With goalkeepers	26.58 m.	8.67 m.
Without goalkeepers	23.17 m.	7.33 m.

Periods 30+ seconds

	Avg. distance between all players	Avg. distance to nearest player
With goalkeepers	27.45 m.	9.74 m.
Without goalkeepers	24.24 m.	8.46 m.

Table 4.1

As you can see, the goalkeepers' trajectories have a large influence on both values, due to the fact that they are far away from most of the players when teams have ball possession.

We will be using the average of the Euclidean distances to the nearest player, for every player on the field, as value for determining the distance threshold. Because goalkeepers can be considered as outliers, we will leave them out. The values we are going to use as basis for determining our distance threshold, are also shown in Table 4.1 in the column `Avg. distance to nearest player`. For periods of 20-30 seconds, the value is 7.33 meters, with a standard deviation of 3.20 meters. For periods longer than 30 seconds this value is 8.46 meters, with a standard deviation of 3.11 meters. With the standard deviation we can decide the range of values we will use for the distance threshold. Considering both averages and standard deviations, we will be looking at all the distance thresholds in the range [4-11] and look at how these values influence the results. The following mining algorithm also requires a distance threshold. We will be using the same values as we computed here.

4.1.2 Which attraction/avoidance relationships do we find and how does the distance threshold influence these results?

In this section, we will give an overview of the attraction/avoidance scores found in the ball possession periods, and see how the distance threshold influences those results. To begin, we will make a distribution of the attraction/avoidance scores between all player pairs, from all the individual ball possession periods. Like

we explained, we will not look at player pairs with goalkeepers. All the resulting distributions are shown in the appendix. In Figure 4.1 and 4.2 two examples of distribution are given, which illustrates the effect of the distance threshold on the results.

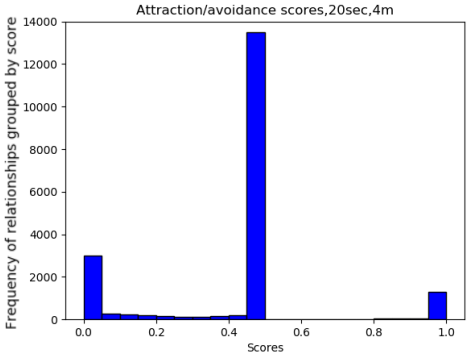

Figure 4.1

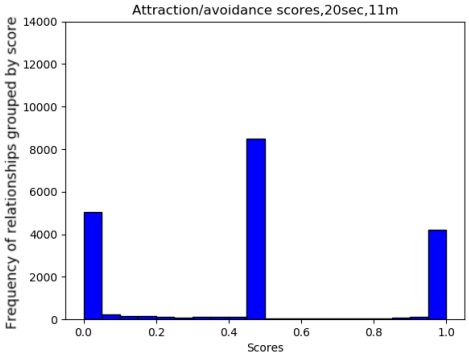

Figure 4.2

As you can see in the figures, when taking all the player pairs on the field (except keepers), nearly all the values are very close to 0.5 or very close to 1 or 0. This is because periods are not long enough to allow for more interactions, so the expected meeting frequency is low. Hence, most player pairs either only have a significant attraction/avoidance relationship or no relationship at all. The fact that most players do not have a relationship at all, is due to the fact that we are comparing every player pair and most players are not close to each other. However, if we increase the distance threshold, more significant attraction/avoidance relationships will be found. This makes sense, because raising this thresholds means more players are spatially close, according to the algorithm. Finally, if we take a look at Figure 4.1 and 4.2, and all the distributions for the other distance thresholds, then we can see that there are way more avoidance relationships than attraction relationships. Because we are comparing all the player pairs with each other we can not draw many conclusions about the distributions. We do not know what causes certain results yet. This will become clear when only looking at certain player pairs (e.g. pairs belonging to one of the teams) and/ or looking at when teams have ball possession vs. when teams do not have ball possession.

The goal right now is to find the best distance threshold which we will be using for our algorithm. First, let us get a better understanding of the influence of the distance threshold on the attraction/avoidance scores. During this research we will use a confidence level of $\alpha = 0.05$, so we will call avoidance relationships significant if they are below 0.025 and attraction relationships if they are above 0.975. 0.05 is a standard value for determining the confidence interval [19]. We will take a look at the average distance between players who have:

1. an attraction **or** avoidance relationship (score <= 0.025 or >= 0.975) .
2. only an attraction relationship (score >= 0.975).
3. only an avoidance relationship (score <= 0.025).
4. no significant relationship (0.025 <score <0.975).

These results are shown in Figure 4.3 for periods of 20-30 seconds. The plots are named, according to the four categories mentioned above. We look at the average distance values for every distance threshold used by the algorithm.

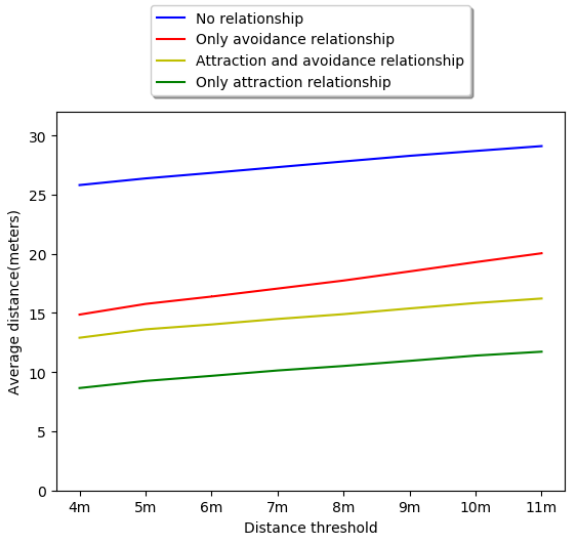

Figure 4.3: Average distance between players for certain relationship types, per distance threshold (20-30 seconds)

The distributions for periods of 30+ and 20-30 seconds follow the same pattern. However, for periods of 20-30 seconds, all the values are slightly lower than periods of 30 seconds and longer. In Figure 4.3 we show that players who have an attraction relationship, are on average moving much closer to each other than players who have an avoidance relationship. This makes sense, because those players tend to avoid each other. However, players who do not have a relationship with each other at all, are even further away. This is because those players are too far away to say anything about them. On average for every distance threshold, when players are $\pm$28 meters away, they do not have an attraction or avoidance relationship at all.

Furthermore, we can look at the attraction/avoidance scores and the amount of relationships found, given a certain distance threshold. Given Figure 4.3 and the distributions of the scores, we expect to see more relationships of any kind, when we increase the distance threshold. This is because more players will be spatially close to each other and will have a higher probability of a attraction or avoidance relationship. Let us first take a look at the scores, given the four categories mentioned above. In Figure 4.4 those are given for every distance threshold, for periods of 20-30 seconds. We can see when regarding players who have an attraction or an avoidance relationship, that whilst increasing the distance threshold, there are more attraction relationships to be found relatively to avoidance relationships. If we look at Figure 4.5, we see the patterns that we expected; when increasing the distance threshold, we see more relationships of any kind and relative to the avoidance relationship, we see more attraction relationships.

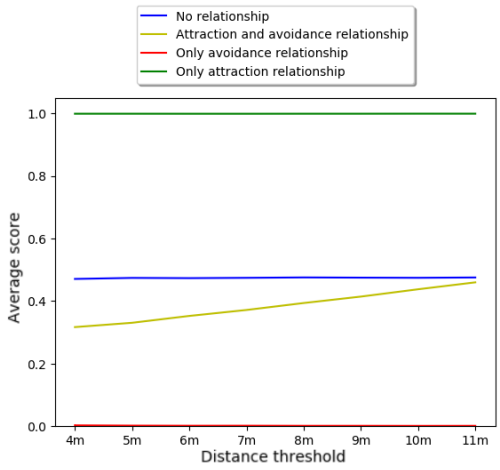

Figure 4.4: Scores for different relationship types (20-30s)

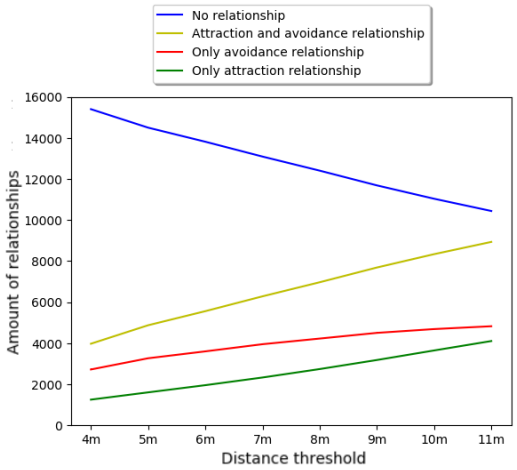

Figure 4.5: Amount of different relation types (20-30s)

To conclude, we have seen the influence of the distance threshold on the results. Considering the scores, there was no big difference. Considering the amount of relationships there was. However, we can not easily say that value x is the "right" amount of relationships found. If the distance threshold is too big, we might find relationships of players that are too far away and their relationship may be accidental. If we choose a small distance threshold, there is a higher possibility of missing attraction or avoidance relationships. That is why we will choose a value in the middle. Based on the average of the closest player for every player, we will choose a value of eight meters.

4.1.3 How do the attraction/avoidance relationships correlate between ball possessions?

In this section we will look at the differences in attraction/avoidance relationships, depending on having ball possession or not. For this, we will look at *interactions within teams*, during the ball possession periods.

We will first show the distributions of all the attraction/avoidance relationships for when teams have ball possession and do not have ball possession. Furthermore, we will take a look at average scores and count the relationships found. It is important to realize that we currently only look at interactions between players belonging to one team. Thus, for every match we look at the relationships between players belonging to a team and assign those to the distribution of the relationships found in ball possession or the distribution of the relationships found not in ball possession. We do this because this is the only way we can separate players from whom their team has ball possession, from players from whom their team does not. Otherwise, if we would look at inter-team relationships we have a player pair where one player has ball possession and the other one does not. The amount of player pairs we look at are the same for in ball possession and not in ball possession, so resulting values are comparable. The distribution of attraction and avoidance scores for periods of 20-30 seconds is given in Figure 4.6. Again, those are the distributions of all the relationships within teams, from all the matches. The average score and amount of relationships is shown in Tables 4.2 and 4.3.

The distributions in Figure 4.6 show that there are both more attraction- and avoidance relationships to be found, when teams do not have ball possession. In other words, there are more significant relationships to be

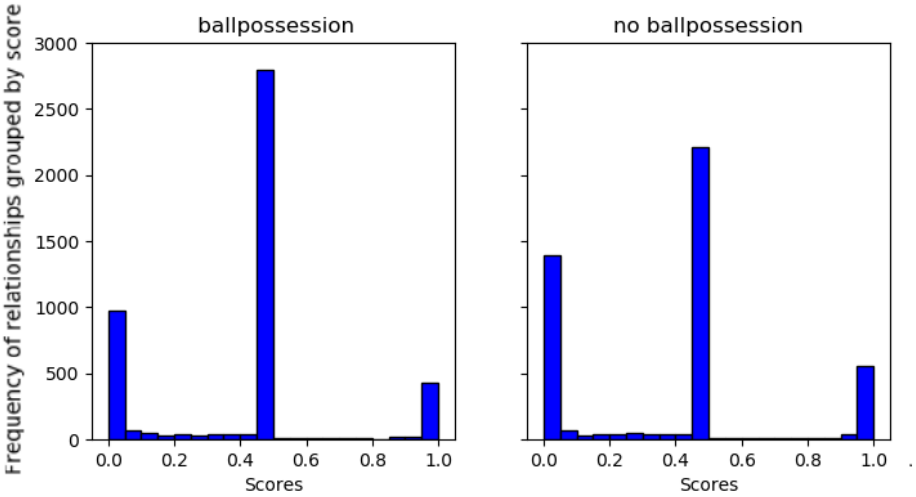

Figure 4.6: Attraction/avoidance scores distribution

	Avg. score	Amount of relationships
Attraction or avoidance relationship	0.313	1335
Only attraction relationship	0.999	417
Only avoidance relationship	0.001	918
No relationship	0.474	3255

Table 4.2: Attraction/avoidance scores when teams have ball possession (20-30 sec)

found between player pairs. There is a logical explanation for this; the distance between players is lower when they do not have ball possession. Like Figure 4.3 told us, the higher the distance between players the higher the chance they will not have a attraction/avoidance relationship at all.

The world-famous Dutch ex-footballer and -coach Johan Cruyff once said: "If you have the ball you must make the field as big as possible, and if you dont have the ball you must make it as small as possible." In general this is an unwritten law in soccer; in ball possession players should move away from each other to create more space to pass and make sure the opponent has to defend a larger area. When teams do not have ball possession they need to stand more compact, in order to make space smaller for the opponent and being able to defend better. Hence, they will move towards each other.

Next to being a well-known phenomenon in soccer, there is also proof in the data that substantiates the statement. The average distance between all players from the same team, when their team has ball possession is 25.72 meters, for periods of 20-30 seconds. When their team does not have ball possession this is 21.09 meters. For periods of 30+ seconds it is even more clear: the values are respectively 27.99 meters and 20.99 meters. This explains why we find more significant relationships when teams do not have the ball.

Next, we want to know if certain interactions are more likely to occur, given the fact that teams have ball possession or not. From now on, we will only consider the **significant** relationships. We already know that there are both more significant attraction- and avoidance relationships when teams do not have ball possession. The mean of periods of 20-30 seconds of all significant relationships is 0.313 for teams in ball possession and 0.287 for teams not in ball possession. This suggests that there are relatively more attraction- than avoidance

	Avg. score	Amount of relationships
Attraction or avoidance relationship	0.287	1827
Only attraction relationship	0.999	536
Only avoidance relationship	0.002	1336
No relationship	0.472	2718

Table 4.3: Attraction/avoidance scores when teams do not have ball possession (20-30 sec)

relationships when teams have ball possession, compared to when teams do not have ball possession, because the value is closer to one. We can compare both means of the distributions with each other by performing a t-test; this way we can see whether there are significantly more (or less) attraction- than avoidance relationships, when comparing ball possession with no ball possession.

Like we mentioned, we will look at two distributions: all significant relationships when teams have ball possession, and all significant relationships when teams do not have ball possession. We still only look at the periods of 20-30 seconds. Our null- and alternate-hypotheses are as follows:

1. **Null-hypothesis**: The ratio of attraction- to avoidance relationships will be the same for ball possession and not in ball possession.

2. **Alternate-hypothesis**: The ratio of attraction- to avoidance relationships will *not* be the same for ball possession and not in ball possession.

We will use a two-sided test with confidence threshold $\alpha = 0.05$. The t-value we got from comparing both means is ± 3.433. The corresponding p-value is ± 0.001. This is smaller than our threshold α. Hence, we can reject the null-hypothesis and say the mean is significantly different for the ball possession distribution. Because we have a positive t-value, we can say that there are significantly *more* attraction relationships compared to avoidance relationships when in ball possession.

For periods of 30+ seconds the means of both distributions are closer to each other. Namely they are ± 0.199 for in ball possession and ± 0.204 for not in ball possession. Here the mean is higher for teams not in ball possession. However, we want to know if this is significant or not. The t-value we got from our test is $\pm$-0.564 and the corresponding p-value is ± 0.573. This is higher then our α, so we can accept our null-hypothesis. This means that the means of both distributions are not significantly different. Thus there is not much evidence supporting the fact that the ratio of attraction- to avoidance relationships is different for ball possession and no ball possession in periods of 30+ seconds.

To conclude, we have seen that there are more significant relationships of any kind, when teams do not have ball possession. We also explained that this was due to the lower distance between players. Next, we compared the mean score of both distributions, to see whether there are *relatively* more/less attraction relationships compared to avoidance relationships, when comparing ball possession and no ball possession. For periods of 20-30 seconds it was very clear that there were relatively more *attraction*- than avoidance relationships. This is not something we would expect, given the fact that the distance between players is bigger in ball possession. Namely, in Figure 4.3 we saw that the *avoidance* relationship occurs more often between players with a bigger distance between them, instead of the attraction relationship. However, despite the distance between players

being higher, there are significantly more attraction relationships to be found *in ball possession*, at least for the periods of 20-30 seconds.

We can explain the results we found. Like we mentioned, players move away from each other when having ball possession, in order to create space. This only leads to the fact that many players are too far away from each other to have a relationship at all. Hence, we saw less significant relationships of any kind. For the players that are close enough for a relationship, more players are likely to move towards a teammate, rather than moving away from them. A logical explanation for this is, that players in ball possession that are close enough for an attraction - or avoidance relationship, ask their teammates for the ball, by approaching them. This leads to the fact that we relatively find more attraction relationships than to avoidance relationships in ball possession, compared to teams not in ball possession.

4.2 The following relationship

In this section we take a look at the different following relationships during the ball possession periods. Furthermore, we want to see how ball possession influences those relationships and if we can discover certain patterns in following relationships. For an extensive explanation of how the following mining algorithm works, we refer to the Chapter 2.

4.2.1 Which following relationships do we find and how do the different parameters influence these results?

We will first give an overview of all the following scores found when looking at all the ball possession periods, between any two players on the pitch. Also, we will show a distribution of the length of the following relationships. There are a few things we need to keep in mind. First, a player pair can have more than one following relationship during a period of ball possession. This occurs when multiple following patterns are discovered for a player. Furthermore, the algorithm needs two input values: a time threshold and a distance threshold. The time threshold specifies until which time lag one can say that individuals are still temporarily close. Hence if this value is set at five seconds, a player may arrive at a previous position of another player within five seconds, in order to call those players temporarily close. The distance threshold determines whether individuals are spatially close or not. In Figure 4.7 an overview is given for the periods of 20-30 seconds with a distance threshold of eight meters and a time threshold of five seconds. The distributions of periods longer than 30 seconds are mentioned in the appendix.

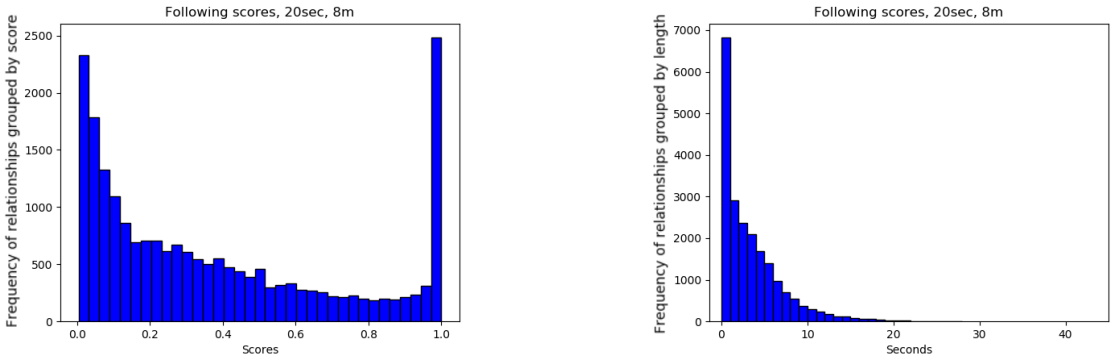

Figure 4.7: Distributions of scores and length for following relationships in periods of 20-30 seconds

When looking at the following scores for periods of 20-30 seconds, we see that most scores are near to zero. However, there are also many "perfect" following relationships with a score of around 1. If we look at periods of 30+ seconds, there are relatively less perfect scores compared to lower scores. Considering the length of the following relationships, we see that most relationships only are about one second long. Some other properties are given in Table 4.4.

	Periods 20-30 seconds	Periods 30+ seconds
Average score	0.383	0.280
Std. dev. score	0.336	0.288
Average length	3.402 sec.	3.353 sec.
Std. dev. length	3.661 sec.	3.772 sec.
Maximum length	28.5 sec.	42.4 sec.

Table 4.4: Properties of the following relationships found

For periods of 30+ seconds, the following relationships are on average a bit longer and have a lower score. Nevertheless, the average following score is $\pm$ 0.3 and the average length is $\pm$ 3.4 seconds. The longest following relationship found with the current thresholds was 42.4 seconds.

Now that we have an overview of the following relationships we found in the date, we will look at the different parameters that influence the results we find. The four different parameters we are going to look at are:

- The minimum *length* of a following relationship
- The minimum *score* of a following relationship
- The *distance threshold* we had to specify for the following mining algorithm
- The *time threshold* we had to specify for the following mining algorithm

Minimum length threshold

At first, let us take a look at the minimum length of the following relationships. As I showed in Figure 4.7, most following relationships are very short. Moreover, the most frequent length for a following relationship is between 0 and 1 seconds. There are even some following relationships that only have a length of 0.1 seconds, which corresponds to one point in the trajectory for both follower and leader in the relationship. This is not representative at all. This is why from now on we will only be considering a following relationship if it is *at least one second* long. This changes some properties of the data by much, as shown in Table 4.5. As you can see, the average scores are much higher, when ignoring all the following relationships shorter than 1 second. This implies that the relationships shorter than one second already had a low score. Furthermore, the average length per relationship goes up, which of course is very logical.

Minimum score threshold

Secondly, we will take a look at the impact a score threshold has on the following results. To understand the differences in following relationships for a high score and a very low score, we give two examples of following

	Periods 20-30 seconds	Periods 30+ seconds
Average score	0.494	0.361
Std. dev. score	0.314	0.284
Average length	4.846 sec.	4.733 sec.
Std. dev. length	3.642 sec.	3.839 sec.

Table 4.5: Properties of the following relationships that are over one seconds.

relationships with a length of five seconds. The first following relationship in Figure 4.8 has a score of 0.160 and the seconds following relationship in Figure 4.9 has a score of 0.962. The most left image in both figures is the start of the following relationship. The most right image is the end of the following relationship. The red dot in the figures is the follower and the blue dot the leader. It is very obvious that the following relationship with higher score is much clearer and less random. Also, the following relationship with the higher score has a lot more points in both trajectories that are temporarily and spatially close.

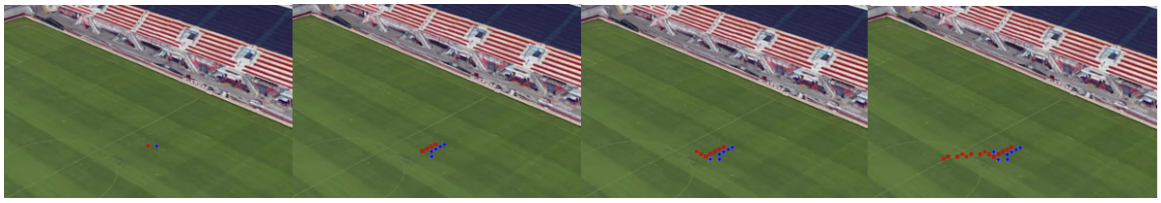

Figure 4.8: Example of a following relationship of five seconds with a low score

Figure 4.9: Example of a following relationship of five seconds with a high score

We want to know how a certain score threshold will influence the results. It goes without saying that the higher the minimal score threshold, the less following relationships we will find. It is also clear that the average score will be higher. However, we are interested in how the length of the following relationships will be influenced. Earlier, we saw that if we took the following relationships of lower than one second away, the quality of the relationships went up. We will still only look at periods of at least one second long. However, we are interested to see if it is also the other way around; if we increase the score, we want to know if the length of the period increases. In other words, we want to know if the score and length are correlated with each other. In Figure 4.10 and 4.11 we see that, if we raise the score threshold, the average length of the following relationships goes up, until the threshold reaches a certain value. For periods of 20-30 seconds, this value is 0.5 and for period of 30+ seconds this value is 0.4. For periods of 30+ seconds, the score threshold even has a higher impact on the average length. This is probably due to the fact that there are less examples of periods of 30+ seconds. As score threshold, we will take a **value of 0.4**, because this gives for periods the maximum average length of 30+ seconds, and for periods of 20-30 seconds nearly the optimum. Also, if we would pick a higher score threshold there would be less examples.

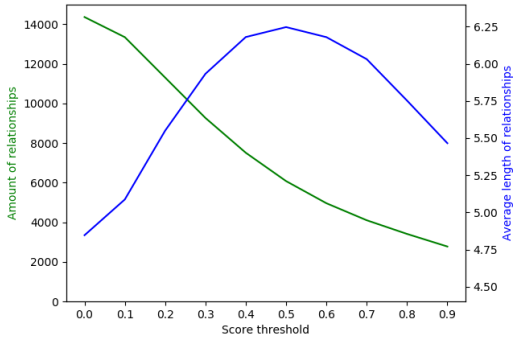

Figure 4.10: Periods of 20-30 seconds

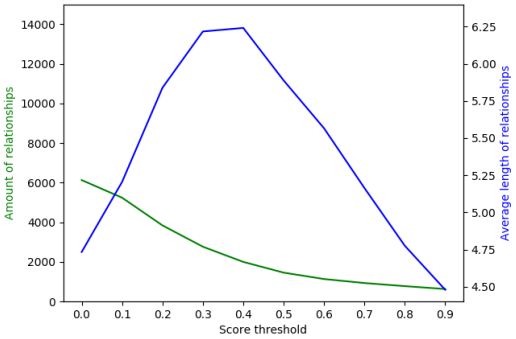

Figure 4.11: Periods of 30+ seconds

Amount of following relationships and the average length of the following relationship, given a certain score threshold

Distance threshold

As a third, we will look at the impact of the distance threshold on the following relationships. We expect to see more following relationships when we increase the threshold, because more players will be spatially close. However, we do not yet know what the impact will be on the scores and average length of the relationships. As said, we will skip relationships of under one second. The time threshold we will be using for now is five seconds. The distance threshold will be a variable ranging through the values four to eleven. These are the same values as we used with attraction/avoidance. In Figure 4.12 and 4.13 the following properties per distance threshold are given: the amount of following relationships, the average score and the average length.

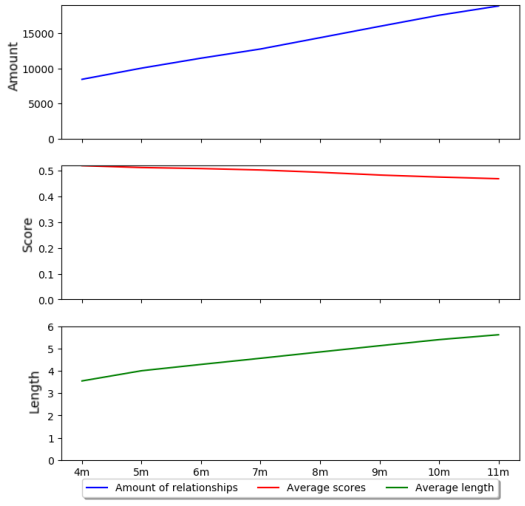

Figure 4.12: Periods of 20-30 seconds

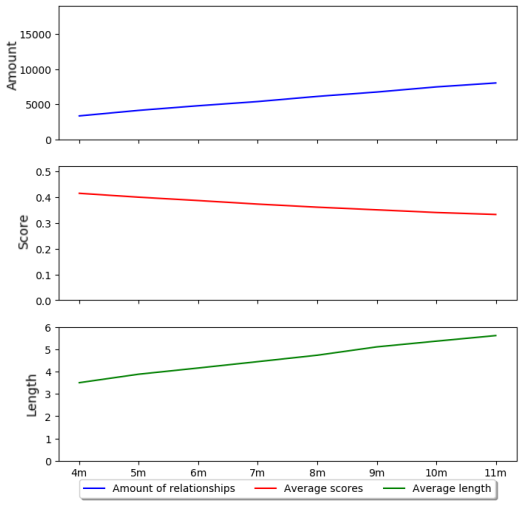

Figure 4.13: Periods of 30+ seconds

Influence of the distance threshold on certain properties in the data

As we expected, the amount of relationships found when increasing the distance threshold does go up. Furthermore we see that the following score decreases as the distance threshold increases. This could be due to the fact that when the distance threshold is higher, more players are "accidentally" close for a few subsequent timestamps, but this does not necessarily mean that this player actually is following another player. Hence, the

following relationship is not clear and the score is low. Finally, we see that the average length of the following relationships increases as we increase the distance threshold. We can conclude from this that the higher the distance threshold, the more following relationships with a higher length and lower score we will find. We need to choose between a high score on the one hand, and more relationships with a higher length on the other hand. Because all three are important criteria for different reasons, we will choose some value in between. As you can see, especially for periods of 20-30 seconds, the score threshold does not decrease by much until a distance threshold of seven meters. Hence, a threshold of seven meters seems to be the most logical choice; this is the highest distance threshold, where the score is not influenced too much.

However, there is still something we have to consider here. As we have seen until now the average score drops and the amount of relationships increases when we increase the distance threshold. However, like we explained in the section about the score threshold, we are interested in scores of 0.4 and higher. When we only consider those relationships, the question is if we still find more relationships, when we increase the distance threshold. In Figure 4.14 and 4.15 we see some properties of following relationships with the same thresholds as in Figure 4.12 and 4.13, but the relationships must have a score higher than or equal to 0.4. Here, we still see the same trends as in Figure 4.14 and 4.15, but then with less relationships with higher scores and lengths. Hence, we will choose a value that is in the middle, where the average score is not too low: seven meters.

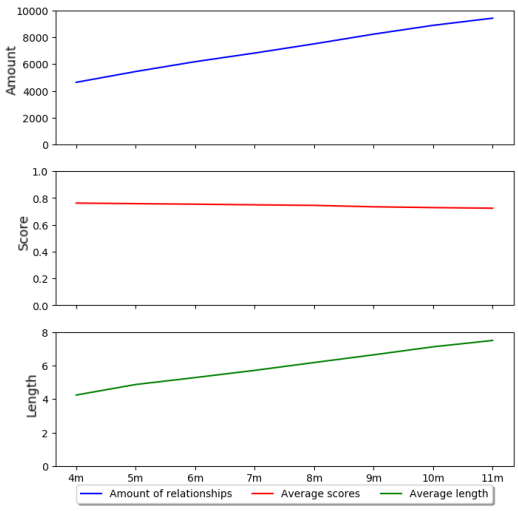

Figure 4.14: Periods of 20-30 seconds)

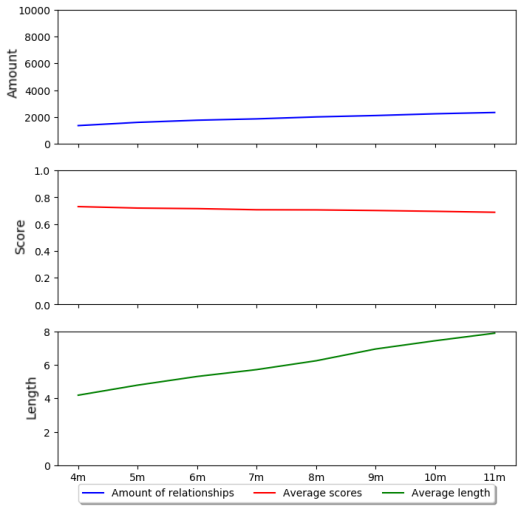

Figure 4.15: Periods of 30+ seconds)

Influence of the distance threshold on certain properties in the data, where the minimum score is 0.4

Time threshold

Finally, we will take a look how the time threshold influences the results. Like with the distance threshold, it is logical that the lower the time threshold, the less relationships will be found. This is due to the fact that less players will be temporarily close to each other. If there is a lot of time passing (high time threshold) before a player is spatially close to another player, chances are bigger that this is a coincidence rather than that this player is following some other player. This is also why we expect that with a high time threshold, the average following scores will be lower; players will be walking by "accidentally" and are not following the other player.

To determine this threshold, we will let the value run from one till ten seconds and see whether there are (i) enough periods (ii) those relationships have a high enough score and (iii) length. We will use a distance threshold of seven meters and leave relationships of under one second out. In Figure 4.16 we see properties of all the relationships when varying the time threshold, for periods of 20-30 seconds. In Figure 4.17 we see properties of relationships when the score is above 0.4, for periods of 20-30 seconds. The figures regarding the periods of 30+ seconds are mentioned in the appendix.

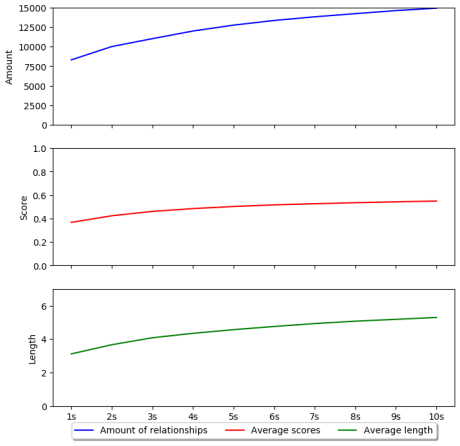

Figure 4.16: All following relationships

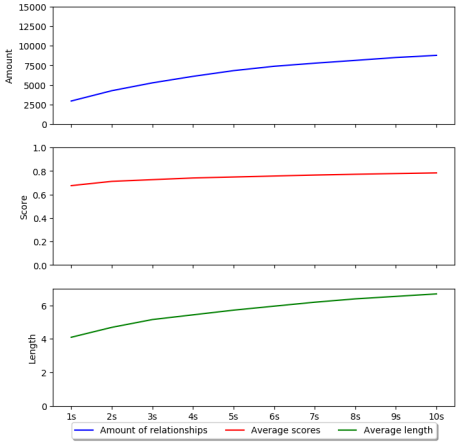

Figure 4.17: Following relationships with a score higher than 0.4

Influence of the time threshold on certain properties in the data

First, we see that the length of the relationships and the amount of relationships go up as we increase the time threshold, as expected. When considering the score, we see that in the beginning the average score is influenced much by the time threshold, but at a certain point the score does not increase by much anymore. For the relationships with a score higher than 0.4 this is also the case, but to a lesser extend. However, we expected the score to drop at a certain point. This does not happen. On the contrary, the score keeps increasing, although not by much. An explanation for this is that if the time lag is bigger for players, they have a higher chance to be spatially close on a certain point in that period, which causes the score to go up. However, this still does not guarantee that a player is actually following another player. We do not want to find accidental following relationships, so we want to keep the time threshold as low as possible, but we also want an average score that is high enough, relationships that are long enough, and enough relationships to examine. In Figure 4.16 we see that the score increases a lot in the beginning and than stagnates a little bit at a time threshold of five seconds. For Figure 4.17 the score stagnates even earlier. Also the amount of relationships and the length of the periods increase much in the beginning, and stagnate at a time threshold of five. Hence, we will take a time threshold of *five* seconds; this is the lowest value after which the three criteria do not increase by much anymore.

4.2.2 How do the following relationships correlate between ball possessions?

Now that we know which threshold values we are going to use, we can focus on the actual results. In this section we will look at the differences in following relationships depending on ball possession.

At first, we have to explain how we are going to look at the following relationships. Namely, we want to research the following relationship in the same way as the attraction/avoidance relationship; we want to look at **player pairs**, where pairs have a certain value specifying their following relationship during a ball possession period. The only difference is, that the following relationship between players can go in two directions; a player can either be a follower or a leader in a pair. Hence, player pairs can have two values. This is how we will look at the following relationship: the algorithm outputs the top ten following relationships in a certain period[1], for each player in a player pair. So, it outputs the top ten following relationships found when player "A" is the follower and player "B" is the leader, and vice versa. What we will be looking at, is the sum of the length of those following relationships, when the following score is above a certain threshold. In other words, when a following relationship is significant enough (so, the score is above 0.4), then we will take this one into consideration. What we will have as end result is the total *length* in seconds that a player followed another player, during a period.

First, we will look at the amount of following relationships in ball possession compared to not in ball possession. We are firstly interested in the amount of following pairs we can find in ball possession compared to not in ball possession. For periods of 20-30 seconds the amount of player pairs with a following relationship is 1125 in ball possession and 1730 not in ball possession. For periods of 30+ seconds, the amount of following pairs is 334 in ball possession and 483 not in ball possession. Thus, it is very clear that when teams do not have ball possession that there are more following relationships. An explanation for this, is that teams that do not have ball possession have to defend in a disciplined manner by staying close to each other. So, if for example one defender goes to the left, the defender next to him/her has to move along in order to stay compact. Players in ball possession can move more freely; sometimes they will come asking for the ball and sometimes they will run forward. Furthermore, the distance between players not in ball possession is smaller, like we already explained in the previous section. This means that players are more likely to be spatially close to each other, increasing the probability of a following relationship to occur. The distributions of the following relationships for in ball possession and not in ball possession in periods of 20-30 seconds are given in Figure 4.18. For 30+ seconds the distributions are mentioned in the appendix.

[1]It will never happen that there are more than ten following relationships between a player pair, so the algorithm will find all the following relationships in a period, for a player pair.

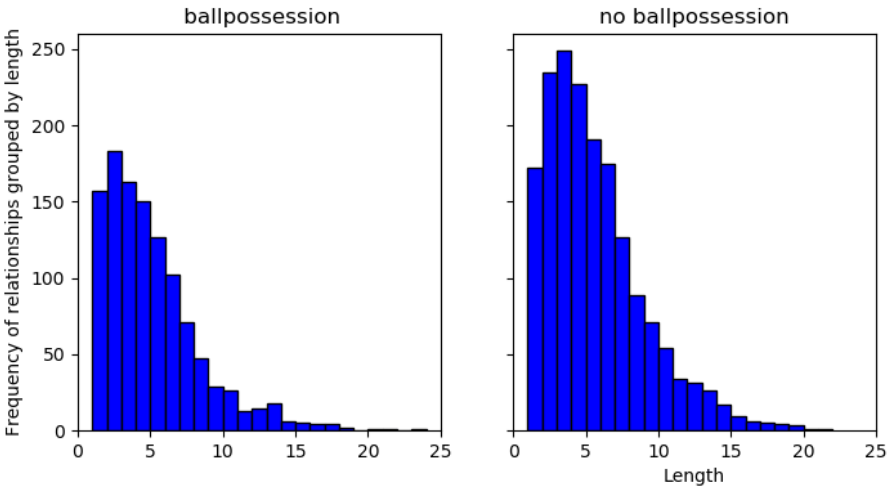

Figure 4.18: Length of following relationships distributions

So, we know that there are less following pairs in ball possession. Next, we want to know if the length a player followed another player in a following pair, will be different in ball possession, compared to not in ball possession. If we look at the average length of the following relationship for both distributions, we already see that relationships for in ball possession are shorter: for periods of 20-30 seconds the means are respectively 5.065 and 5.617 and for periods of 30+ seconds 4.571 and 5.671. As we can see, for periods of 30+ seconds, the difference in means is even higher. To see whether the means are significantly different, we will use a t-test to compare the means of both distributions (in ball possession and not in ball possession).

Our null- and alternative-hypotheses are:

1. **Null-hypothesis**: The length of a following relationship are the same for in ball possession and not in ball possession.

2. **Alternate-hypothesis**: The length of a following relationship is lower for ball possession compared to not in ball possession.

We will be using a one-sided test, because we expect the scores to be lower and not higher, given the means of the distributions. The confidence level we will be using is $\alpha = 0.05$. Hence, if the p-value we get is lower than 0.05, we can say that the length of the following relationship is significantly lower for ball possession. The test-statistics for periods of 20-30 seconds and 30+ seconds are respectively -1.251 and -1.458. The corresponding p-values are 0.106 and 0.069. Thus, we can conclude that for both periods, the means of the distributions are not significantly different and we can reject our alternate hypothesis. However, for both period intervals and especially for periods of 30+ seconds the value is very close to 0.05. This means that, although the differences are not significant, we can say that in most cases the length of the following relationships when teams have ball possession are indeed lower compared to when teams do not have ball possession.

We are interested how we can explain this behavior. So, we want to know what the reason is, that players follow each other for a longer period of time when they do not have ball possession. First of all, when teams have ball possession, it is not smart for a player to follow players from the same team. This is because it will be

easier to defend them; defenders can focus on defending the leader in the following pair and at the same time do not have to cover another area in order to defend the follower. Furthermore, there is also a mathematical reason to explain why players follow other players for a longer period, when their team does not have ball possession. Namely, there are less following pairs to be found in ball possession. This also indicates that there are less following relationships to be found. This leads to the fact that the length of a following in a pair will probably also be smaller. This is because we take the sum of the following relationships found in a period between two players. Hence, the sum of the following relationships will also be lower. In other words, players not in ball possession will on average follow other players from their team longer during periods. The fact that we could find this in the data, is proof that there are more following relationships when teams do not have ball possession.

Chapter 5

Conclusions

In this thesis we have researched two relationship types: the attraction/avoidance relationship and the following relationship. The goal was to see if we could find such interactions and maybe patterns of those interactions in soccer data. Furthermore, we wanted to research if there were any differences in those interaction types, depending on the condition of having ball possession or not.

First of all, we looked at the attraction/avoidance relationship. We found that the Euclidean distance between players highly affects whether players had an attraction, avoidance, or neutral relationship. Also, we discovered what the influence of the distance threshold on the attraction/avoidance relationship was. At last, we saw the influence of having ball possession on the attraction/avoidance relationship: when having ball possession, players in a team are more likely to have an attraction relationship, compared to when a team does not have ball possession.

Secondly, we researched the following relationship. We saw that most relationships were very short and also had a very low score. That is why we decided to leave out relationships of under one second. Furthermore, we looked at the effects of the score-, distance- and time threshold on the following relationship. At last, we researched the differences for ball possession and no ball possession, like we did with the attraction/avoidance relationship. We saw that there were much more following relationships when teams do not have ball possession, and players also follow other players for a longer time when they do not have ball possession.

5.1 Future work

The research in this thesis is still only "the tip of the iceberg" compared to what still can be done into analyzing the attraction/avoidance- and following relationship and their correlation with ball possessions. In this research we only look at the interactions between all the players and see if we can discover interesting insights. Furthermore, we only look at team-team interactions (we do not compare individual players with each other).

The research can be expanded by looking at the position of players in a team; we can for example analyze inter-line relationships and see if we can discover certain differences. Another possibility is look at inter-team interactions; how do defenders of teams interact with attackers of other teams?

Furthermore, it is possible to look at some sort of performance measure, like the duration a team holds ball possession or the amount of shots on target. There might be a connection between this performance measure and the interactions between players.

Bibliography

[1] M. Lewis, *Moneyball*. Ww Norton & Co, first ed., 2004.

[2] L. Steinberg, "Changing the game: The rise of sports analytics," *Forbes*, 2015.

[3] L. Vilar, D. Arajo, K. Davids, and C. Button, "The role of ecological dynamics in analysing performance in team sports.," *Sports medicine*, vol. 42(1), pp. 1–10, 2012.

[4] M. Kempe, A. Grunz, and D. Memmert, "Detecting tactical patterns in basketball: comparison of merge self-organising maps and dynamic controlled neural networks.," *European journal of sports science*, vol. 15(4), pp. 49–55, 2014.

[5] D. Memmert, *Teaching Tactical Creativity in Sport*. Routledge, first ed., 2015.

[6] D. Memmert, K. A. P. M. Lemmink, and J. Sampaio, "Current approaches to tactical performance analyses in soccer using position data," *Sports Medicine*, vol. 47(1), pp. 1–10, 2017.

[7] P. D. Jones, N. James, and S. D. Mellalieu, "Possession as a performance indicator in soccer.," *International Journal of Performance Analysis in Sport*, vol. 4(1), pp. 98–102, 2004.

[8] Z. Li, B. Ding, F. Wu, T. K. H. Lei, R. Kays, and M. C. Crofoot, "Movemine 2.0: Mining object relationships from movement data," *Proceedings of the VLDB Endowment*, vol. 7(13), pp. 1613–1616, 2014.

[9] F. Wu, T. K. H. Lei, Z. Li, and J. Han, "Attraction and avoidance detection from movements," *Proceedings of the VLDB Endowment*, vol. 7(3), pp. 157–168, 2013.

[10] Z. Li, F. Wu, and M. C. Crofoot, "Mining following relationships in movement data," *2013 IEEE 13th International Conference on Data Mining*, vol. 7(3), pp. 458–467, 2013.

[11] "Woods hole oceanographic institution." [Online; accessed 1-April-2018].

[12] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xie, "Discovering spatio-temporal causal interactions in traffic data streams," *KDD*, pp. 1010–1018, 2011.

[13] J. A. Long, "Quantifying spatial-temporal interactions from wildlife tracking data : issues of space, time, and statistical significance.," *Elsevier*, vol. 26, pp. 3–10, 2015.

[14] J. Perl, A. Grunz, and D. Memmert, "Tactics analysis in soccer an advanced approach," *International Journal of Computer Science in Sport*, vol. 12(1), pp. 33–44, 2013.

[15] A. Grunz, D. Memmert, and J. Perl, "Tactical pattern recognition in soccer games by means of special self-organizing maps.," *Human movement science*, vol. 31(2), pp. 334–343, 2012.

[16] X. Yu, C. Xu, H. W. Leong, Q. Tian, Q. Tang, and K. Wan, "Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video," *ACM Multimedia*, vol. 11(3), pp. 11–20, 2003.

[17] X. Yu, C. Xu, H. W. Leong, Q. Tian, Q. Tang, and K. Wan, "Detection of individual ball possession in soccer," *Computer Science in Sports*, pp. 103–107, 2015.

[18] C. Lago and R. Martin, "Determinants of possession of the ball in soccer.," *Journal of sports sciences*, vol. 25, pp. 69–74, 2007.

[19] A. Agresti and B. Finlay, *Statistical Methods for the Social Sciences*. Pearson, fourth ed., 2014.

Chapter 6

Appendix

6.1 Attraction and avoidance relationship

6.1.1 Distributions of all attraction/avoidance scores (periods 20-30sec and 30+ sec)

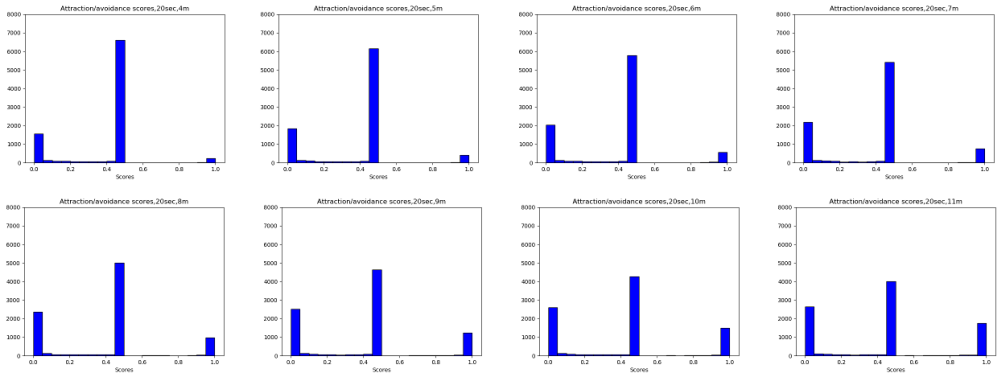

Figure 6.1: Periods of 20-30 seconds

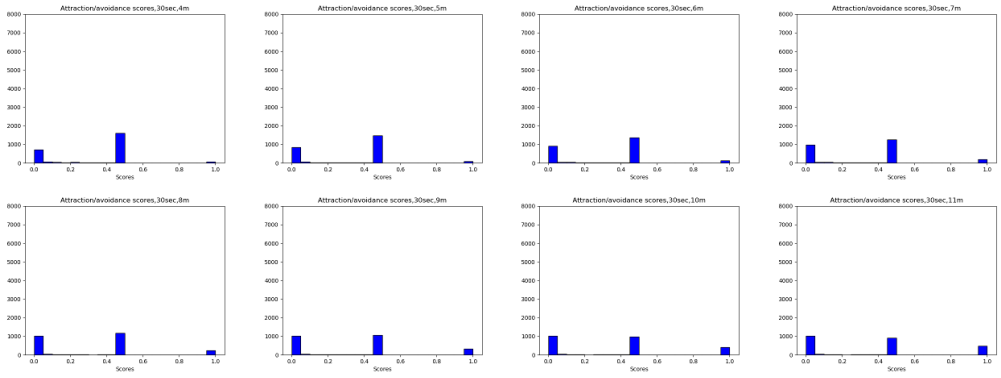

Figure 6.2: Periods of 30+ seconds

6.1.2 Avg. distance between players with certain attraction/avoidance scores

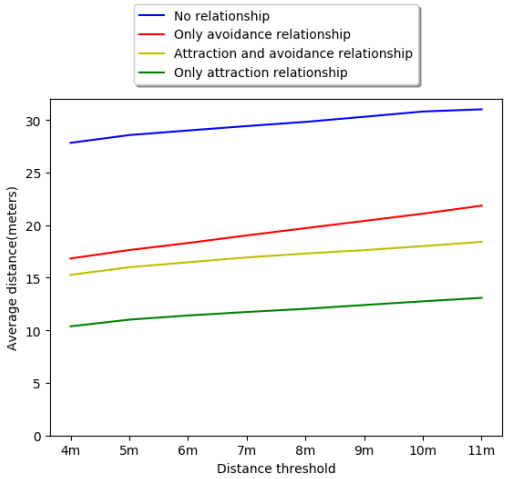

Figure 6.3: Average distance between players for certain relationship types, per distance threshold

6.1.3 Attraction/avoidance scores and amount of relationships when teams have ball possession or not (periods 30+ sec)

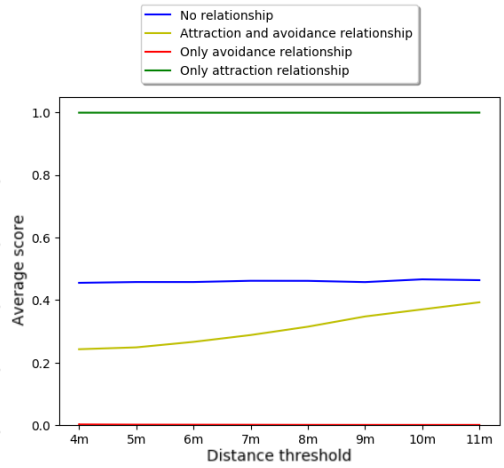

Figure 6.4: Scores for different relationship types (30+s)

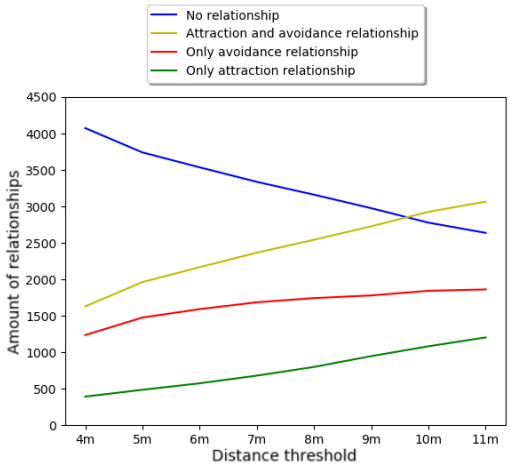

Figure 6.5: Amount of different relationship types(30+s)

6.1.4 Distributions and properties of the attraction/avoidance scores in ball possession and not in ball possession (periods 30+ sec)

	Avg. score	Amount of relationships
Attraction or avoidance relationship	0.199	485
Only attraction relationship	0.999	96
Only avoidance relationship	0.002	389
No relationship	0.459	865

Table 6.1: Attraction/avoidance scores when teams have ball possession (30+ sec)

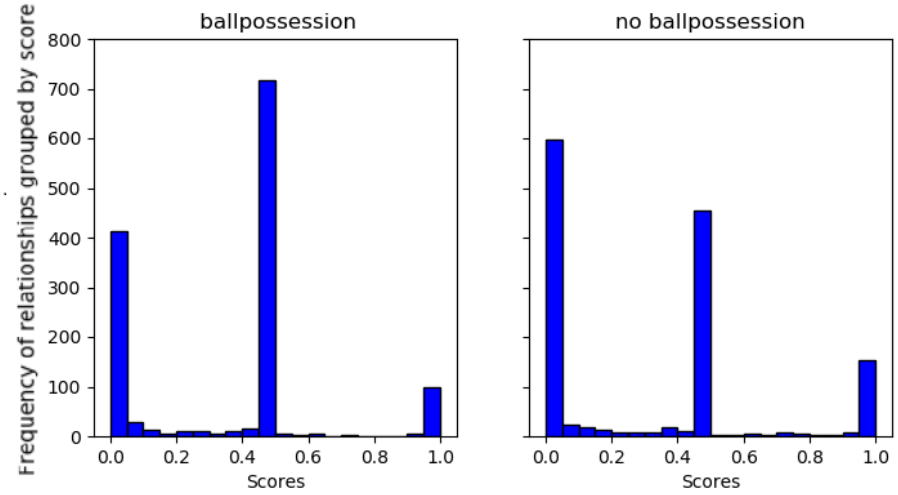

Figure 6.6: Attraction/avoidance scores distribution

	Avg. score	Amount of relationships
Attraction or avoidance relationship	0.204	726
Only attraction relationship	0.998	148
Only avoidance relationship	0.001	578
No relationship	0.459	624

Table 6.2: Attraction/avoidance scores when teams do not have ball possession (30+ sec)

6.2 The following relationship

6.2.1 Overview of the scores and length of all following relationships found (periods 30+ seconds)

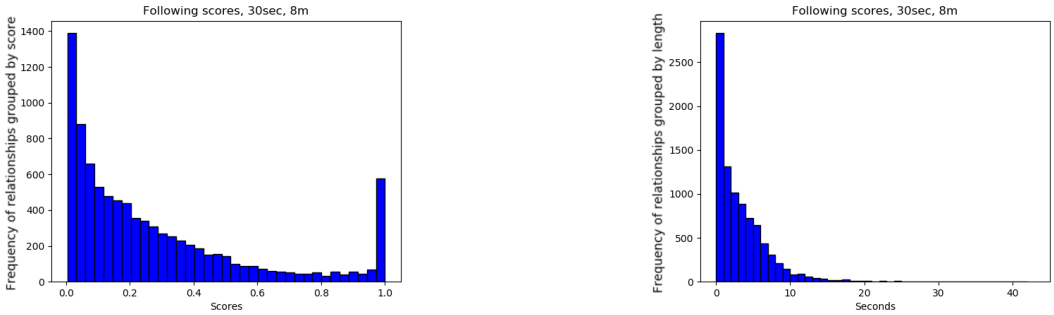

Figure 6.7: Distributions of scores and length for following relationships in periods of 30+ seconds

6.2.2 Time threshold (periods 30+ seconds)

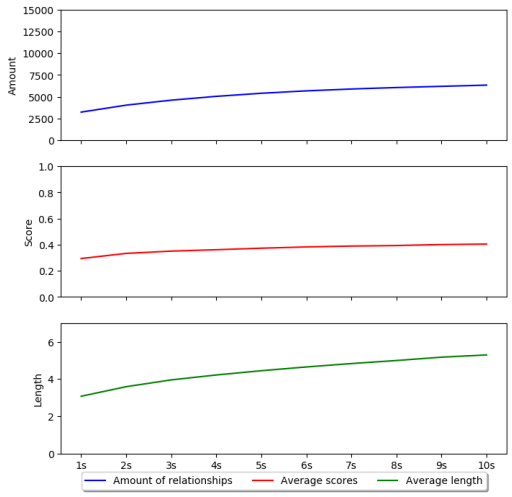

Figure 6.8: All following relationships

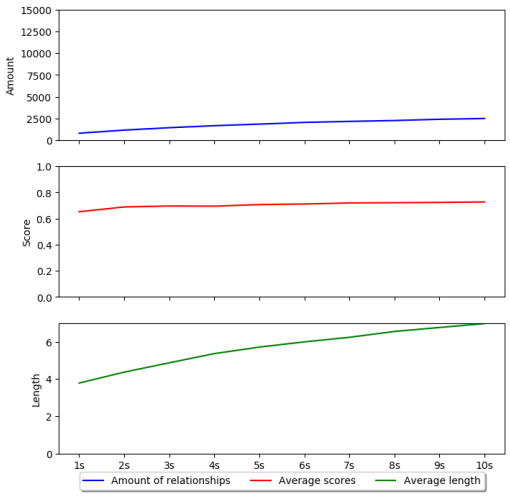

Figure 6.9: Relationships with a score higher than 0.4

6.2.3 Distributions of the following relationship lengths in ball possession and not in ball possession (periods 30+ sec)

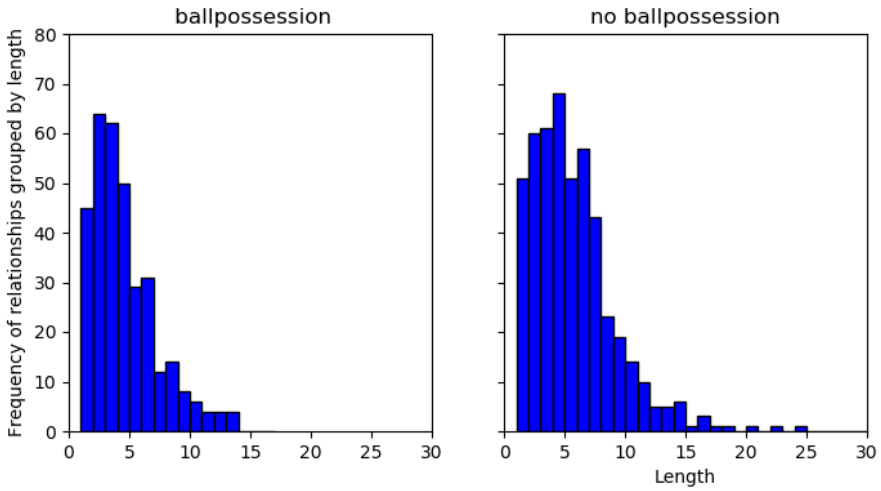

Figure 6.10: Length of following relationships distributions