Universiteit Leiden
The Netherlands

Opleiding Informatica

Analyzing the behaviour of

players in soccer prior to a shot on target.

Rose Browne

Supervisors:
Mitra Baratchi & Arie-Willem de Leeuw

BACHELOR THESIS

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

02/07/2018

Abstract

Developments regarding tracking and sensor technologies have led to large amounts of movement data within the field of sports. Performing spatio-temporal data mining on this sports data has become of interest for several reasons, such as the development of new spatio-temporal mining techniques or the commercial interest in the potentially usefull information that lies in this data. In this thesis, we were interested in the behaviour of the players who shoot on target, and in particular the behaviour prior to that shot. We were looking to find whether these so called shot on target players show deviant behaviour compared to other players. To investigate the behaviour of the players three pairwise relations were used: attraction, avoidance and following relations. An attraction relation between a pair of players is used to indicate whether these players are drawn to each other, an avoidance relation is used to indicate whether two players are dodging each other and a following relation tells us whether a certain player is going after some other player. By using these pairwise relations we found that the shot on target player shows deviant behavior in comparison to other players. Based on this research we can conclude that when a player makes a shot on target his behavior differs from his usual behavior. In terms of soccer the shot on target players behavior indicated that direct play is a successful tactic in creating goal scoring opportunities.

Contents

1 Introduction **1**

1.1 The need for mining soccer data . . . 1
1.2 The problem to be solved . . . 1
1.3 Related work . . . 2

2 Definitions **4**

2.1 Spatio-temporal data mining . . . 4
2.2 Soccer . . . 4
2.3 The dataset . . . 5
2.4 Interaction patterns between players in soccer . . . 7
2.5 Movemine . . . 8
2.6 Movemine functions and the football patterns they represent . . . 10

3 Data preprocessing **12**

3.1 Extracting trajectories . . . 12
3.2 Performing Movemine's functions . . . 14
3.3 Collecting the results . . . 16

4 Results **17**

4.1 Dependencies between the results and the used parameters . . . 17
4.2 Finding the number of relations . . . 20
4.3 Overall behaviour of the shot on target player . . . 25
4.4 Detailed behaviour of the shot on target player . . . 27

5 Conclusions **35**

Bibliography **37**

A Appendix **39**

A.1 Collecting the results . . . 39
A.2 Dependencies between the results and the used parameters . . . 41
A.3 Finding the number of attraction and avoidance relations . . . 42

A.4 Finding the number of following relations 49

Chapter 1

Introduction

1.1 The need for mining soccer data

Due to the rise of digital techniques, such as GPS, but also the tremendous increase in the use of apps that collect tracking information, large amounts of movement data is being collected. Developments regarding tracking and sensor technologies in sports have led to large amounts of movement data within the field of sports as well. Performing spatio-temporal data mining on sports data has become of interest for several reasons. First of all, the data leads to the development of new spatio-temporal mining techniques and is therefore of interest for the acadamic world. There are several reasons why sports or soccer in specific is chosen for this type of research, instead of, for example, behaviour of animals.

First, sports like soccer are very complex since it involves a large numbers of players resulting in many dependencies and factors on which failure of success depends upon. Another reason is that sports in general, but soccer especially are very popular among a broad audience. Using sports as a topic of research can help promote the public's interest in science. Robocup is a good example of this. It's an international initiative promoting robotics and artificial intelligence by means of the Robocup project. The aim of the project is to develop a team of robots that can defeat the world champions in soccer by the year of 2050.

Another reason for spatio-temporal data mining research within sports is the commercial interest in the potentially useful information that can be found. Improving a player's or a team's tactics can earn many people a lot of money.

1.2 The problem to be solved

The main research question for this research was: "Does the shot on target player show deviant behaviour prior to a shot on target?" An answer to this question can be useful for soccer coaches and players as it can give them information on how to create goal scoring opportunities in order to potentially score more goals

and thus win matches. To answer this question it is necessary to classify the behaviour of the players. In this research this was done by distinguishing three types of behaviour: attraction, avoidance and following behaviour. These types of behaviour can tell us a lot about an individual player, but also about the way he behaves towards certain other players of his own or the opposing team. For example, which players are drawn to each other, but also which players try to stay away from each other. After defining the types of behaviour, several questions need to be answered. First, it is important to relate these types of behaviours to events in soccer matches. We will get into this in section 2.6.

Once we know what the behaviour entails in terms of soccer, we need to compare the behaviour of a player who made a certain shot on target with the behaviour of other players in order to see if his behaviour is deviant prior to a shot on target. Before we can perform this comparison, we need to find the right players to compare it with. The behaviour of these players can then be seen as the usual behaviour of a player who made a certain shot on target. Simply comparing the behaviour of the player who makes a shot on target with each other player on the field is not useful, because players are expected to show different behaviour based on their position in the field and the objective that comes with that position.

Next, several questions need answering. We need to know what the usual behaviour in terms of attraction, avoidance and following of the players we compare the player who made the shot on target looks like. We need to know how many of these relations they have and with which players they have them. Obviously, we also need to know what the behaviour of the player who made the shot on target looks like in terms of attraction, avoidance and following. We again want to know the number of players that he has each of these relations with, but also with who they are. Once we have all this information, we want to find whether there is a significant difference between the relations of the players responsible for the shot on targets and the other players we use for this comparison. If they indeed are significantly different, we need to determine what this deviant behaviour of the players making the shot on targets entails in terms of soccer.

1.3 Related work

A lot of research based on this type of data has been performed within the field of soccer. There are different approaches for performing data analysis within soccer. A common perspective is the analysis of passing sequences. In 2004, S. Hirano and S. Tsumoto [HT04] investigated passing sequences leading up to goals. They clustered the sequences based on the features they found in the sequences. By performing this clustering they tried to find pass patterns. Another research, conducted by M. Hughes and I. Franks [HF05], looks at the relation between the length of a passing sequence and the chance at scoring. They found that most sequences resulting in goals were of shorter length.

A completely different approach is that of Oliveira et al. in the paper "A datamining approach to solve the goal scoring problem" [OAC+13]. They developed a decision system for determining the right time and direction to shoot the ball towards the goal in order to enhance the chances of scoring. Also, research has been

conducted to finding the best place on the field to score from. In 2017 Smith and Lyons [SL17] used data from FIFA world cup championships between 2002 and 2014 to find zones in a soccerfield from where a significant number of goals where scored.

Other papers propose methods and approaches for spatio-temporal data analysis in general or specificly for sports data. In Stein's et al. [SJS+17] it is stated that "The challenge of analyzing team sport data is that movement is restricted by a pitch and rules, driven by the predetermined objective and influenced by the movement of own and opposing team players." This influencing factor of movement is further divided in three types.

1. A group influencing an individual and vice versa.
2. A group influencing a group.
3. One individual influencing one individual.

In this paper analysis is performed from the third perspective. Namely, how do soccer players behave towards each other prior to shot on targets. The research differs from previous researches in the manner in which this "individual influencing individual" perspective is used. Up to now, no research has been conducted into pairwise behaviour of players in terms of attraction, avoidance and following in particular.

In order to find these attraction, avoidance and following relations in the data a tool called Movemine [WLLH14] is used in this research. Movemine is mostly used for the analysis of relations and behaviour of animals, such as Friedemann's et al. [FLK+16] research into foraging behaviour during breeding amongst predators in which they used the attract/avoid function to look into the behaviour of birds.

Although no research has been conducted in the field of soccer or sports in general, using attraction, avoidance and following relations, other researches have shown that these relations can be used successfully in other domains as well. A good example is Cacho et al. [CMFE+16] paper on smart city planning. In the research, a mobile touristguide was used to collect data about tourists behaviour in a Brazilian city. The functions were used to analyse the data resulting from the mobile guide. For example, for the detection of places and times in the city where flocks occurred.

Chapter 2

Definitions

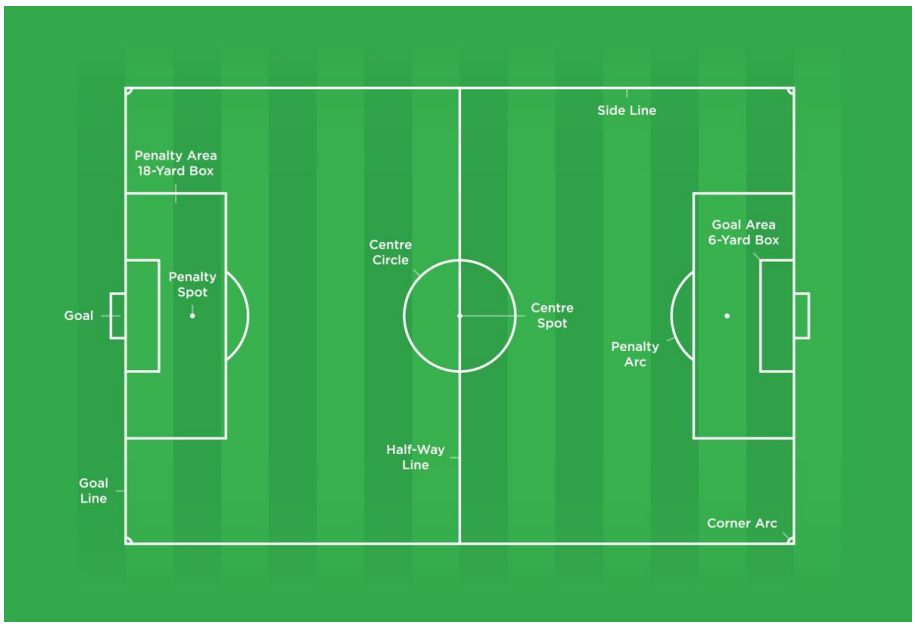
2.1 Spatio-temporal data mining

Spatio-temporal data mining is the mining of data that consists of geospatial and temporal components. Geospatial data is the location of the object in question, for example given by latitude and longitude coordinates. Temporal data gives the moments in time, often in the form of timestamps, on which an object found itself at a certain location. The objective of mining spatio-temporal data is to find previously unknown and potentially interesting patterns in the movement of the objects included in the data. These objects can be anything from animals to farmlands or in the case of this thesis soccer players.

Due to the growth in available movement data, the interest in data mining techniques suitable for this type of data is growing as well. This is because the application possibilities for spatio-temporal data are very broad and the data can contain interesting information. However, classical data mining approaches often perform poorly on spatio-temporal data. This is due to the fact that this data is often more complex and mostly contains continious data rather than the discrete data contained in classical datasets.

2.2 Soccer

Soccer is a popular team sport that involves two teams of eleven players each and a ball. The purpose of the game is to play the ball in the so-called goal of the opposing team, referred to as scoring. A match consists of two halves of forty-five minutes each. The team that scores the most goals is the winner of that game. In case both teams score an equal number of times then it is a draw.



The game is held on a rectangular grass field, often 68 meters wide and 105 meters long. The half-way line indicates the sides of each team. If a team plays on one side of the playing field than it has to score by playing the ball in the goal on the other side of the field.

The players are allowed to use any part of their body when playing the ball except for their hands and arms. The goalkeepers form an exception, they are allowed to use their hands in order to stop a ball from going in the goal.

The rules of the game require that a goalkeeper is assigned. For the rest of the players on the field assigning roles is not mandatory, but definitely common practise. There are roughly three types of players

1. Attackers or strikers: these players are closest to the goalkeeper and goal of the opposing team and their main task is to score goals.

2. Midfielders: these players are placed in between the attackers and the defenders and therefore can perform both attacking or defensive tasks.

3. Defenders or backs: these players are closest to their own goalkeeper and goal in order to prevent the opposing team from scoring.

The positions of the players can be subdivided further based on their horizontal placement on the field. A player can be placed on the right or the left side of the field or centered. A midfielder that plays on the right side of the field is then referred to as a right midfielder.

2.3 The dataset

The data was collected during the UEFA Women's European Championship of 2017. Figure 2.1 gives a graphical representation of the dataset. The data is divided in two types of datasets, namely event data and

position data. The upper four entities in figure 2.1 are from the position data and the lower two entities in figure 2.1 are from the event data. The event data mostly contains background information for each match and was not used very extensive in this research except for finding the size of the playingfield, the line-ups and the positions of the players. The position data however contained a whole lot more match specific information and formed the base of this research.

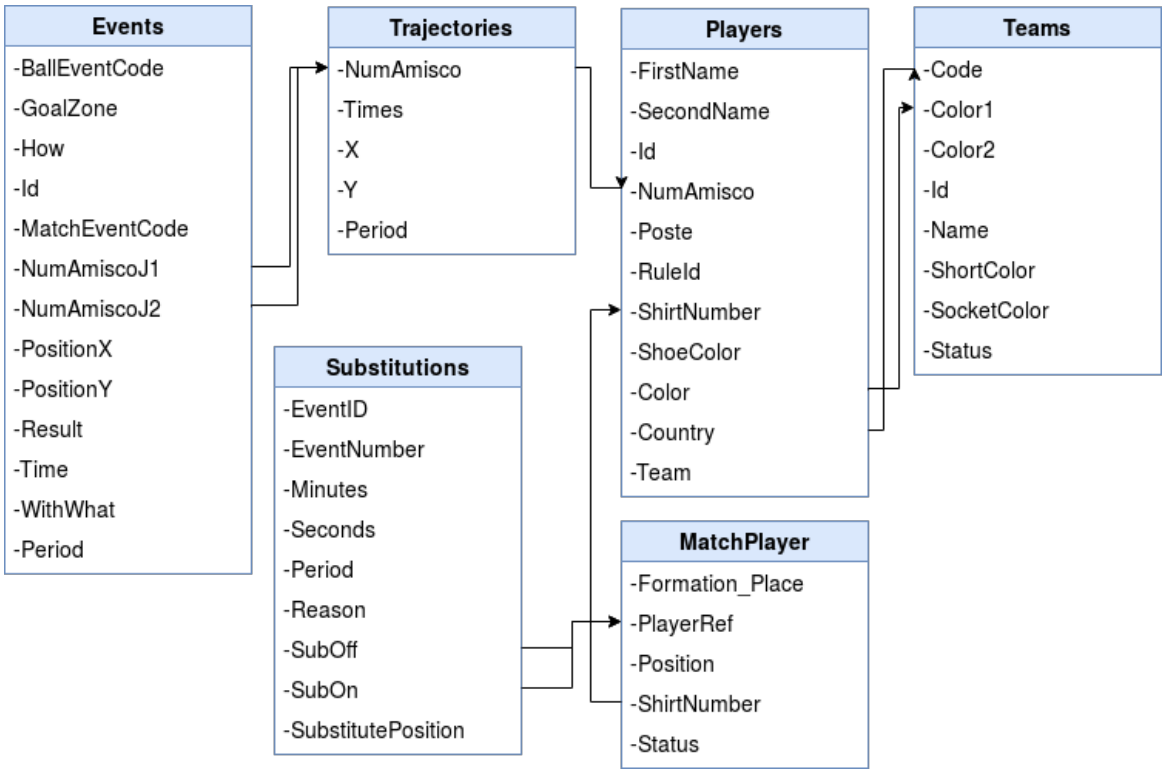


Figure 2.1: A graphical representation of the entities in the dataset that was used.

The position data set contained the coordinates for each player on the field as well as for the ball. The coordinates are relative to the playingfield with coordinate (0, 0) representing the centre spot of the field. Coordinates were measured with a rate of ten coordinates per second, making it possible to precisely track the path each player travelled in a match. Besides these so-called trajectories, the data also contained a list of events that occurred in a specific match. Examples of occurring events are ”Pass”, ”Reception”, ”Deep forward pass”, ”Cross” and ”Shot on target”. For each event, the data contains details such as which player was responsible for the event, a coordinate to indicate where the event took place and at what time it happened. Furthermore, the data contained some basic information about teams and players such as the t-shirt color of a team or the team a player belongs to.

In total the dataset contained data from six matches. All the attributes shown in figure 2.1 were collected for both the home and away team in each match. The general information about the players was collected for both the players that actually played as well as for the substitutes. The trajectory data was obviously only available for players that played during a certain match. In total 64 shot on targets were found in the data of all six

matches.

2.4 Interaction patterns between players in soccer

When thinking of interaction between soccer players several patterns can be found between both allies and opponents.

- Dodging
  A common event between opponents is dodging. For example, when the attacker of one of the teams has ball possession and starts attacking by taking the ball to the other side of the field in order to score. On his path to the other side of the field, players of the opposing team will try to stop him in his attack and regain possession of the ball. In order to prevent this from happening, the attacker will try to dodge his opponent.

- Spreading out
  Suppose when one of the teams initiated an attack and has succeeded in making it to the other side of the field while remaining in ball possession. The opposing team will try to prevent this and therefore have followed the attackers to their side of the field, resulting in a crowded situation. In order to deal with this, the players of the attacking team can spread out to make themselves more available for receiving passes and ultimately score.

- Allies coming up together
  Coming up together is the event in which two or more players of the same team decide to move to the same direction of the field at the same time. For example, if the attackers of one of the teams start attacking, it is common practice for the midfielders to follow their attackers shortly after the start of an attack in order to back them up.

- Opponents coming up together
  Opponents coming up together might seem illogical, because when a certain player is moving towards his opponent, the opponent will almost always most likely move away from him in order to either remain ball possession, be available for passes or for some other objective. However, events in which opponents are coming up together can occure due to coincidences. An example of such a coincidence can be that one of the teams started an attack and some midfielder of that same team starts heading towards the attackers of his team in order to back them up. Meanwhile, the defenders of the opposing team start heading towards the attackers as well in order to stop the attack.

- Covering
  In a covering event a player of one of the teams is trying to prevent a player of the opposing team to get near one of his allies or, in most cases, the goal. For example, when the attacker of one of the teams initiates an attack and makes it to the other side of the field, then one of the defenders of the opposing

team will try to prevent him from scoring. The defender can do so by chasing this attacker in order to steal the ball away from him, or block him from the goal.

- Backing
  In a backing event the player of one of the teams if following one of his allies in order to provide support to this teammate. For example, if an attacker of one of the teams starts an attack and moves to the opponents side of the field, then midfielders of the attacking team will follow this attacker in order to back him up and provide support.

2.5 Movemine

Movemine is a software tool designed to mine spatio-temporal data. Users can upload their data in the form of csv files. These files should contain a list of trajectories for a number of objects. The program requires these trajectories to contain a coordinate in the form of a latitude and a longitude, a timestamp formatted as Year-month-day Hours-minutes-seconds-miliseconds and a unique identifier ID for each object. Once the file is loaded the program offers several options.

Functions

The program offers four functions. The plotting function is disregarded here, since it was not used in this thesis.

- Distance calculation. This function creates a matrix with the average distance during the specified time interval for all selected objects. This average distance between two objects A en B is the pairwise Euclidean point distance normalized by the trajectory length of the objects. The formula used for this Euclidean distance is as follows.

$$Dist(A, B) = \frac{1}{n} \times \sum_{i=1,..,n} dist(A_i, B_i)$$

  Hereby dist() stands for the simple Euclidean distance between two locations.

- Attract/Avoid mining. [LDW+13] This function computes the significance value of the interactions (attraction/avoidance) between two objects. If this significance value approaches one, there is an attraction relation between the two. If it approaches zero, there is an avoidance relation between them. The significance value is computed by looking at the meeting frequency of two objects. The meeting frequency is the number of times the two objects meet, i.e., when they are spatially very close. However, determing the interaction relation based on this meeting frecuency alone is not enough. Instead, a number of permutations is constructed and for each the meeting frequency is determined. Then, if the actual meeting frequency is, for example, higher than 95% of the permuted frequencies, we state that there is a significant attraction relationship of 0.95 between the two objects. This approach is based on the assumption that if two movement sequences are independent and thus any meeting event between

them is random, then the meeting frequency between de random permutations and the actual meeting frequency should be similar. Otherwise the two objects are not independent.

- Following mining. [LWC13] This function determines whether there is a following relationship between two objects and the duration of that following relation. Also, within an interval there can be multiple following relations of different lengths. Whether one object is following an other object at some point depends on the spatial and temporal distance between the trajectories of the two objects. In other words, if object B is following object A, then B has to visit places close to the places A has visited and he has to visit them shortly after A has visited them. If this happens, we call this location pair (A_j, B_i) a following pair. All following pairs in an interval are found by means of an indicator function $f(i)$. This function takes the closest location of B to A and, if it is indeed close enough, looks at the difference in time between A en B. If B's time is later than A but still below a specified maximum $f(i)$, it will indicate that there is a following pair at time i. The indicator function f is performed for all timestamps within the interval and the number of following pairs is then counted. Next, the expected number of following pairs is subtracted from this actual number of following pairs. This expected number is computed under the assumption that if A and B are simply moving together, then the chances that a following pair occures at a timestamp i is 50%. The following score for an interval I is then defined as follows:

$$g(I) = f(I) - 0.5 \times |I|$$

Parameters

There are two types of parameters.

1. Parameters related to the uploaded file.
    - Object selection, this option allows the user to select the object it wants to compare with each other.
    - Start time and end time. These parameters determine the interval for trajectories to investigate.
    - Data interpolation parameters. These parameters consist of a *Gap* a *ThresGap*. The gap is specified in minutes and determines the sampling rate of the data after interpolation. For example, if Gap is set to one minute, then the program will use a sampling rate of 1 sample per 1 minute. The ThresGap is specified in hours and gives the constrains on a possible time gap between to samples. This means that if ThresGap is set to one hour, only locations with a time difference smaller than 1 hour will be used. This way the program can deal with missing data. Data interpolation is required for the distance and attract/avoid functions, but optional for the following function.
2. Parameters related to the choosen mining function.
    - Attract/avoid parameters. These parameters consist of *Rounds* and *Dist thres*. Rounds indicates the number of permutations to be computed to estimate significant values. This means the more rounds, the more accurate the estimation of the significance value. Dist thres specifies the distance constraint to determine whether two point are spatially close.

- Following parameters. These parameters consist of *Time thres*, *Dist thres* and *Min Interval Length*. Time thres specifies the time constraint to determine whether two points are temporally close. Dist thres specifies the distance to determine whether two points are spatially close. The minimum interval length is used for the visualization graph Movemine creates. If the duration of a following relation is shorter that the minimum specified length, then it will not be depicted in the graph.

2.6 Movemine functions and the football patterns they represent

Now we have explained how the attraction, avoidance and following relations of Movemine work, we can relate them to the interaction patterns discussed in section 2.4. Later on, this will enable us to relate the results to their meaning in terms of soccer.

- Avoidance

  Avoidance takes place when one player is trying to avoid contact or stay away from a certain other player. Both dodging behaviour and spreading out behaviour comply with the definition of avoidance. In the case of dodging behaviour, the player that is trying to dodge some other players will have an avoidance relation with that player. In the case of spreading out behaviour, the allies that are spreading out will have avoidance relations with each other as a result of the distance that they are trying to create between one another.

- Attraction

  An attraction event takes place when two players move along approximately equal paths at the same time. These paths do not have to be exactly the same, but they should be parallel to each other and have a certain maximum distance between them. The main difference between attraction and following is that in an attraction event two players are moving along similar paths at the same time. As for a following event, one player is moving along a similar path of the player he is following, a few moments later after the player he is following took that path. The event of coming up together between both allies and opponents complies with the definition of attraction. When two players are coming up together they are moving towards the same direction of the field at the same time, causing attraction relations to occur between those allies. In the event of two opponents coming up together, attraction relations occure for the same reason. However, as mentioned before about opponents coming up together, the occurence of this event is rather a coincidence than on purpose. Attraction relations between opponents therefore do not indicate that these players are seeking each other out as is the case for attraction relations between allies.

- Following

  As already briefly explained above, a following event in soccer occurs when a certain player is following a path very similar to the path of some other player, shortly after this other player took that path. We say that the player following the path last has a following relation with the other player, but not the other way around. Both covering and backing behaviour comply with the definition of following behaviour. In

the case of covering behaviour, the player that is following one of his opponents has a following relation with that opponent. This is caused by the fact that the opponent is trying to move away from the player that is blocking him, while the player that is showing the blocking behaviour keeps chasing him. In the case of backing behaviour, the player that is chasing one of his allies in order to support him in his attack, has a following relation with that ally. The reason we don't speak about attraction behaviour between these allies is not because one teammate tries to avoid the other teammate, but because the player that is following his teammate walked the same path as that teammate, but only after that teammate did so.

Chapter 3

Data preprocessing

3.1 Extracting trajectories

In order to analyse the pairwise relations between players prior to shot on targets, the trajectories of all players have to be collected first. To do so, a time interval has to be determined to indicate up to what point coordinates would be collected. Two time intervals were chosen here in order to compare their output later on in the process. The first interval is set to 30 seconds and the second to 60 seconds. We chose an interval of 30 seconds, because within 30 seconds there is still a substantial amount of relations that is related to the attack leading up to the shot on target. Whereas for a time interval of 3 minutes, we could almost be certain that a large part of the relations occurring in that interval would not be related to (the attack prior to) the shot on target at all. The time interval of 60 seconds is chosen for exploratory purposes in order to see whether the relations that we find in the time interval of 30 seconds are also occurring 30 seconds prior to that.

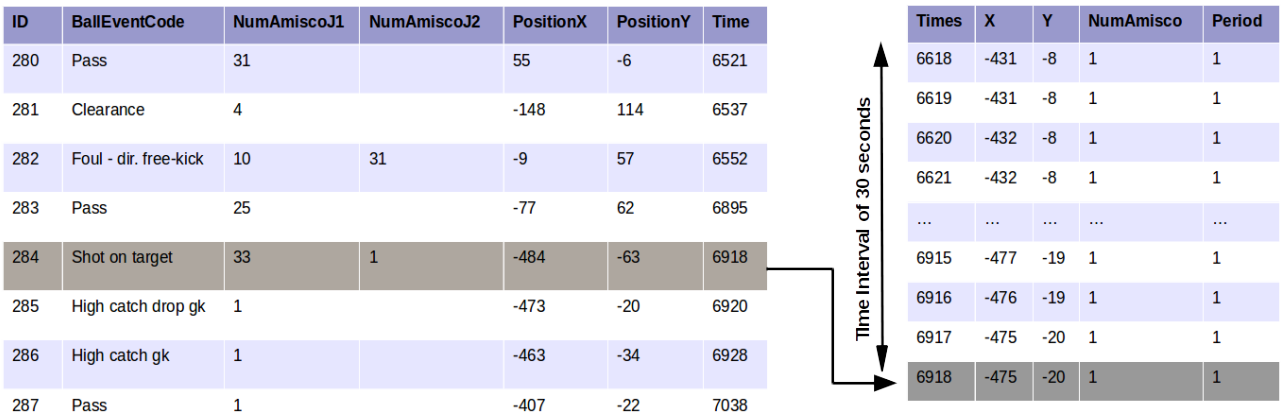


Figure 3.1: A graphical representation of the process of extracting the trajectories of the soccer players prior to shot on targets that were found in the data.

Next, all the events of type "shot on target" have to be located in the data. For each event, several properties

are specified, among which is a timestamp x. The timestamp x of a found "shot on target"-event is then used to collect coordinates from the trajectory data. Namely, for all players in a trajectory data file the coordinates with a timestamp between $(x - timeinterval)$ to x are extracted and placed in a new file. An example of a trajectory data file is shown on the right-hand side of figure 3.1. The player in this file is identified by the column "NumAmisco". The trajectory prior to this specific shot on target for player 1 is thus made up of all coordinates within the time interval shown.

In order to compare the pairwise relations of the players later on, a more general identification than an ID is needed. For this purpose, all players get a position assigned to them. In the original data set the positions of the players were already given. However, the position types only distinguish between goalkeeper, defender, midfielder and striker. In order to distinct between different types of attackers, midfielders and defenders as well, the players are assigned a more detailed position based on the already given position and the provided lineups. Figure 3.2 shows an example of the way this was done. All position names are abbreviated both in figure 3.2b as in the results.

(a) The position names for a 4-2-3-1 lineup before specification.

(b) The position names for a 4-2-3-1 lineup after specification.

Figure 3.2

These abbreviations are made up of three components.

- A player's horizontal placement on the field.
- A player's vertical placement on the field, i.e., their original position.
- In case of a center position, whether the player is left, right or central to the center itself.

The last one may seem unnecessary, but is needed to make a distinction between all players within the same team in order to perform pairwise comparisons for all possible combinations of players. Based on this structure,

player 7 in figure 3.2b gets the position:

Center	Midfielder	Left
C	M	L

The table in figure 3.3 shows an overview of all used abbreviations and the positions they correspond with.

Player's abbreviation	**Player's position**
G	Goalkeeper
LB	Left back
CBL	Center back left
CBC	Center back center
CBR	Center back right
RB	Right back
LM	Left midfielder
CML	Center midfielder left
CMC	Center midfielder center
CMR	Center midfielder right
RM	Right midfielder
LS	Left striker
CSL	Center striker left
CSC	Center striker center
CSR	Center striker right
RS	Right striker

Figure 3.3: All used abbreviations and the corresponding player positions.

At the end of this fase their are two files for each shot on target in each match. One of the files contains the trajectory data of all the players 30 seconds prior to a shot on target and the other file contains the trajectory data of all the players 60 seconds prior to a shot on target.

3.2 Performing Movemine's functions

Movemine offers a user interface that lets its users upload a single file containing trajectory data of at least two objects. After uploading a file a few more steps have to be taken.

1. **Object selection**

   After uploading a file, Movemine identifies and collects all the objects in the data based on the object ID's for each trajectory. Here, the detailed position name of a player combined with the name of that player's team suffices to uniquely identify each player. When Movemine found all the objects in the data, in this case all the players, it lets the user select the objects that it wants to compare relationships for.

2. **Setting the start and end time**

   In case the user only wants part of the trajectories to be compared, it can use these parameters to set a start and ending time. In the case of this thesis this function was not used, since the shot on targets and specified time intervals demarcated the interval to investigate.

3. **Data interpolation**

   Movemine also provides a function to interpolate the uploaded data in order to deal with missing data, as explained in section 2.5. As mentioned before, interpolation is required for the distance calculation and attract/avoid mining functions and therefore also is applied to the results in this thesis. The minimum value for the gap parameter of 10 minutes creates a a problem for the initial dataset, since the collected trajectories only have a length of 30 and 60 seconds. In this case, Movemine would take only one coordinate out of the trajectory of each player, making the execution of the functions completely useless. For this reason, each tenth of a seconds is converted to 10 minutes in the dataset, in order to make Movemine use all the coordinates in a trajectory.

4. **Mining functions**

   The next and most important step is to select the desired mining function. In this thesis the functions for distance calculations, attract/avoid mining and following mining were used.

5. **Function specific parameters**

   Based on the chosen function, certain parameters have to be set before Movemine can execute them. In this thesis the functions were performed using several parameter settings in order to compare the differences in their output later on. We will briefly sum up the choice of parameters used in this thesis as the purpose of the parameters has already been elaborated on in 2.5.

   Distance calculation

   This function does not require any parameters.

   Attract/avoid mining

   The number of rounds to be used for the permutations has to be a multiple of hundred with a minimum of one hundred and a maximum of a thousand rounds. In all results regarding attract and avoidance relations, this parameter was set to thousend rounds for accuracy reasons. The more rounds the more accurate the results.

   Determining the distance threshold was slightly more difficult. After trying out different distance thresholds, it was found that the lower the distance threshold, the less relations there are and the higher the distance threshold, the more relations there are. Therefore, we have used distance thresholds of 5, 10 and 15 meters. This way we can compare the output of the different thresholds later on in the results. These values are chosen based on the size of the soccer pitch. The soccer fields used in the matches had a length of 105 meters. The players of both teams are positioned across the area in between the goals. This area has a length of approximately 94 meters. Initially, the players of a team are spread out over their half of the field, which has a length of around 47

meters, in roughly three groups, the defenders, the midfielders and the attackers. From this we can conclude that in their starting position, teammates have an average distance between each other of around 23.5 meters. Once the game starts, these distances will of course fluctuate, based on the positions of all the players and the objectives that come with those positions. However, the distance thresholds of 5, 10 and 15 meters overall seem to be suitable values based on the size of the field.

Following mining

For the time threshold this function requires, a value of 5 seconds is chosen. This value seemed to be a good guideline. After performing the function for several time thresholds and analyzing the results, it indeed appeared that a time threshold above 5 seconds was to high for almost all the player combinations to have a following relation.

For the distance threshold, the same values have been chosen as for the attract/avoid mining function. Again, this seemed the most reasonable value based on the average player distances and the size of the pitch, but also to make it easier to compare the results of the following mining function with those of the attract/avoid mining function.

For each pair of players Movemine, creates a csv file containing all following relations between those two players. Besides the csv file, it also creates a graphical representation of these following relations in the form of a diagram. The minimum interval length is a value in seconds that specifies how many seconds a following relations must take, before it may be included in this diagram. Therefore, this parameter is not that relevant, since only the csv files were used in this thesis.

3.3 Collecting the results

After running the functions in Movemine, Movemine has produced one csv file for both the distance calculation and the attract/avoid mining function for each shot on target and all different combinations of parameter settings. However, for the following mining function it produces a csv file for each combination of players. All these files together result in a structure that is very unclear and not suitable for analysis. To give an idea, the structure is depicted in figure A.1 in the appendix. For this reason, certain parts of the data have to be combined in this fase of preprocessing. In order to join results, all output with the same parameter settings is combined in a large table. Meaning, all the data below the "DistThres_xm" files in figure A.1 is now all combined in one table. An example of such a table is shown in figure A.2 of the appendix.

Chapter 4

Results

After generating the tables for each of the relations and all combinations of parameters, a list of the players who made the shot on targets was drawn. Based on the resulting tables and this list of players, several graphs were generated. These graphs were created using a Python library called plotly [Inc15]. In these graphs, the behaviour of the player who made the shot on target was isolated from the data of the other players in order to see if she behaves differently than the other players. The graphs in section 4.3 show the overall attraction, avoidance and following relations of the player who made the shot on target. In the graphs, we make a distinction between teammates and opponents in order to compare the behaviour that the player responsible for the shot on target has with both of them. These graphs were made for all used combinations of time intervals and distance thresholds in order to look at the influence of these parameters on the results in section 4.1. In section 4.4, the relations of the player who made the shot on target are worked out in more detail in order to see with which other players she has relations and what these relations look like. In the rest of this chapter, we will refer to the player who made the shot on target as the shot on target player.

4.1 Dependencies between the results and the used parameters

As mentioned in chapter 3, the analysis was performed for two time intervals, 30 seconds and 60 seconds, in combination with three different distance thresholds of 5, 10 and 15 meters. Before looking at the results, some remarks have to be made about the dependencies between these parameters and the results.

The time interval

In the results, we found that the number of relations of a certain player increases when the time interval goes up. This is actually quite obvious, because all players have 30 seconds more in which they can have relations with other players if the time interval increases to 60 seconds. If the number would not increase, this would mean that there are no relations in the first 30 seconds prior to a shot on target. This would be rather suprising, since there are plenty of relations occurring in the last 30 seconds prior to a shot on target.

The distance threshold

In all the results, we found that the higher the distance threshold, the more attraction and following relations there are. This is a consequence of the fact that the distance threshold forms a constraint in determining whether two players are spatially close. If the distance threshold becomes larger, this constraint becomes weaker and therefore more players will have a following or attraction relation. This positive effect of the distance threshold on the number of relations however is slightly different for the avoidance relations. This has to do with the fact that the distance threshold has the opposite meaning when using it to find avoidance relations. Namely, for two players to have an attraction relation, they need to be within a distance from each other that is lower than the distance threshold. On the other hand, two players that have an avoidance relation, need to be within a distance from each other that is higher than the distance threshold. With this definition one might think that increasing the distance threshold should always lead to less avoidance relations instead of more relations. This is also not the case, because the actual number of found avoidance relations is normalized by an expected number of avoidance relations as explained in section 2.5. So, although the total number of found avoidance relations might go down when increasing the distance threshold, it is still significantly higher than what would be expected for that given threshold. For example, figure 4.10 shows us that the number of avoidance relations increases when the distance threshold increases and the time interval is set to 30 seconds. However, when we look at the number of avoidance relations for a time interval of 60 seconds this is not true anymore. In fact, the number of avoidance relations with opponents decreases when the threshold increases.

Despite the fact that an increasing distance threshold leads to an increase in the number of attraction and following relations as well, the distribution of the number of relations that a certain player has with each other player roughly remains the same. This does not completely hold for an increase in the time interval. This has to do with the fact that in the first 30 seconds of an attack different relations occure than in the last 30 seconds of an attack. To illustrate this, we display the number of attraction an following relations the right striker has with each of his opponents in figures 4.1 and 4.2. The number of relations is the total number of all shot on targets that were found in all six matches. The right striker is the striker on the attacking team, i.e., the striker that is on the shot on target player's team. The upper halves of both graphs contain an outline of the lower halves of the graphs in order to make it easier to see that for all three distance thresholds, the distribution of the relations that the striker has with each player follows the same trend.

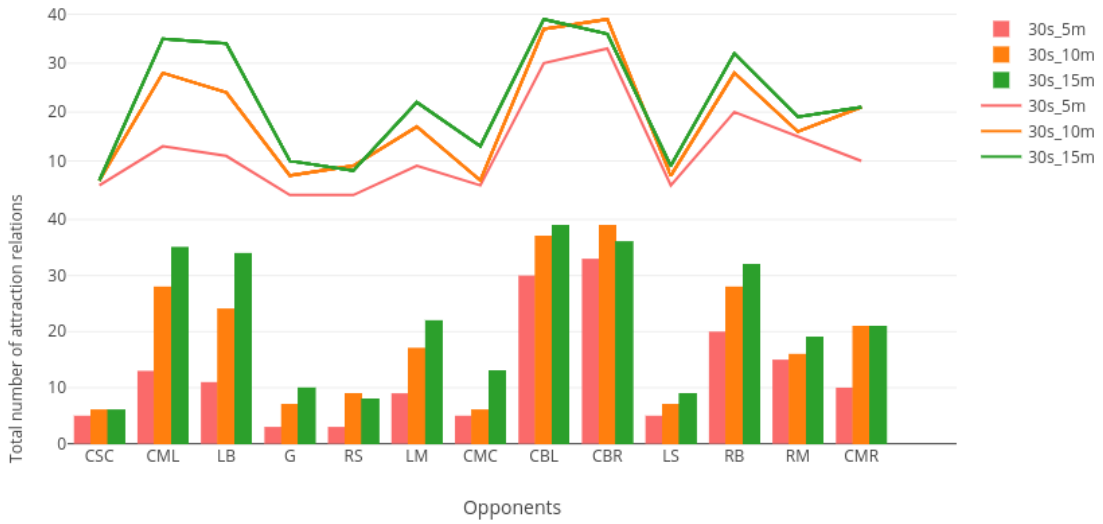


Figure 4.1: The total number of attraction relations that the right striker on the attacking team has with his opponents, for all distance thresholds and a time interval of 30 seconds.

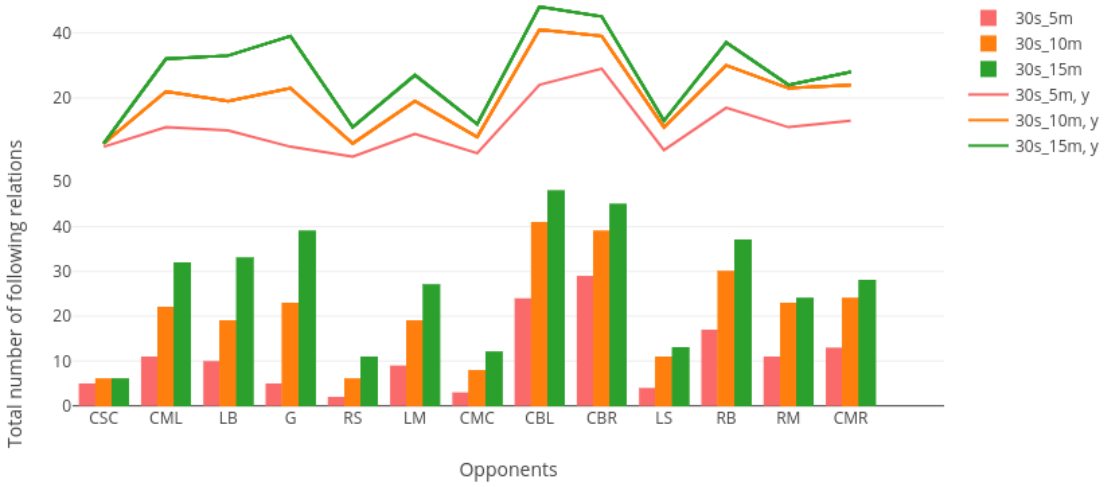


Figure 4.2: The total number of following relations that the right striker on the attacking team has with his opponents, for all distance thresholds and a time interval of 30 seconds.

The graphs in figures 4.1 and 4.2 were also generated for the relations between teammates in combination with a time interval of 60 seconds. They have been placed in the apendix in section A.2 for the reader to verify that the distribution also remains the same for teammates and a time interval of 60 seconds.

4.2 Finding the number of relations

Before looking at the overall and the detailed behaviour of the shot on target player, it is good to explain how the results were used to find the total number of attraction, avoidance and following relations for a certain player.

Attraction and Avoidance

As explained in section 2.5, the attract/avoid mining function returns a significance value for each pair of players. The significance value is a number between zero and one. The closer this value is to zero, the more likely two players have an avoidance relation. The closer this value is to one, the more likely two players have an attraction relation. In order to see what the distribution of the significance value looks like, several boxplots were created. The boxplots contain the significance values of three types of players. Namely, the shot on target player and the left midfielder and right striker that are on the shot on target player's team. The reason for choosing these specific players is because these are also the players that are going to be compared in section 4.4. Next, for each of these players all significance values of all the relations that they have with other players are collected. Within these significance values a distinction between the relations with the teammates and the opponents is made. This is done, because the relations that a player has with his teammates and the relations that she has with her opponents, are expected to differ more than relations among teammates of relations among opponents. We created the boxplots for all used combinations of distance thresholds and time intervals. Two will be discussed here and the other four are placed in the appendix in section A.3 for the reader to verify.

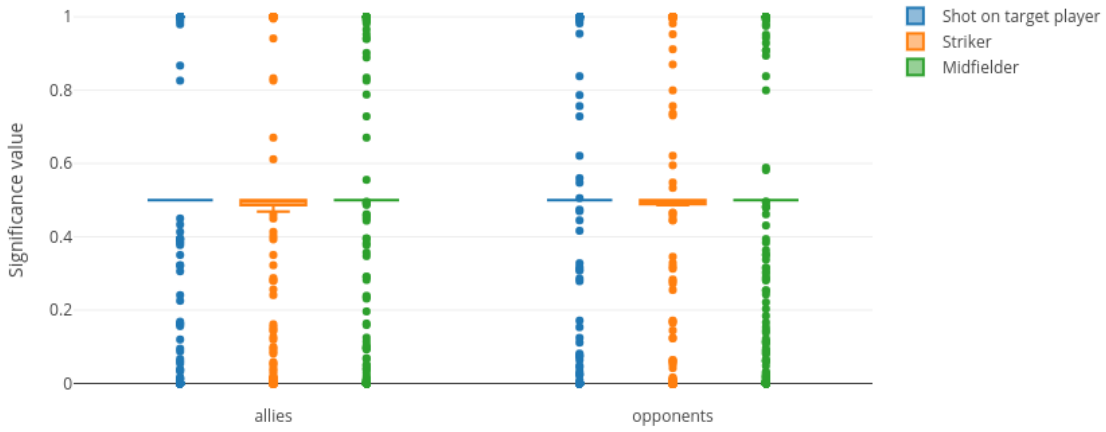


Figure 4.3: The significance values of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 5 meters together with a time interval of 30 seconds.

Figure 4.3 contains two boxplots for each of the three types of players mentioned above. The boxplots on the left contain the significance values that the three types of players have with all teammates and the boxplots on

the right contain the significance values that the three types of players have with all their opponents. Figure 4.3 tells us that a significance value of 0.5 is by far the most occurring value for alle three types of players. If we were to take the means of these significance values, we would get an average value of 0.5 which tells us that there is neither attraction nor avoidance. However, we also see a lot of outliers for all three players that do implicate attraction and avoidance relations. By using the mean of the significance values all this information would be lost.

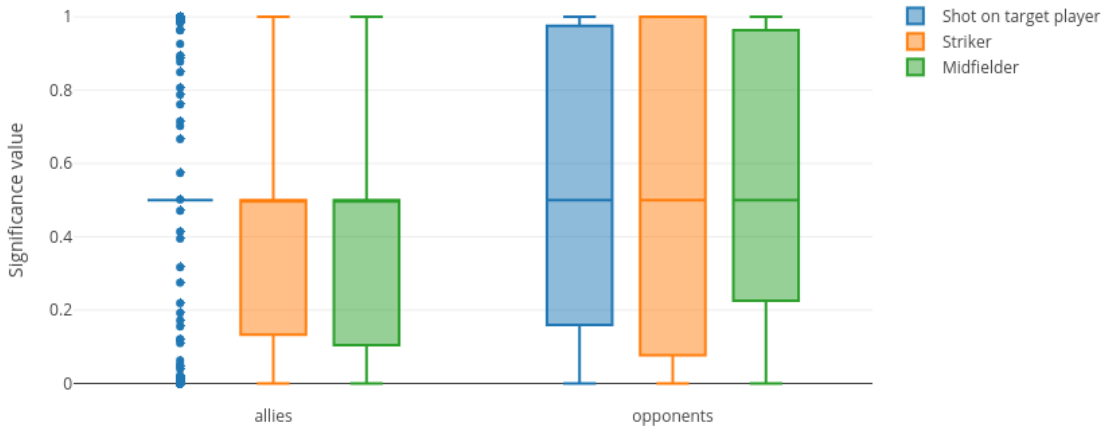


Figure 4.4: The significance values of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 10 meters together with a time interval of 30 seconds.

The boxplots in figure 4.4 already look a lot different from the boxplots in figure 4.3. The boxes that contain the middle 50% of all significance values are a lot higher. These larger boxes show that the number of significance values that are higher and lower than 0.5 increases when the distance threshold increases from 5 meters to 10 meters. For both the striker and the midfielder, it holds that 75% of the significance values of the relations that they have with their teammates is between 0.0 and 0.5 and the other 25% is above 0.5. This indicates that these two players have more avoidance relations with their teammates than attraction relations. However, it again holds that if we would take the average of these significance values, we would lose all the information about the attraction relations. This is due to the fact that the average would be 0.5 or smaller because of the high number of lower significance values that will push the average in their direction. When looking at the boxplots of the relations that the players have with their opponents, the significance values are more equally spread out between 0 and 1. The players have more attraction relations with their opponents, than they had with their teammates, as is shown in the boxplots on the left of the graph. Although, these boxplots again look different from the ones already discussed, it again holds that taking the averages of these values would lead to information loss in terms of attraction and avoidance relations. This is due to the fact that the range of the boxplot again goes from the lowest to the highest possible significance value.

In order to find the number of attraction and avoidance relations, significance values above or below a certain threshold are counted. If the significance value for a certain pair of players is lower than or equal to 0.3, it is seen as an avoidance relation. If the significance value for a certain pair of players if higher than or equal to 0.7, it is seen as an attraction relation. However, a significance value of 0.9 indicates a stronger attraction relation than a significance value of 0.7. Therefore, we want to see whether their is a big difference in the strength of the attraction and avoidance relations of the players that we are going to compare in section 4.4. Therefore, new boxplots were created from the boxplots in figure 4.4. Figure 4.5 contains all the significance values from the boxplots in figure 4.4 that are higher than or equal to 0.7. Figure 4.6 contains all the significance values from the boxplots in figure 4.4 that are lower than or equal to 0.3.

Figure 4.5 shows that the vast majority of the significance values that indicate an attraction relations are 1. There are some outliers that indicate slightly less stronger attraction relations, but overall the attraction relations of all three players with both their teammates and opponents are equally strong.

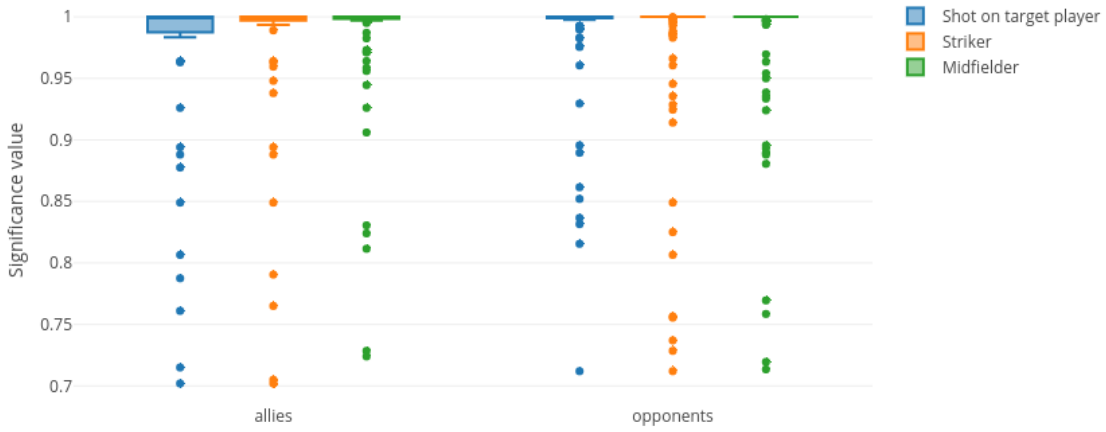


Figure 4.5: The significance values that are ≥ 0.7 of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 10 meters together with a time interval of 30 seconds.

In figure 4.6 te vast majority of the lower significance values have a value of 0. The boxplots again contain outliers that still indicate avoidance relations, but are of less strength then the majority. Overall, all three types of players have equally strong avoidance relations with both their teammates and their opponents.

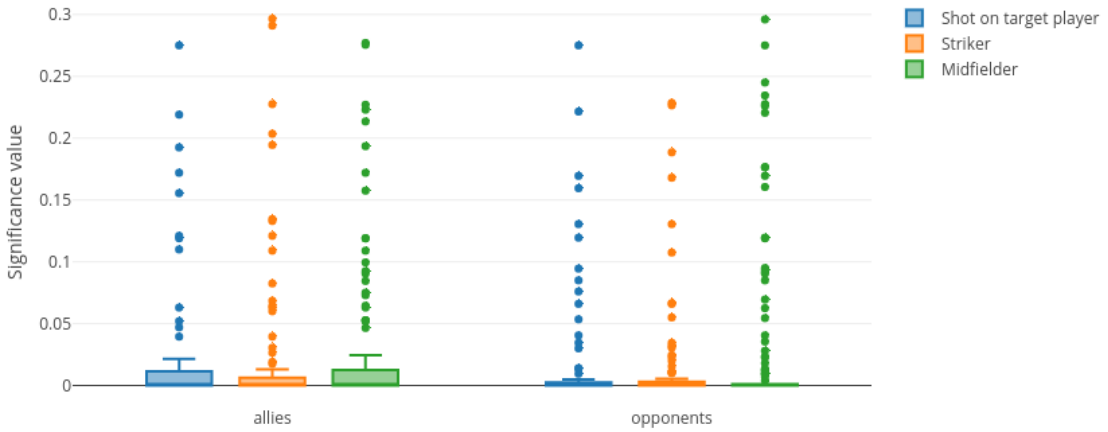


Figure 4.6: The significance values that are ≤ 0.3, of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 10 meters together with a time interval of 30 seconds.

The boxplots in figures 4.5 and 4.6 showed that all three types of players have equally strong attraction and avoidance relations for a distance threshold of 10 meters and a time interval of 30 seconds. The attraction and avoidance relations of these players remain equally strong for all other combinations of distance thresholds and time intervals as well. The boxplots corresponding with these combinations are placed in the appendix in section A.3 for the reader to verify.

Following

As mentioned in section 2.5, the output of the following mining function is quite different from the attract/avoid mining function. Whereas the attract/avoid mining function gives a single number back for each pair of players, the following mining function returns an array for each pair instead. This array contains a list of all the times that those two players are following each other within the specified time interval. Each following record in the array contains the duration of that following relation in seconds. In order to get an idea of the distribution of these durations, the durations of all following relations have been placed in the boxplots in figures 4.7 and 4.8. The values in these boxplots were obtained in the same way as for the significance values in the boxplots above, meaning we again used all the following relations found in all shot on targets throughout the six matches that were analyzed. The types of players used in the boxplots are also the same as before, namely the shot on target player, the right striker and the left midfielder.

Despite the outliers in figure 4.7 that indicate that there are some following relations lasting longer than 10 seconds, the vast majority of the following relations fluctuate around 5 seconds. This holds for the following relations of all three types of players, with a distance threshold of 10 meters and a time interval of 30 seconds.

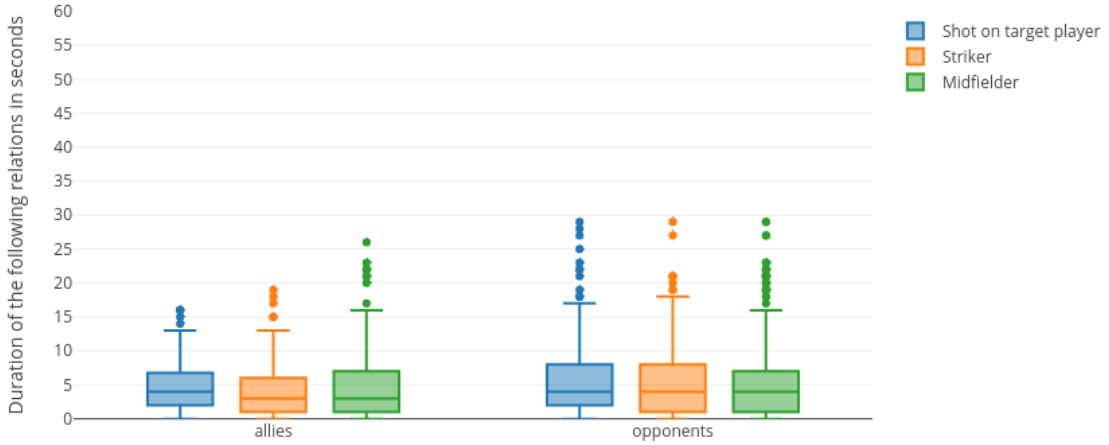


Figure 4.7: The durations of the following relations of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 10 meters together with a time interval of 30 seconds.

The boxplots in Figure 4.8 show that for all three types of players the vast majority of their following relations fluctuates around 5 seconds for a distance threshold of 10 meters and a time interval of 60 seconds as well.

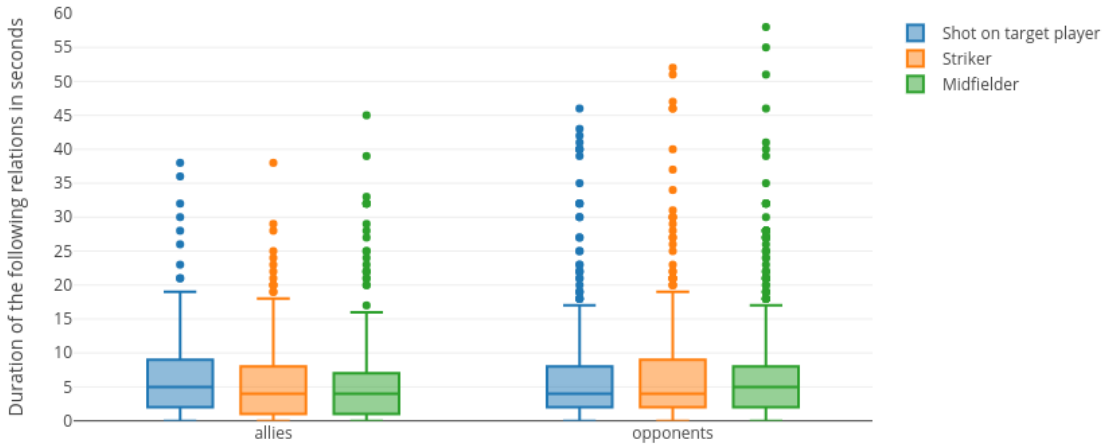


Figure 4.8: The durations of the following relations of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 10 meters together with a time interval of 60 seconds.

The durations of the following relations for the other parameter combinations used in the research fluctuate around 5 seconds as well. The boxplots corresponding with these following relations are placed in the appendix in section A.4.

To compare players in section 4.4, we simply want to know whether they have a following relation with each of their teammates and opponents or not. Therefore, we want to use the arrays for a binary classification whether a pair of players has a following relations. To this end, we count the number of following arrays that contain at least one following trajectory that lasts at least 5 seconds. The threshold of 5 seconds is based on the chosen time threshold that determines whether to players are temporally close, as explained in section 3.2.

4.3 Overall behaviour of the shot on target player

Attraction

Figure 4.9 shows that the shot on target player has more attraction relations with her opponents than with her teammates. As already mentioned in section 2.6, drawing conclusions from attraction relations between opponents is a little bit difficult, because it is opposite to the definition of attraction in terms of soccer. Despite this, seeing a lot of attraction relations occurring between the shot on target player and her opponents prior to a shot on target makes sense, since the opposing midfielders and defenders are most likely to move in the same direction as the shot on target player in order to prevent the shot on target. However, we should not really see them as attraction here since the shot on target player is not on purposely seeking the opposing midfielders and defenders out. It is more likely that the attraction between them is caused by the fact that the shot on target player is unable to avoid them because of the resulting crowded situation prior to a shot on target.

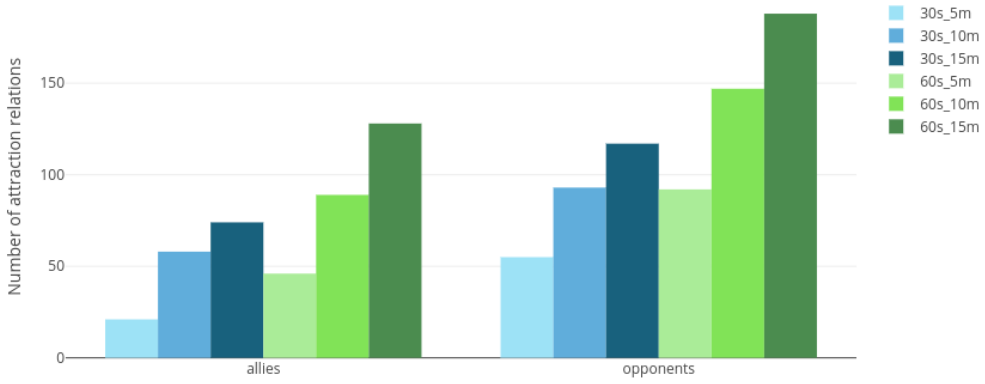


Figure 4.9: The total number of attraction relations that the shot on target player has with her teammates and her opponents.

Avoidance

Figure 4.10 shows that the shot on target player has slightly more avoidance relations with her opponents than with her teammates. This is mostly the case for distance thresholds of 5 and 10 meters, because as mentioned in section 4.1, an increase in the distance threshold causes the number of avoidance relations to decrease

slightly. Observing that the shot on target player has more avoidance relations with her opponents than with her teammates, indicates dodging behaviour prior to a shot on target. This is what we expected beforehand, since the shot on target player heads in the direction of the goal while her opponents try to block him or steal the ball, resulting in the need for the shot on target player to dodge or avoid these opponents.

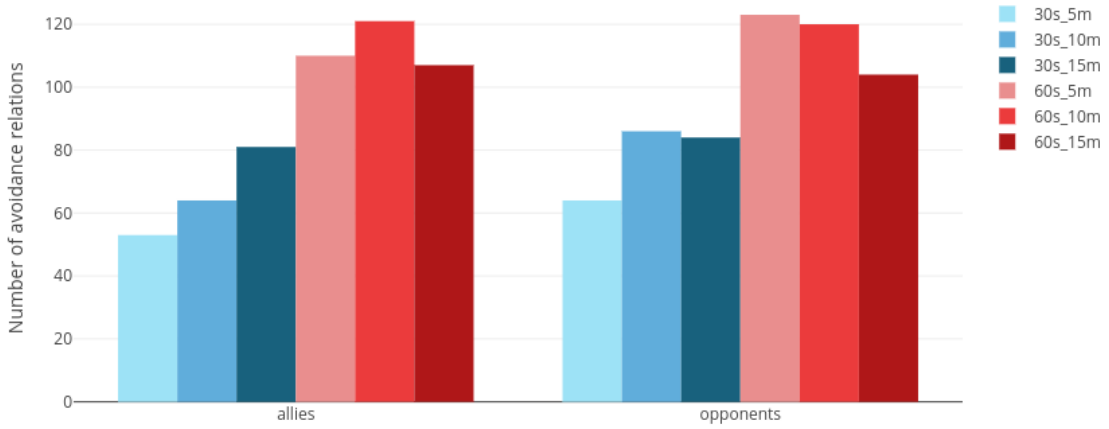


Figure 4.10: The total number of avoidance relations that the shot on target player has with her teammates and her opponents.

Following

When looking at the following relations of the shot on target player in figure 4.11, we again see that the shot on target player has more following relations with her opponents than with her teammates. In section 2.6, we mentioned backing behaviour as an example of following relations between teammates prior to a shot on target. This supporting behaviour occures when certain players follow their teammates during an attack. However, this backing behaviour also causes the players that follow their teammates, to have following relations with their opponents as well. This is caused by the fact that the opponents are following the attackers in order to prevent them from scoring. In doing so, the opponents will approximately take the same path as the attackers that they are following. Players that are on the same team as the attackers and are following the attackers during their backing behaviour, take approximately the same path as the attackers as well. These similar paths results in following behaviour between opponents and explains the higher number of following relations with opponents in this graph.

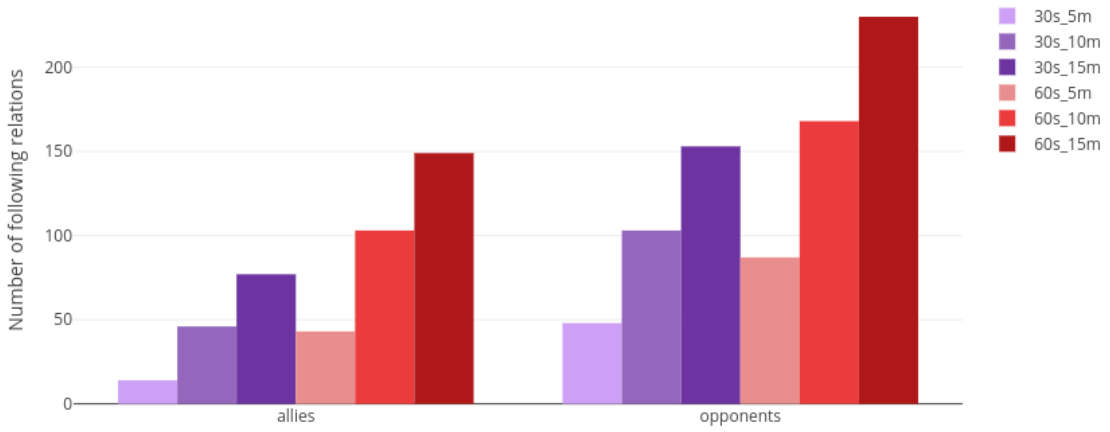


Figure 4.11: The total number of following relations that the shot on target player has with her teammates and her opponents.

4.4 Detailed behaviour of the shot on target player

In section 4.3, we showed that the shot on target player indeed behaves as expected prior to a shot on target in terms of attraction, following and avoidance relations. However, the most important question remains, namely: does the player who makes the shot on target show deviant behaviour prior to the shot on target? In order to look into this, the number of attract, avoid and following relations that the shot on target player has with each other player were calculated and placed in the graphs below. However, based on this information alone we can not say anything about whether her behaviour is deviant or not. For this comparison, we also need to determine the behaviour of certain other players. To determine which players to use for this, the types of players responsible for the shot on targets were found.

Figure 4.12 shows by which types of players the shot on targets were made. It was found that 57.8% of the shot on targets were made by a striker and 35.9% by a midfielder, leaving only 6.3% of shot on targets made by a defender. This means that in the majority of the shot on targets, the shot on target player is a striker and in a smaller part of the shots she is a midfielder. Therefore, we can expect to see the behaviour of the shot on target player to look like that of the strikers, but also with a slight deviation towards the behaviour of the midfielders. If this is indeed what her behaviour looks like, then it is not deviant. To look into this, the relations of the strikers and the midfielders are incorporated in the graphs in this section as well. For the behaviour of the strikers, the relations of the right striker were used, since the right striker made most shot on targets of all strikers. In some matches no right striker was found, due to the formation that was used. In that case the data of the central striker was used. This is because when the formation does not contain a right striker, then it also does not contain a left striker but only a centre one. For the behaviour of the midfielders, the relations of the

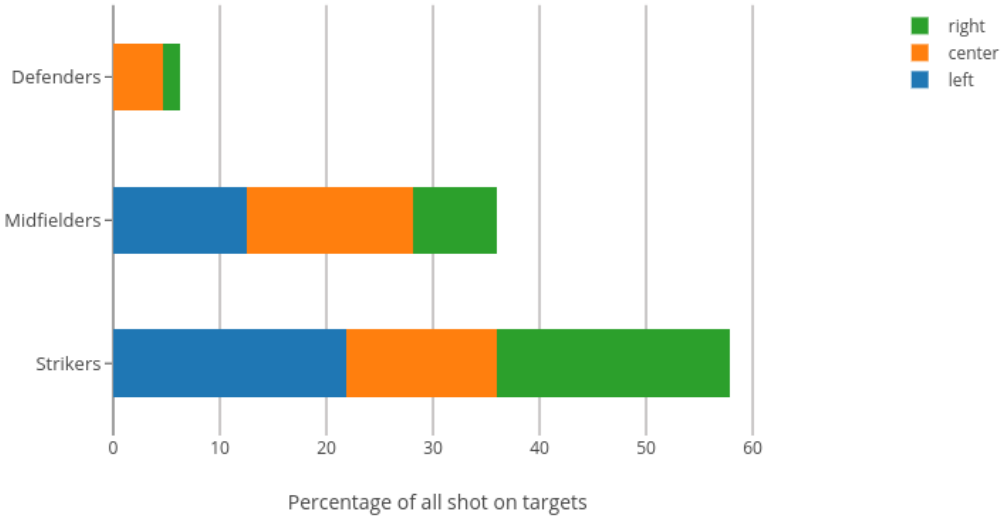


Figure 4.12: The types of players responsible for the shot on targets.

left midfielder were used, although the centre midfielder was responsible for more shot on targets. This is done because the difference in shot on targets made by centre and left midfielders is really small, but most importantly in order to have both left- and right-sided behaviour incorporated in the graphs. Also, because the right striker and left midfielder were used in the graphs below as the 'striker' and 'midfielder', the graphs do not contain relations between the striker and the right striker as well as the midfielder and the left midfielder when looking at teammates. The data used in the graphs below was based on the results for a time interval of 30 seconds and a distance threshold of 10 meters. A time interval of 30 seconds was chosen in order to see the relations that occurred most prior to the shot on target instead of early in the beginning of an attack. The choice for a distance threshold of 10 meters was based on the observation that increasing thresholds causes the number of relations to increase as wel. Although this is not necessarily a bad thing, we chose to slightly limit this effect by chosing the distance threshold of 10 meters, which still contains some of this influence but not as much as with a distance threshold of 15 meters.

Furthermore, each graph is accompanied by a table showing the probability values p resulting from the t-tests that were done in order to see whether the shot on target player's relations are significantly different from the midfielder's and striker's relations. These unpaired t-tests were performed using STHDA's t-test tool [STH]. If the p-value is lower than 0.05, the null hypothesis, stating that there is no significant difference between the relations of two players, is rejected.

Attraction

As mentioned, we would expect the behaviour of the shot on target player to closest resemble that of the striker but also to inherit some behaviour of the midfielder. In terms of the graphs, this would mean we would expect to see the number of relations of the shot on target player to be somewhere in between the number of relations of the midfielder and the striker, but more closely to the number of the striker. Suprisingly, figure 4.13 immediately shows that the behaviour of the shot on target player, in terms of attraction with her teammates,

differs from both the striker and the midfielder. The p-values in table 4.1 verify that the shot on target player's attraction relations with her teammates are indeed significantly different from those of the midfielder and the striker, since both probabilities are below 0.05. For the relations with almost all players, except for those with the central and left striker, the shot on target player has less attraction relations than both the midfielder and the striker. This indicates that unlike the average midfielder and striker, a shot on target player does not or hardly come up with her teammates prior to a shot on target. This suggests that the shot on target player is ahead of all her teammates prior to a shot on target, indicating that she is the one leading the attack.

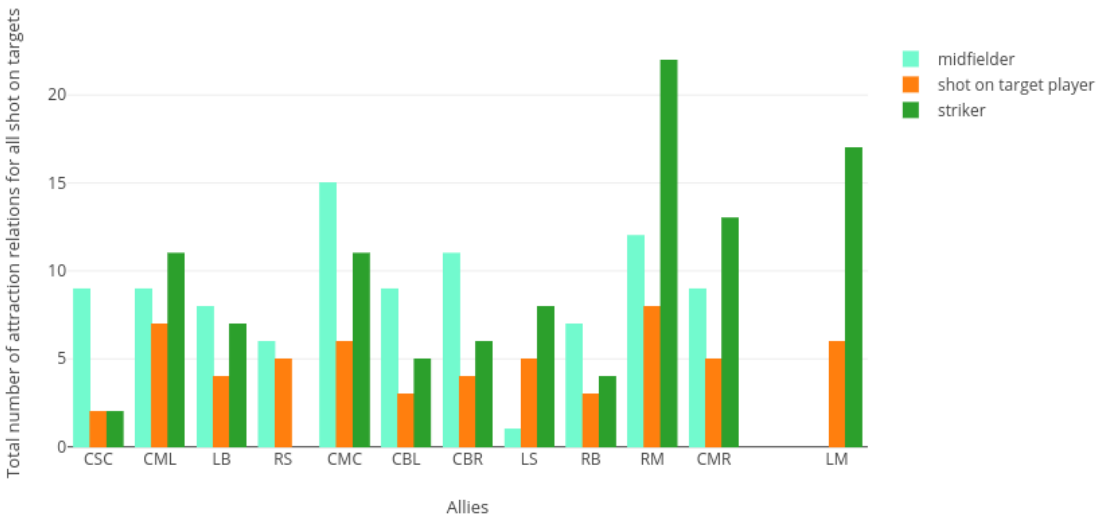


Figure 4.13: The total number of attraction relations that the shot on target player, the striker and the midfielder have with their teammates for a distance threshold of 10 meters and a time interval of 30 seconds.

Group 1	**Group 2**	**P-value**
Shot on target player's attraction with teammates	Midfielder's attraction with teammates	0.00282
Shot on target player's attraction with teammates	Striker's attraction with teammates	0.01435

Table 4.1: The p-values of the t-tests for the shot on target's attraction relations with her teammates versus the midfielder's and striker's attraction relations with their teammates.

In figure 4.14, we again observe that the shot on target player has less attraction relations with a lot of her opponents than the midfielder and striker have. This lower number of attraction relations with her opponents suggests that she is ahead of her opponents prior to a shot on target as well, again indicating that she is leading the attack.

There are also some opponents with which she does have more relations than the midfielder, for example the left back as well as the right and left centre back. The p-values in table 4.2 confirm this. The p-value for the t-test of the shot on target player's relations with her opponents versus the striker's attraction relations with her opponents indicates that there is a significant difference between them. The p-value of the t-test between the number of attraction relations with teammates, of the shot on target player and the midfielder however, is

slightly above 0.05. This means we can not reject the null hypothesis between these two types of players. The difference between the shot on target player and the midfielder in the number of relations with these players is however smaller than the difference between the shot on target player and the striker. This is remarkable because, as mentioned before, a shot on target player is a striker in the majority of the shot on targets and is therefore expected to resemble closest with the striker. The striker has a high number of attraction relations with her opposing backfielders. Although we stated that it is difficult to draw conclusions from attraction relations between opposing teams, we also mentioned that it is logical for the attacking and the defensive parties (the striker and the backfielders in this case) to have attraction relations prior to a shot on target. This is because these are the types of players that head towards the opponents side of the field during an attack, resulting in a crowded situation between them. The fact that the shot on target player has a lot less relations, but still some, with this attacking force indicates that she is also more ahead of the opposing backfielders on her path towards the goal.

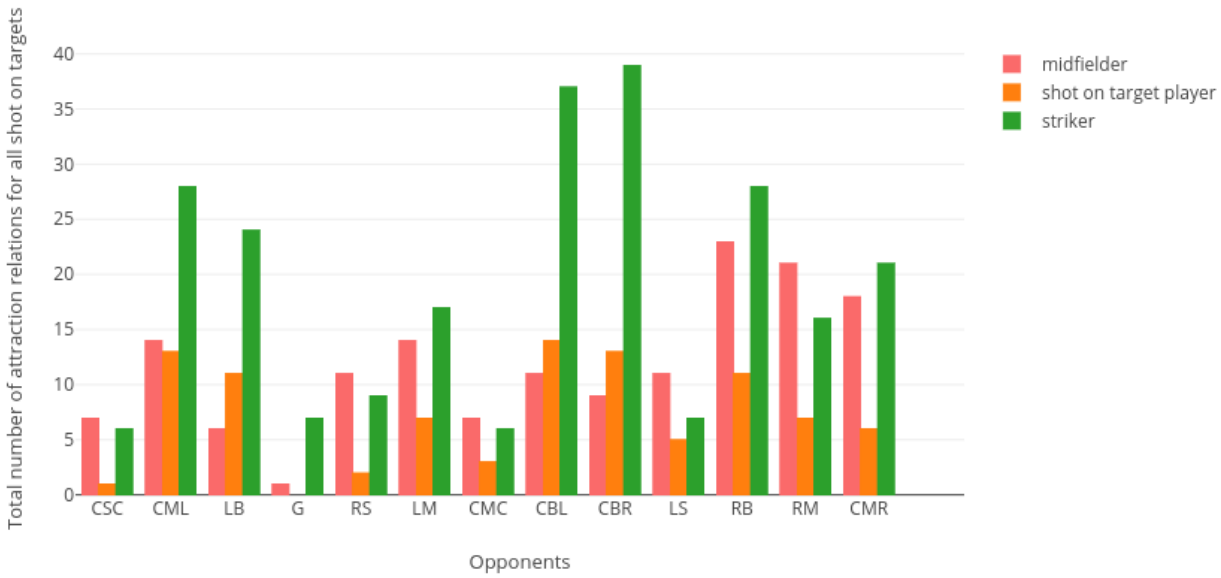


Figure 4.14: The total number of attraction relations that the shot on target player, the striker and the midfielder have with their opponents for a distance threshold of 10 meters and a time interval of 30 seconds.

Group 1	**Group 2**	**P-value**
Shot on target player's attraction with opponents	Midfielder's attraction with opponents	0.08029
Shot on target player's attraction with opponents	Striker's attraction with opponents	0.005493

Table 4.2: The p-values of the t-tests for the shot on target's attraction relations with her opponents versus the midfielder's and striker's attraction relations with their opponents.

Avoidance

The probabilities shown in table 4.3 indicate that the avoidance relations that the shot on target players has with her teammates differ significantly from the avoidance relations that the striker and the midfielder have

with their teammates. When talking about avoidance between teammates, we related it to spreading out prior to a shot on target in order for a player to make himself more available to receive passes. The fact that in figure 4.15 the shot on target player again has less avoidance relations with her teammates than the midfielder and striker do therefore might seem odd, since we would expect the player responsible for the shot on target to participate in this spreading out behaviour as well. However, figures 4.13 and 4.14 showed that the shot on target player is ahead of both her teammates and her opponents. If she is already ahead of them, the need to spread out disappears because this would already make him more available to receive passes. This is one reason why the shot on target player might not participate in spreading out behaviour. Another possible explanation is that she might have ball possession. If she indeed already has possession of the ball, there is also no need to spread out in order to be available for passes. After all, she already has the ball in that case.

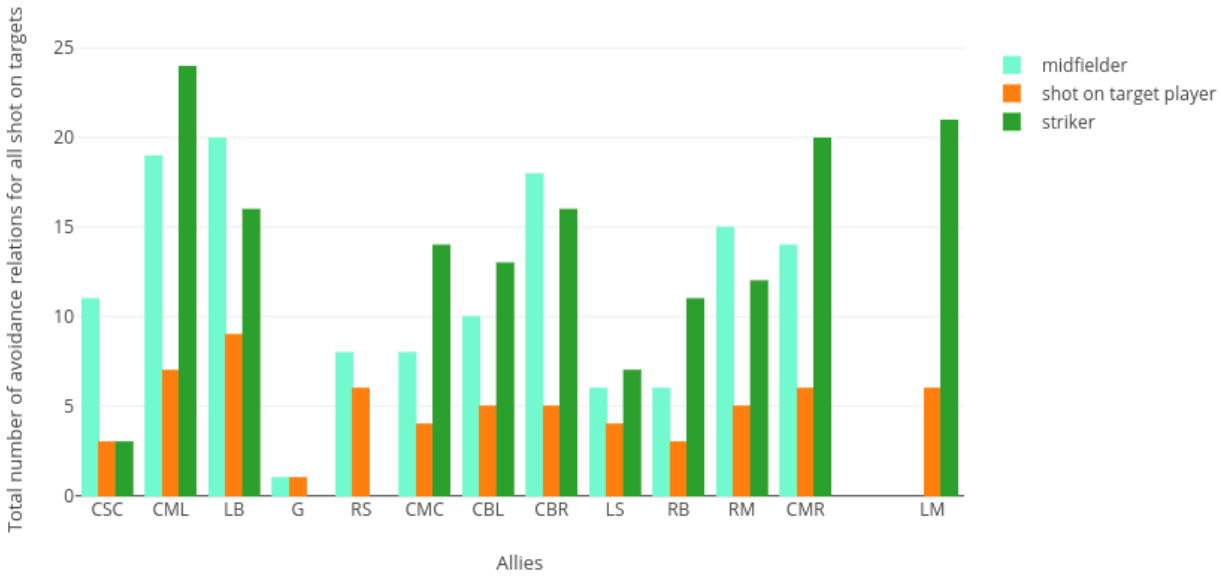


Figure 4.15: The total number of avoidance relations that the shot on target player, the striker and the midfielder have with their teammates for a distance threshold of 10 meters and a time interval of 30 seconds.

Group 1	Group 2	P-value
Shot on target player's avoidance with teammates	Midfielders's avoidance with teammates	0.001242
Shot on target player's avoidance with teammates	Striker's avoidance with teammates	0.00003315

Table 4.3: The p-values of the t-tests for the shot on target's avoidance relations with her teammates versus the midfielder's and striker's avoidance relations with their teammates.

Figure 4.16 shows us that the shot on target player has fewer avoidance relations with her opponents than the midfielder and striker do. Table 4.4 again confirms that this difference between the number of avoidance relations with opponents, between the shot on targets player and the striker and midfielder is significant by looking at the p-values of 0.008985 and 0.00002722, which are both far below 0.05. In section 2.6, we related avoidance between opponents with dodging behaviour, for example when the attackers head to the opponents

side of the field in order to score while trying to avoid the opposing players who are trying to block them or steal the ball. With this definition, we might expect the shot on target player to show dodging behaviour prior to the shot on target and thus have a relatively high number of avoidance relations. However, the opposite is true and she in fact has very few avoidance relations with most of her opponents. The reason for this might be the same as for the low number of avoidance relations in figure 4.15. Namely, from figures 4.13 and 4.14 it followed that the shot on target player leads the attack and is ahead of both her teammates and opponents. If she is indeed ahead of her opponents, there is less need to dodge them since she already got passed their defense.

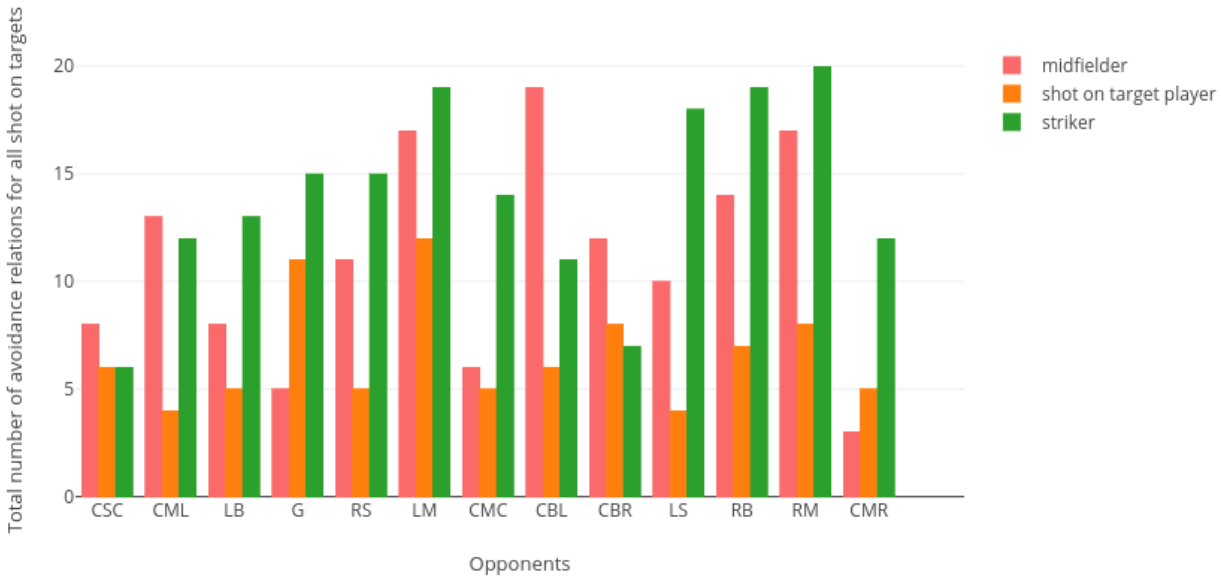


Figure 4.16: The total number of avoidance relations that the shot on target player, the striker and the midfielder have with their opponents for a distance threshold of 10 meters and a time interval of 30 seconds.

Group 1	**Group 2**	**P-value**
Shot on target player's avoidance with opponents	Midfielders's avoidance with opponents	0.008985
Shot on target player's avoidance with opponents	Striker's avoidance with opponents	0.00002722

Table 4.4: The p-values of the t-tests for the shot on target's avoidance relations with her opponents versus the midfielder's and striker's avoidance relations with their opponents.

Following

We related following behaviour between teammates as backing each other up during an attack. This meant that if a certain player initiated an attack, other players would respond to this attack by following their attacking teammate in order to back him up. In graph 4.17, we see that the striker and midfielder both show this backing behaviour. The midfielder shows slightly more backing behaviour than the striker, which is what would be expected since the midfielder by the nature of his position initially is more behind the attack, causing him to follow more players. The probabilities in table 4.5 show that the difference in the number of following

relations with teammates between the shot on target player and the striker and the shot on target player and the midfielder, is significant. The fact that the shot on target player overall has less of these following relations with her teammates than the striker and midfielder is not so suprising after looking at the graphs above for attraction and avoidance. All graphs up to this point indicated that the shot on target player is ahead of many players during an attack, making him the player leading the attack. This leading position immediately explains the lower number of following relations. Afterall, if someone is leading up front than there is no one left in front of him to follow.

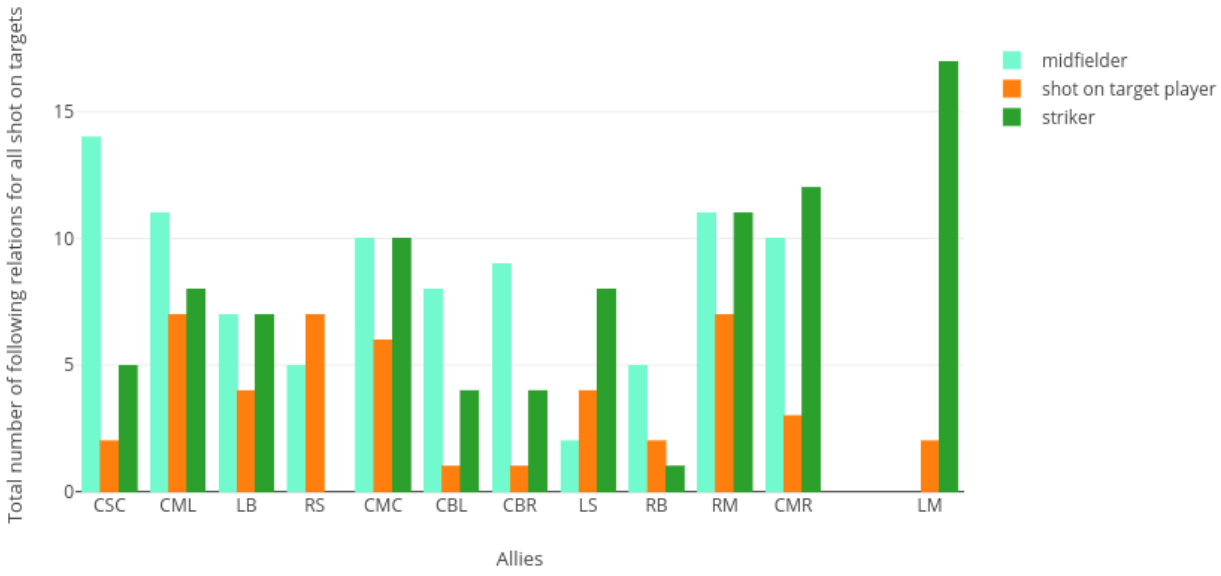


Figure 4.17: The total number of following relations that the shot on target player, the striker and the midfielder have with their teammates for a distance threshold of 10 meters and a time interval of 30 seconds.

Group 1	**Group 2**	**P-value**
Shot on target player's following relations with teammates	Midfielder's following relations with teammates	0.00125
Shot on target player's following relations with teammates	Striker's following relations with teammates	0.0117

Table 4.5: The p-values of the t-tests for the shot on target's following relations with her teammates versus the midfielder's and striker's following relations with their teammates.

Figure 4.18 shows that the shot on target player has a relatively low number of following relations with her opponents, compared to the striker and the midfielder. Yet again confirmed by the probabilities in table 4.6 for both the midfielder and the striker. The explanation for this is the same as for figure 4.17 and can again be found in all other graphs that we have seen up to this point. Namely, all previous figures indicated that the shot on target player is ahead of her opponents prior to a shot on target. Since the shot on target player is ahead of most of her opponents, she does not have many opponents in front of him left to follow, explaining the relatively low number of following relations in figure 4.18.

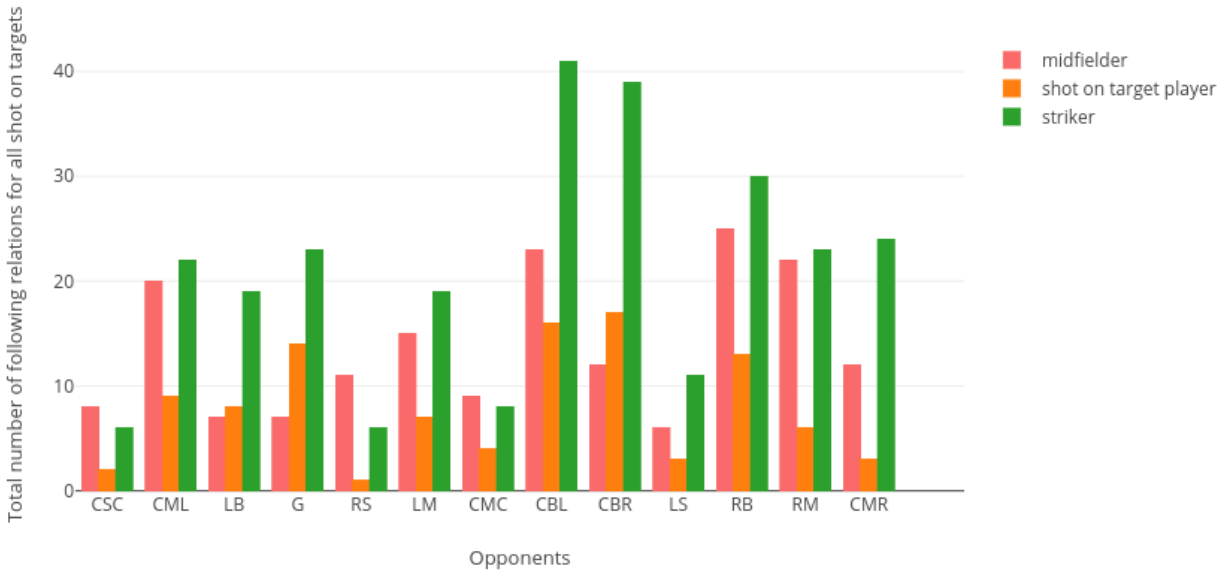


Figure 4.18: The total number of following relations that the shot on target player, the striker and the midfielder have with their opponents for a distance threshold of 10 meters and a time interval of 30 seconds.

Group 1	**Group 2**	**P-value**
Shot on target player's following relations with opponents	Midfielder's following relations with opponents	0.02648
Shot on target player's following relations with opponents	Striker's following relations with opponents	0.001147

Table 4.6: The p-values of the t-tests for the shot on target's following relations with her opponents versus the midfielder's and striker's following relations with their opponents.

Chapter 5

Conclusions

In this thesis, we have investigated whether the player who makes a shot on target shows deviant behaviour prior to that shot on target in comparison with other players. First, we had to find the right players to compare the shot on target player's behaviour with. When looking into the behaviour of the shot on target player, we found that he is a striker in about 58% of the shot on targets and a midfielder in 36% of the shot on targets. Therefore, we chose the behaviour of these players in the comparison with the shot on target player.

In order to compare their behaviour, we had to find the number of attraction, avoidance and following relations for all three of these types of players. In section 4.2, we explained how these total number of relations were found. Based on these number of relations we performed comparisons in section 4.4. This comparison is based on the results for a distance threshold of 10 meters and a time interval of 30 seconds. The other distance thresholds were left out of consideration in the comparison, due to the fact that the distribution of the number of relations remains the same as shown in section 4.1 and figures 4.1 and 4.2.

We wanted to know whether the behaviour of the shot on target player differed significantly from the behaviour of the striker and the midfielder. The graphs in section 4.4 showed that for all three types of investigated relations, avoidance, attraction and following, the shot on target player overall has fewer of these relations with both his teammates and his opponents than the midfielder and striker do. The p-values resulting from the t-tests that were performed, confirmed that these differences were indeed significant.

In terms of soccer, these differences lead to the conclusion that the shot on target player has one main feature, namely he is leading the larger part of an attack for most of the attacks. Besides his leading position, the low number of avoidance relations with his teammates, shown in figure 4.15, insinuated that the shot on target player does not have to take part in spreading out behaviour because he already has ball possession. This creates an image of the shot on target player as a player who initiates an attack and manages to break through the defensive line of his opponents while maintaining possession of the ball. This type of tactic is different from possession play and is called direct play. The tactical implication of direct play is for the team in possession to move the ball into a shooting position as directly as possible with the least number of passes [HF05]. In the case of the UEFA Women's European Championship of 2017, direct play appeared to be a good tactic

for creating goal scoring opportunities. It is important to note that this does not say anything about the best tactic to actually score goals with, because we did not take into account here whether the shot on targets were actually successfull or not.

Another important finding in this research is the large influence that the distance threshold has on the outcome of the Movemine relation functions. This positive correlation between the distance threshold and the number of relations between players in combination with the fact that soccer players, unlike animals, use predetermined formations, leads us to think that it is a good idea to determine the distance threshold based on the average distance between two types of players. For example, two opposing defenders usually are far apart from each other, due to the nature of their positions and objectives. An opposing defender and attacker on the contrary are fairly close to each other. This research has shown that attract/avoid and following mining functions can be applied successfully to sports, or specifically soccer data. However, for further research of movement data within sports using attract/avoid and following mining functions, it would be good if the distance thresholds for these such combinations of players were to be more tuned to the positions in which they play, in order for the results to be more normalized. If the results are more normalized based on the position of the players, than the outcome of these functions is better suited to compare the individual relations between players with each other, whereas in this thesis the comparison between players is more focussed on their overall relations.

Bibliography

[CMFE+16] Andra Cacho, Luiz Mendes-Filho, Daniela Estaregue, Brunna Moura, Nlio Cacho, Frederico Lopes, and Cristiano Alves. Mobile tourist guide supporting a smart city initiative: a brazilian case study. *International Journal of Tourism Cities*, 2(2):164–183, 2016.

[FLK+16] Guilad Friedemann, Yossi Leshem, Lior Kerem, Boaz Shacham, Avi Bar-Massada, Krystaal M. McClain, Gil Bohrer, and Ido Izhaki. Multidimensional differentiation in foraging resource use during breeding of two sympatric top predators. *Scientific Reports*, 6(2), October 2016.

[HF05] Mike Hughes and Ian Franks. Analysis of passing sequences, shots and goals in soccer. *Journal of Sports Sciences*, 23(5):509–514, 2005. PMID: 16194998.

[HT04] Shoji Hirano and Shusaku Tsumoto. Finding interesting pass patterns from soccer game records. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, *Knowledge Discovery in Databases: PKDD 2004*, pages 209–218, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[Inc15] Plotly Technologies Inc. Collaborative data science, 2015.

[LDW+13] Zhenhui Li, Bolin Ding, Fei Wu, Tobias Kin Hou Lei, Roland Kays, and Margaret C. Crofoot. Attraction and avoidance detection from movements. *Proc. VLDB Endow.*, 7(3):157–168, November 2013.

[LWC13] Zhenhui Li, Fei Wu, and Margaret C. Crofoot. Mining following relationships in movement data. In *2013 IEEE 13th International Conference on Data Mining*, pages 458–467, Dec 2013.

[OAC+13] Renato Oliveira, Paulo J. L. Adeodato, Arthur Carvalho, Icamaan Viegas, Christian Diego, and Tsang Ing Ren. A data mining approach to solve the goal scoring problem. *CoRR*, abs/1305.4955, 2013.

[SJS+17] Manuel Stein, Halldór Janetzko, Daniel Seebacher, Alexander Jäger, Manuel Nagel, Jürgen Hölsch, Sven Kosub, Tobias Schreck, Daniel Keim, and Michael Grossniklaus. How to make sense of team sport data: From acquisition to data modeling and research aspects. *Data*, 2(1):online, 2017.

[SL17] Ronald A Smith and Keith Lyons. A strategic analysis of goals scored in open play in four fifa world cup football championships between 2002 and 2014. *International Journal of Sports Science & Coaching*, 12(3):398–403, 2017.

[STH] STHDA. Student t-test for unpaired samples.

[WLLH14] Fei Wu, Tobias Kin Hou Lei, Zhenhui Li, and Jiawei Han. Movemine 2.0: Mining object relationships from movement data. *Proc. VLDB Endow.*, 7(13):1613–1616, August 2014.

Appendix A

Appendix

A.1 Collecting the results

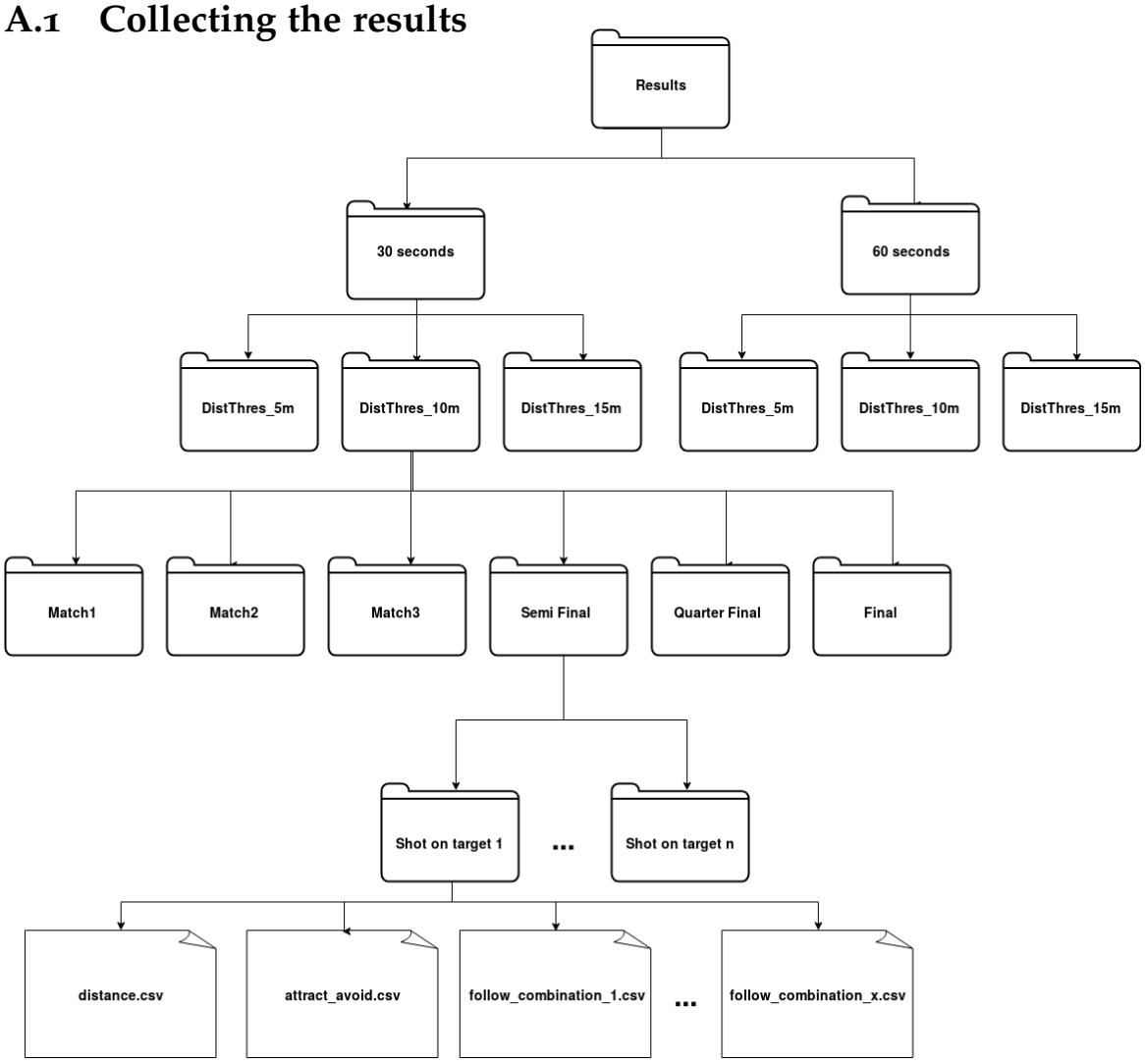


Figure A.1: The results data structure in the last fase of the preprocessing of the data.

shotOnTargetId	Round	Home	Away	Half	Distance_ RB-home_RM-away	Distance_ RB-home_LM-home	SigValue_ RB-home_RM-away	SigValue_ RB-home_LM-home	Follow_ CBL-away_RS-home	Follow_ RS-home_CBL-away
0	Final	HOLLW	DANEW	1	14,2392942313	10,2199381512	0,5	0,5	[']	[']
1	Final	HOLLW	DANEW	1	23,7438798175	9,252521925	0,5	1E-10	[3, 2]	[1]
2	Final	HOLLW	DANEW	1	30,302307325	12,4968079123	0,5	1E-10	[3, 1]	[2]
3	Final	HOLLW	DANEW	1	38,5391098901	22,9394062296	0,5	0,357	[']	[1]

Figure A.2: The structure of the resulting table after collecting data with the same parameter setting. The columns starting with "SigValue_" correspond to the results of the attract/avoid mining function. The results of the following mining function are placed in an array representating the lengths of all the following relations that occur between two players.

A.2 Dependencies between the results and the used parameters

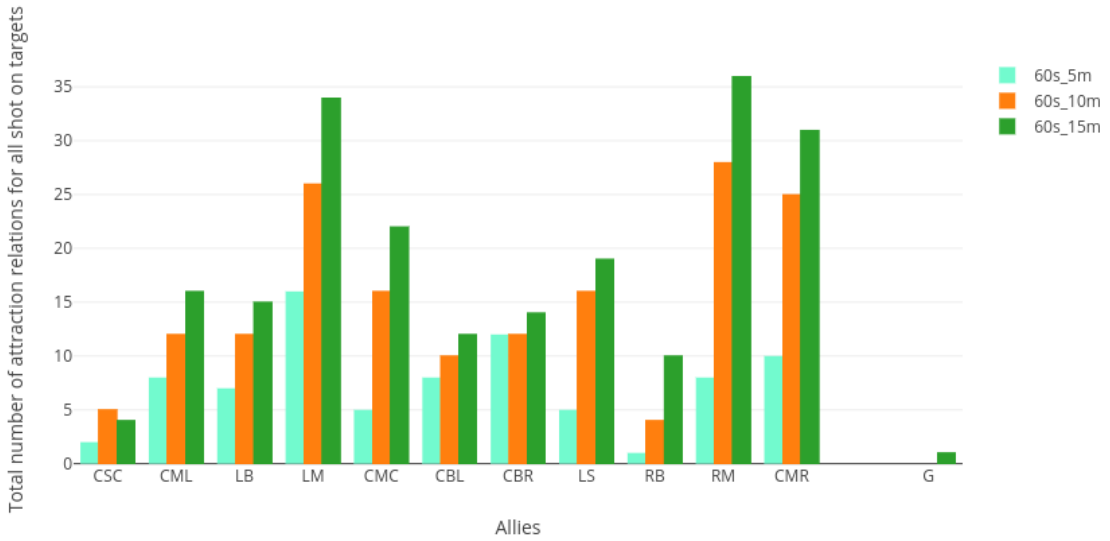


Figure A.3: The total number of attraction relations that the right striker on the attacking team has with his allies, for all distance thresholds and a time interval of 60 seconds. The number of relations is the total number of all shot on targets found in all six matches.

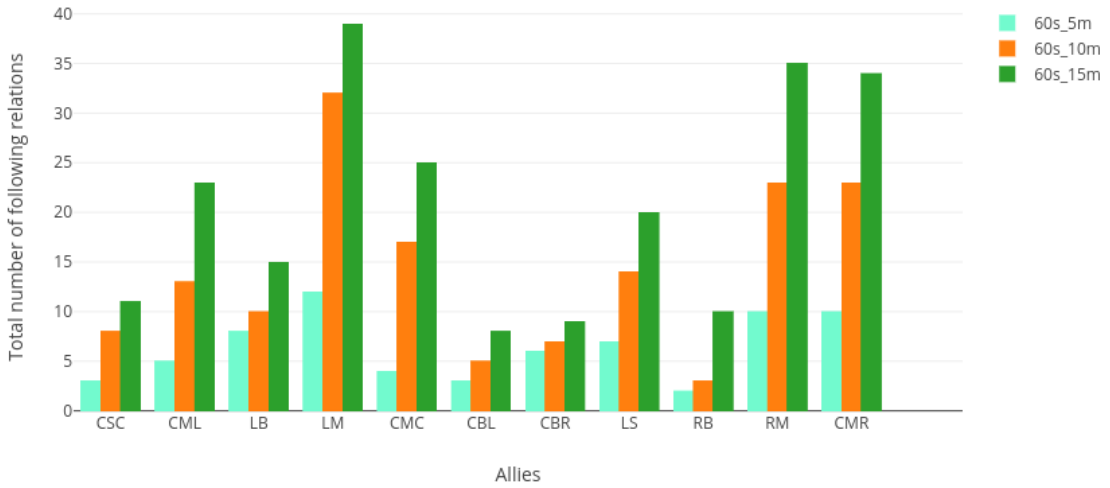


Figure A.4: The total number of following relations that the right striker on the attacking team has with his allies, for all distance thresholds and a time interval of 60 seconds. The number of relations is the total number of all shot on targets found in all six matches.

A.3 Finding the number of attraction and avoidance relations

Significance values

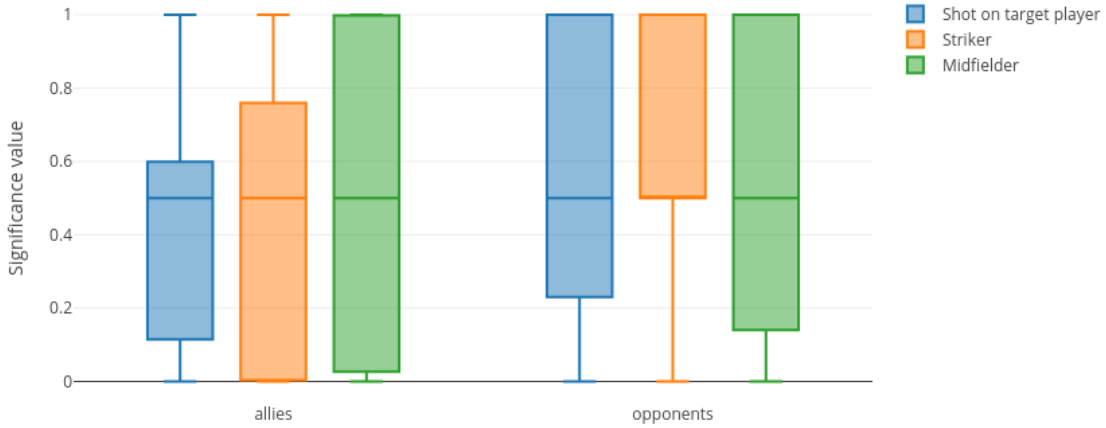


Figure A.5: The significance values of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 15 meters together with a time interval of 30 seconds.

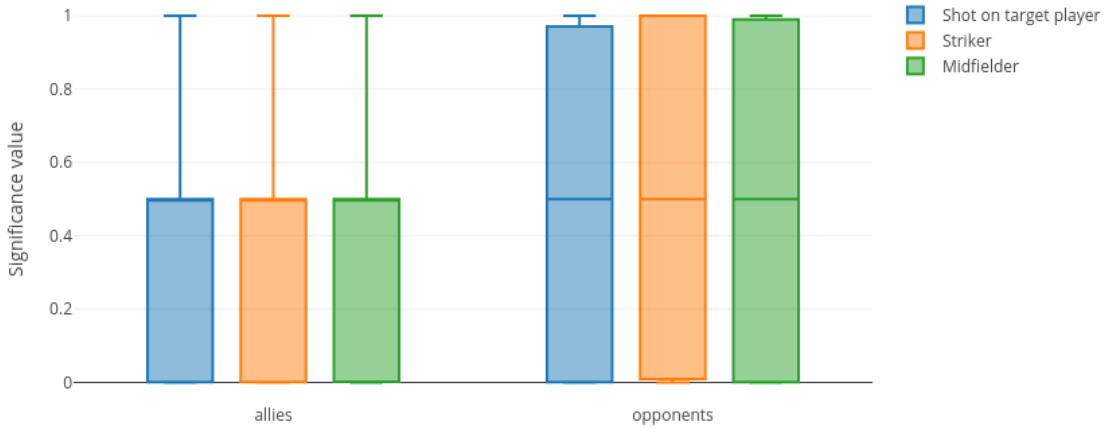


Figure A.6: The significance values of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 5 meters together with a time interval of 60 seconds.

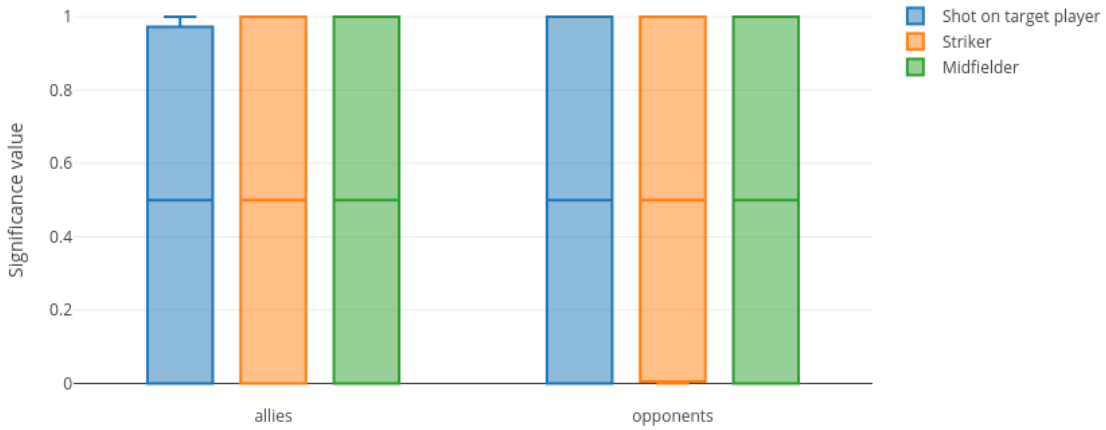


Figure A.7: The significance values of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 10 meters together with a time interval of 60 seconds.

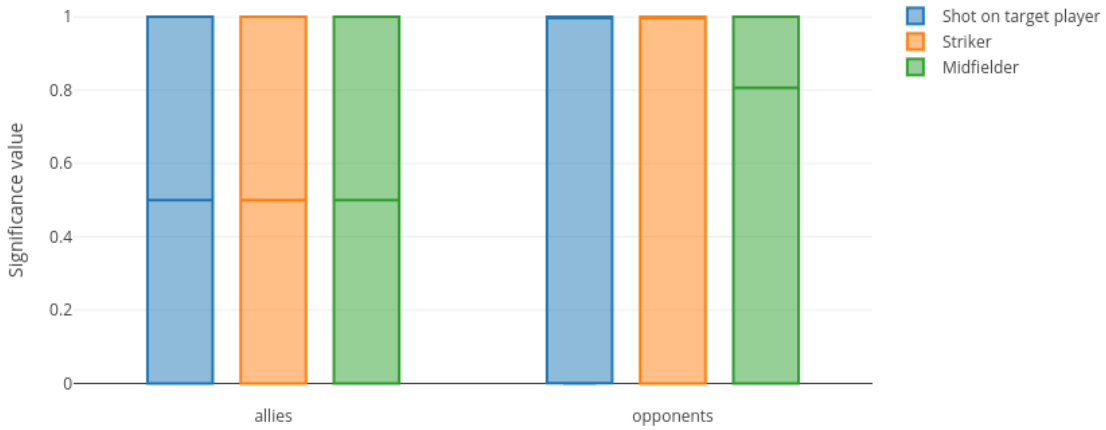


Figure A.8: The significance values of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 15 meters together with a time interval of 60 seconds.

Attraction

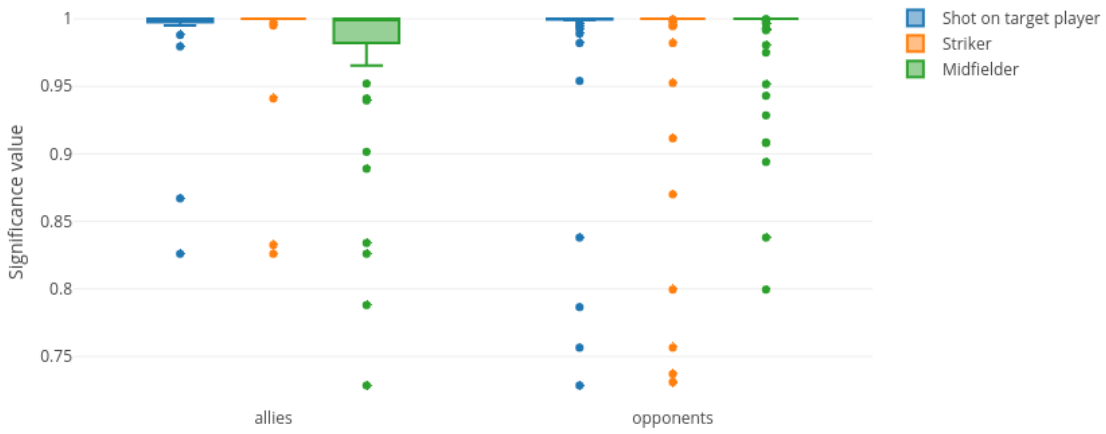


Figure A.9: The significance values that are ≥ 0.7 of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 5 meters together with a time interval of 30 seconds.

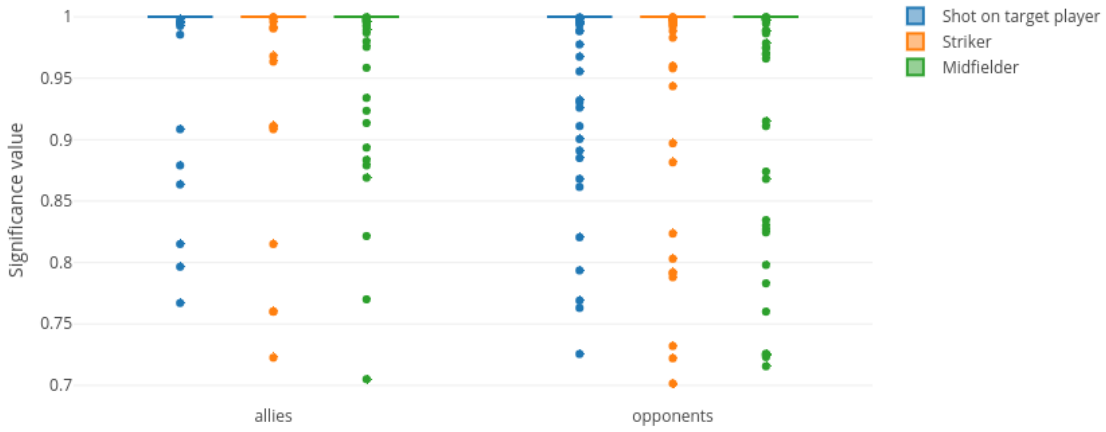


Figure A.10: The significance values that are ≥ 0.7 of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 15 meters together with a time interval of 30 seconds.

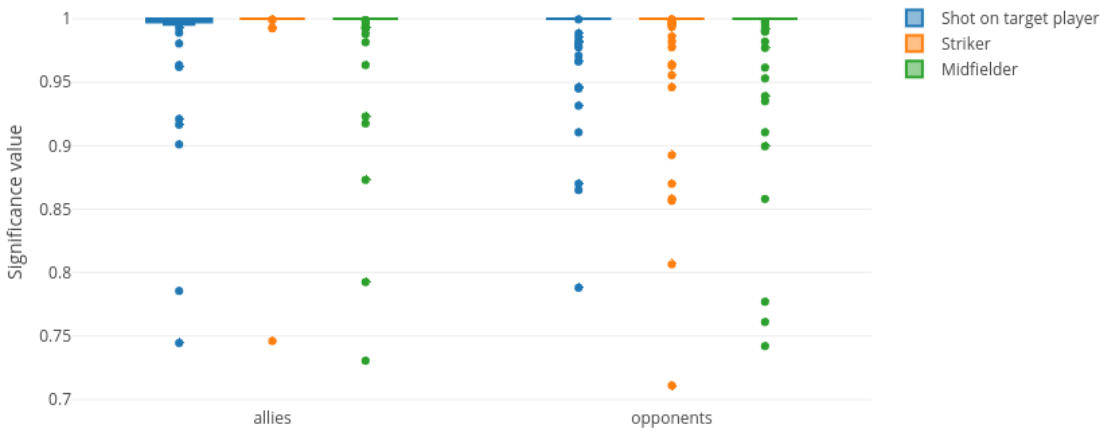


Figure A.11: The significance values that are ≥ 0.7 of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 5 meters together with a time interval of 60 seconds.

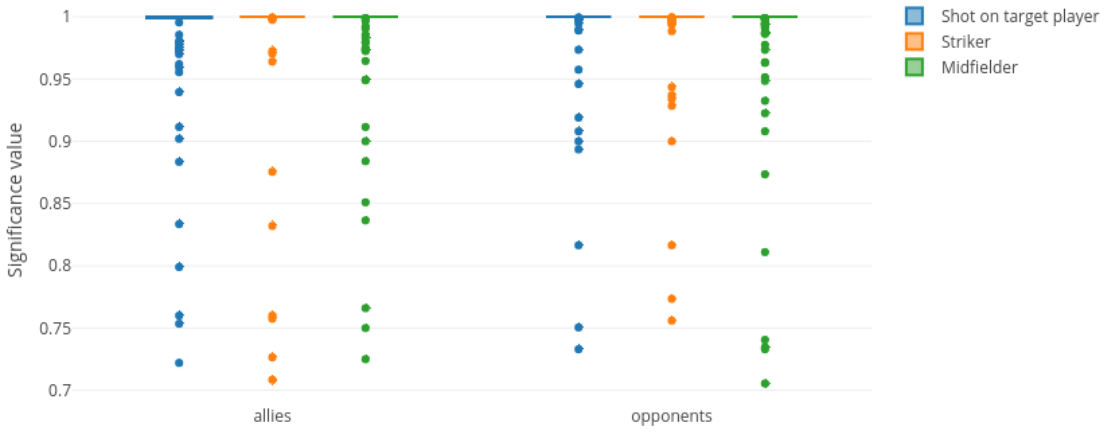


Figure A.12: The significance values that are ≥ 0.7 of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 10 meters together with a time interval of 60 seconds.

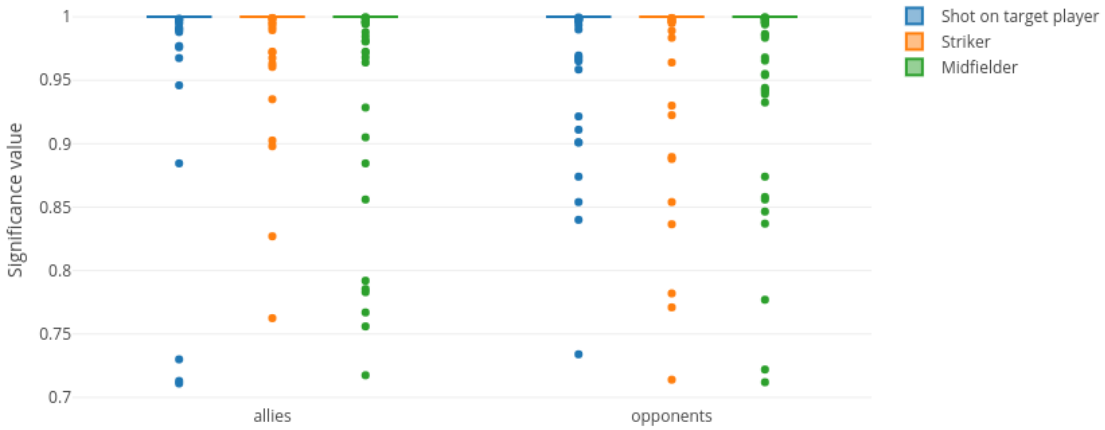


Figure A.13: The significance values that are ≥ 0.7 of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 15 meters together with a time interval of 60 seconds.

Avoidance

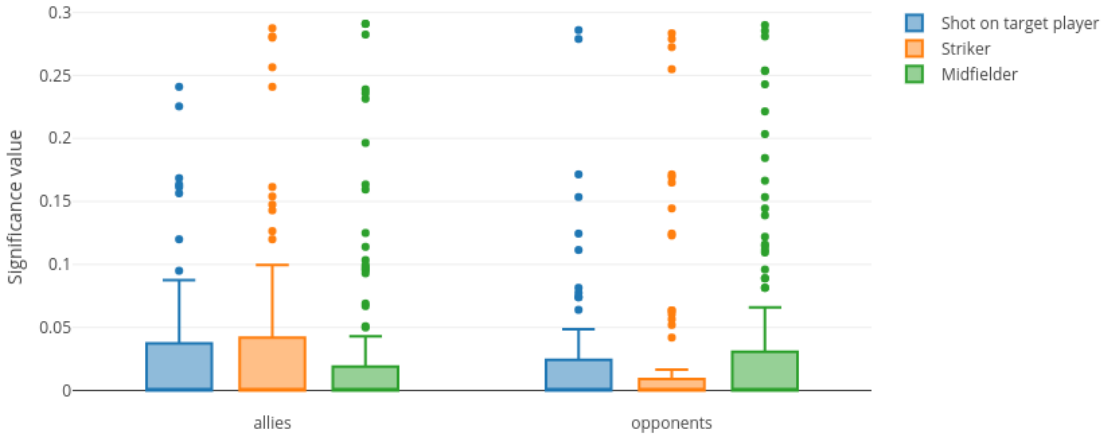


Figure A.14: The significance values that are $\leq$ of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 5 meters together with a time interval of 30 seconds.

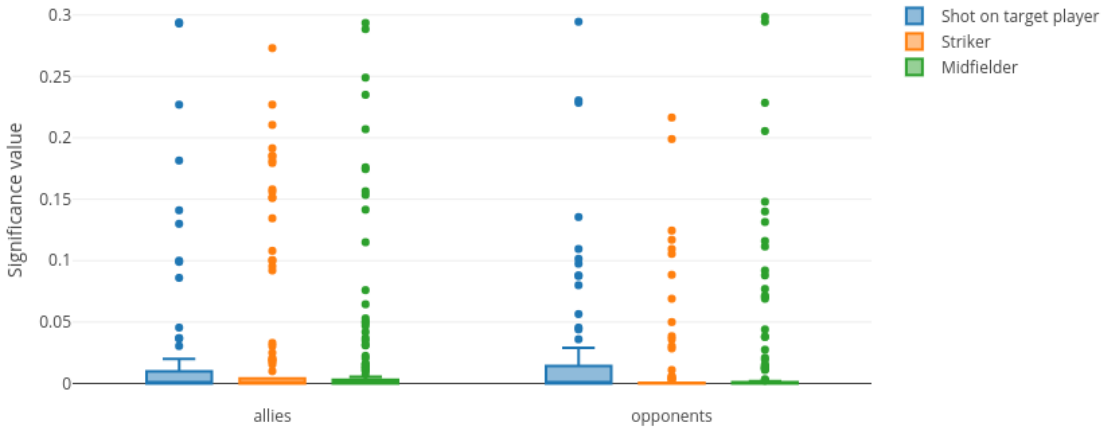


Figure A.15: The significance values that are $\leq$ of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 15 meters together with a time interval of 30 seconds.

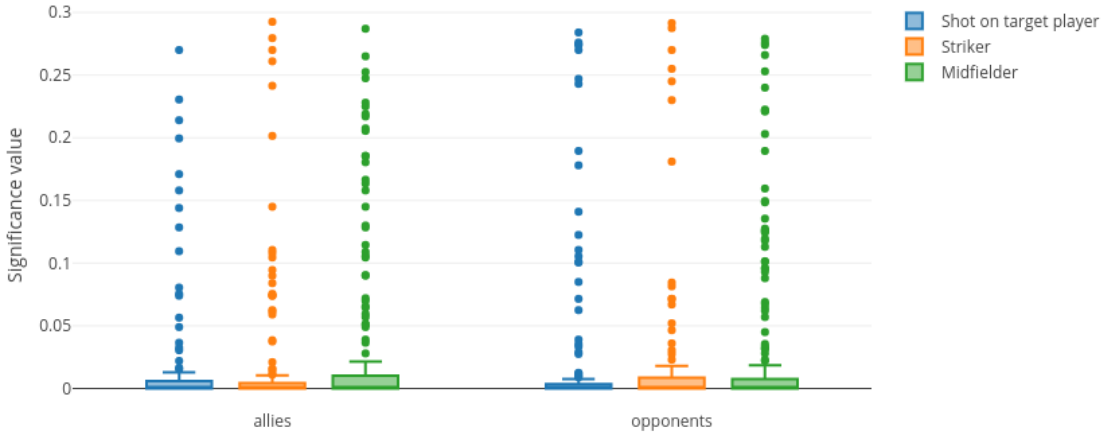


Figure A.16: The significance values that are $\leq$ of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 5 meters together with a time interval of 60 seconds.

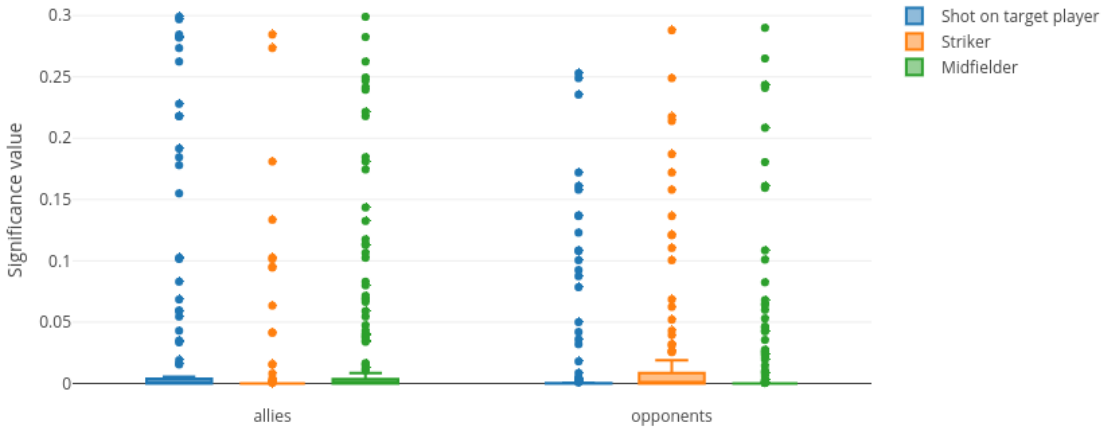


Figure A.17: The significance values that are $\leq$ of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 10 meters together with a time interval of 60 seconds.

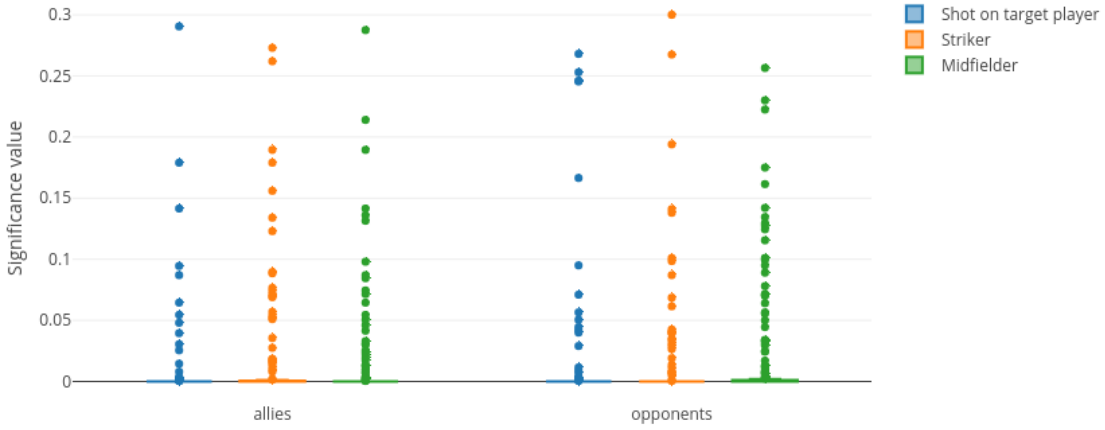


Figure A.18: The significance values that are $\leq$ of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 15 meters together with a time interval of 60 seconds.

A.4 Finding the number of following relations

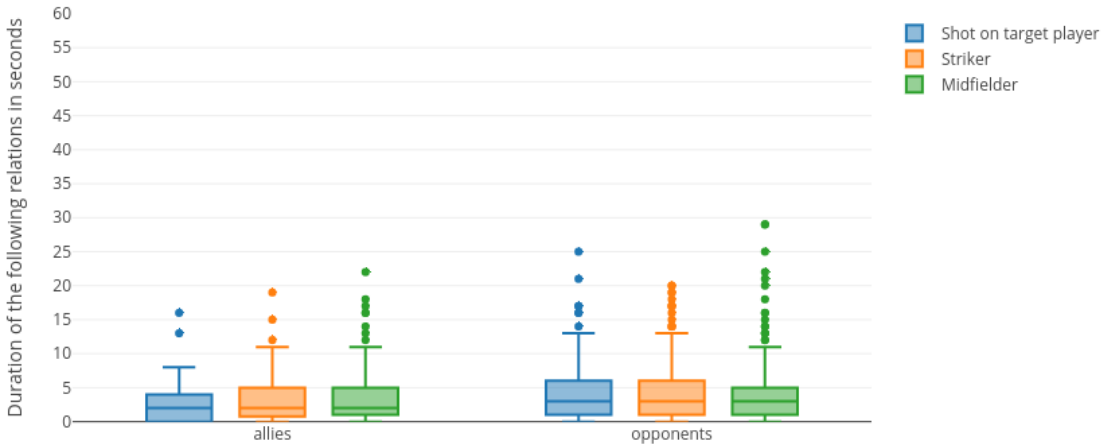


Figure A.19: The durations of the following relations of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 5 meters together with a time interval of 30 seconds.

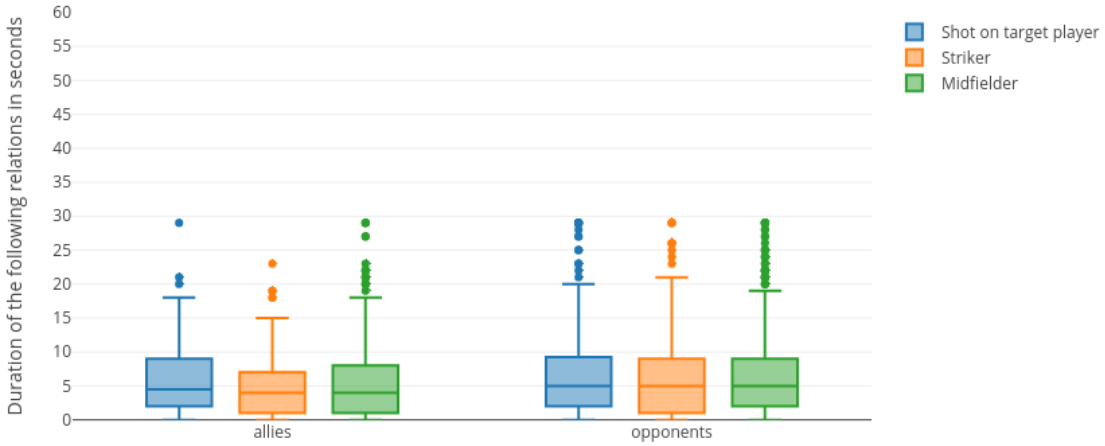


Figure A.20: The durations of the following relations of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 15 meters together with a time interval of 30 seconds.

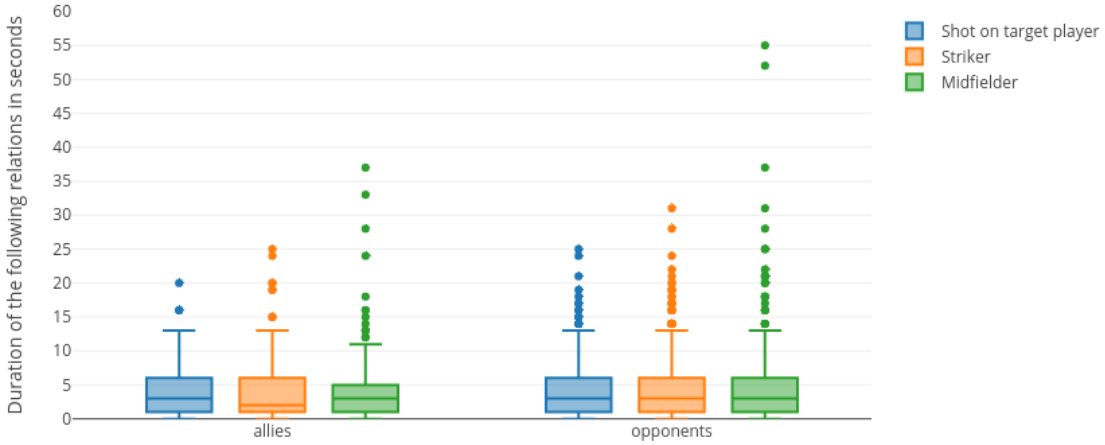


Figure A.21: The durations of the following relations of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 5 meters together with a time interval of 60 seconds.

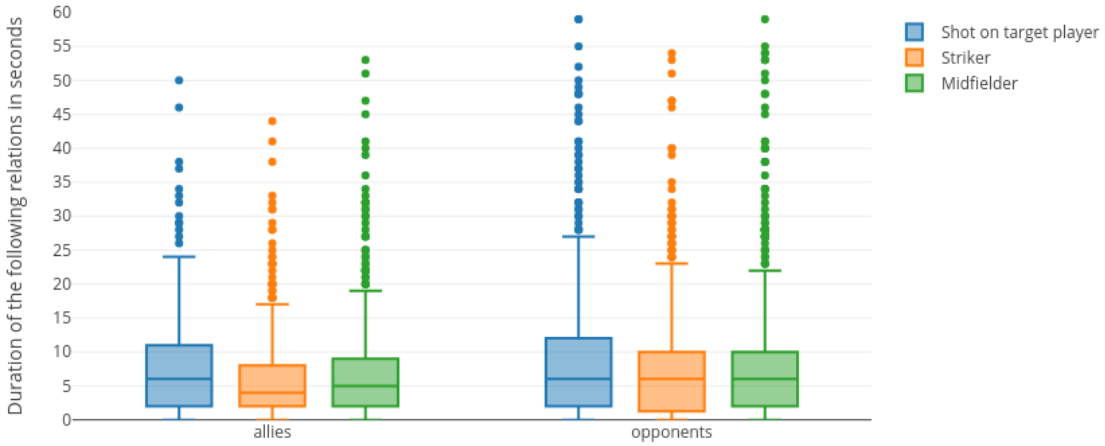


Figure A.22: The durations of the following relations of the shot on target player, the striker and the midfielder, of all shot on targets found in all six matches and a distance threshold of 15 meters together with a time interval of 60 seconds.