Universiteit Leiden
The Netherlands

Opleiding Informatica

System for analysing off-the-ball performance of individual football players using spatiotemporal tracking data

V.J.A. Poslavsky

Supervisors:
A.J. Knobbe (LIACS), L.A. Meerhoff (LIACS) & T. Janssen (KNVB)

BACHELOR THESIS

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

28/01/2019

Abstract

Recent advances in the developments of tools to analyse football matches have intrigued the Dutch Football Association to the extent that they use these tools to increase their knowledge about the game of football. At the moment, players are judged by Sports Scientists who apply a model by hand. However, this is very time consuming and therefore this approach has its limitations. By combining Sports Science with Data Science, it should be possible to create an automated system that objectively rates players on their off-the-ball performance. This will help the coach pick the best players for his starting eleven and might give him insights on which aspects of the team require the most training. This brings us to the core of this paper, which will answer the question:

"How can off-the-ball performance of individual football players be analysed using spatiotemporal tracking data, in a way that generates valuable information for the coach?"

Contents

Chapter 1

Introduction

In this chapter, the research question of this thesis is introduced.

1.1 Data Analysis in Team Sports

Recent progress in sensor development has led to an increased interest in the analysis of the movement of individual players during team sports. In team sports, judging the positioning of players objectively is thought to be more challenging than that of sports that are performed individually. This is because during team sports, players do not solely have to adapt their movement to that of their opponents, but also to the positioning of their teammates. Hence, it is difficult to give a clear description on the criteria that a player has to meet in order to be considered positioned in a way that either satisfies the coach's game plan or in a way a player has the best chance to score. This has caused the interest in the analysis of the corresponding data sets to currently be on the rise [SJS+17] .

In football, professional teams invest substantial resources to analyse their own team's performances and that of their (future) opponents. Most data sets, such as data containing exact movement trajectories, are not made available publicly. However, basic statistics about the distance a player has covered and the number of shots taken by a player, are publicly available for analysis purposes. An example of a company that gathers these statistics is Opta Sports, which is a sports data industry that has already started to establish itself since 1996. At first, data would only be recorded by viewing video footage of the game after the match was played. However, from 1999 onwards, Opta was able to collect match data in real-time [Spo19].

Depending on the available data, analysts execute different analysis tasks on data. This analysis mostly does not focus on the outcome of the action that was performed by the player, but on the question why the player decided to make the decision and thereafter perform the action. For example, the analysts want to get a better understanding of why a player picked a certain teammate to pass to rather than one of his or her other teammates. Moreover, the influence that this decision had on the performance and positioning of both their own and the opposing team is examined. With these results, the discoveries that are made can be trained and discussed with the coach, players and scouts, hereby benefiting to the improvement of the scouting network and the training sessions.

In most sports, data can be divided into three different types. These three different types of data are all obtained from both video footage and sensors. The first type is called movement data. These data describe where an actor or game object is located at a specific point in time. In football, the positions are usually sampled at around 10 times per second (Hz). For each sensor, a timestamp (X, Y, Z)-position, as well as both overall and component-based velocity and acceleration are retrievable. By using passive recognition of patterns from video data, computer science can be used to automatically process movement data of both the players and the ball and provide the user with valuable information about both the individual players and the team. By processing these data with a computer, in combination with the available event data which covers match-relevant actions and happenings during the match, analysts can examine large amounts of spatiotemporal tracking data in a relatively short time span. Thus, more performances of both individuals and teams can be examined for both pre- and post-match analysis.

1.2 Data Analysis in Football

For many years, FIFA [FIF18a] disallowed the implementation of (technological) innovations during professional football matches. However, at the 2018 World Cup in Russia [FIF18b], teams were allowed to use technical resources to analyse their performance during matches.

In March 2018, three months before the World Cup took place, FIFA allowed the use of electronical performance and tracking systems in football matches. Furthermore, they allowed team analysts to transmit data and to communicate with the coaches during the match itself [Dat18]. These forms of technology had already been used by many teams during training but had been disallowed to use in a match environment.

Thanks to the newly introduced performance and tracking system, the physical aspects of players could be examined in real-time during the match. On the basis of a player's heart rate, covered distance and number of high-intensity sprints, analysts can, for example, report to the coach about a player being too tired to be able to fulfill the remainder of the match. Therefore, these measures can help the coach make better informed substitutions. This easily implementable use of spatiotemporal tracking data resulted in the broad incorporation of spatiotemporal tracking data.

Furthermore, in addition to these physical analyses, spatiotemporal tracking data also lend themselves for tactical analyses. For example, the X- and Y-coordinates of the players can be used to calculate the space between the lines of a team. By examining this information, the coach can compare the tightness of the team to either that of the opponent or to what had been discussed in the tactical team talk before the game. By using these data, the coach can not only verify his thoughts, but can also gain new insights on patterns of his team that he or she was not yet aware of.

One company that tries to retrieve and recognise patterns in football data is Opta Sports. Opta Sports is a British company that, in 2018, was the world's leader in sports data [Spo18b]. In football, Opta is known for its advanced metrics to analyse and calculate how both a team and an individual have performed. One of the most well-known metrics that Opta applies to these data is called Expected Goals (xG) [Spo18a]. In football, the aim of the game is to score more goals than the opponent. Nonetheless, the number of goals an attacker has scored during a season might not give a fair reflection of how much of a threat this player might have been to the opponent. Moreover, goals are quite rare in football, which means that over the course of a season, the number of goals that can be used for analysis is relatively low. With the Expected Goals metric, Opta measures the quality of a shot by analysing variables such as assist type, shot angle, distance to the goal and whether it was a headed shot. Hereafter, Opta combines these measures. However, unfortunately, the way Opta combines them is not publicly available. by combining these measures, Opta can determine whether a player or a team has either under- or overperformed during a match or a season. This is achieved by comparing the expected goals to the actual number of goals scored. Apart from this, it provides the coach with a measure that has been built up from more data points than only goals, giving him a better reflection of how well the team has actually performed. For example, if a match ends 2-0, the home team is expected to have been superior to the away team. However, when looking at the expected goals, the away team might have been expected to score more goals than the opponent during the match. If this is the case, the coach of the away team can state that his team has underperformed, with respect to this aspect of the game.

As is the case when examining the number of expected goals, most notational analysis in football is done for the actions of the player in possession of the ball. For example, the pass accuracy, shots and dribbles are noted for each individual during a match. However, at each moment during a game, only one of the at most 22 players on the field is in possession of the ball. Hence, it could be interesting to analyse off-the-ball performance of players. Currently, the only off-the-ball measurements that are done focus on a player's physique. For example, heart rate, number of accelerations, maximum speed and distance covered. The tactical aspect of the game is mostly neglected. Why has a player decided to position himself at a certain place on the field? Is the player able to receive the ball from his teammate from that position and what could the potential impact on the continuation of the attack be if the player receives the ball? Although the answers to these questions would benefit to our understanding of the game of football, they remain unanswered by computer and data scientists. Thus, the impact of data analysis is expected to grow in the future as soon as these types of questions can also be answered by it.

1.3 Data Analysis at the Royal Dutch Football Association (KNVB)

One can read in section 1.2 above that technology has provided football with many new insights and possibilities. Just like all the other clubs and national teams in the world, the Royal Dutch Football Association (KNVB) [Dat18] has access to tools to analyse both their own players and their opponent's players.

At the moment, the KNVB mainly uses the information that can be obtained from wearables and GPS tracking to examine a player's performances during a training. From these wearables, they can extract the effort that a player had make do to keep up with the pace of the session. Data of each individual player is then compared to both their performance in previous training sessions and to that of his or her teammates during the same training session. By doing this, the KNVB can see what the overall fitness of a player is compared to that of the other players in the team. Furthermore, this can show the KNVB which players have been slacking during the training and which players participated enthusiastically. This can then be fed back to the players. Other features that are recorded and notated in the player reports are the number of high intensity sprints a player has done and the player's average heart rate during the exercises. Again, these features are compared to both data from the corresponding player and other team members.

For the analysis of matches, the KNVB does very little with the available data. The only aspect that the association looks at thoroughly is set-pieces (e.g. free kicks and corners). Not only do they do this to examine and tweak attacking set-pieces, but also to view where the opponent usually places the ball during these events. Data analysts inform the coach and the trainers about their discoveries, in order for them to optimally prepare their players for what they can expect during the upcoming fixture.

In the future, the KNVB wants to improve their data analysis by adding metrics that cover tactical aspects of the game of football. Their aim is to achieve this by combining the knowledge of sports scientists and computer scientists and to implement this combined knowledge into an automated analytical system. The system being automated is an essential step towards improving the analysis of the game of football. Firstly, automated systems are faster, which allows more data to be explored and allows coach's to inform the players and the team on their recent performance as soon as possible. Moreover, an automated system would is able to rate players and teams objectively, which would lead to a more valuable analysis. This system should then be able to process spatiotemporal tracking data from matches of the Dutch National Team and objectively analyse various aspects of the game.

1.4 Research Question

This thesis will focus on rating football players objectively thanks to an automated notational analysis system. Since autumn 2012, spatiotemporal tracking data of the Dutch National Team have been recorded. From this moment, a total of 36 matches of which data is available have been played. To analyse these matches by hand would be very time-consuming and therefore not a desirable option. Moreover, analysing the system by hand is more prone for errors than if this is done by an automated system. Hence, it is preferable to create a system that can process these data automatically. This thesis will focus on creating and adding one aspect of the game: off-the-ball performance of players of the attacking team during build-up. More specifically, it will cover the positioning of players during attack with regards to how well players are able to receive the ball from the ball carrier. This system would be a first step towards the automated system desired by the KNVB to analyse tactical aspects of a game of football by processing football data. Therefore, the research question that is going to be addressed in this thesis is:

"How can off-the-ball performance of individual football players be analysed using spatiotemporal tracking data, in a way that generates valuable information for the coach?"

Chapter 2

Related Work

2.1 System for notational analysis in small-sided soccer games

In the past, Van Maarseveen and colleagues [vMOS17] developed a notational analysis system for small-sided soccer games. In their analysis, they quantify the attacking, both with and without the ball, and defensive abilities of football players. Three different roles are identified: attacker with ball, attacker without ball and defender. The model consists of a scheme with scores for every action a player could perform during build-up play. By doing this consistently, an overall performance score for each player can then be calculated by averaging the number of points a player achieved during each trial.

As an initial attempt to understand off-the-ball performance, Van Maarseveen studied a simplified setting: 3 vs. 2 + goalkeeper. Benefits of this test setting were that this would lead to less complex situations, more ball touches per player and would contribute to getting a better view of how well a player masters the basics of football, according to the standards set by the KNVB. The attackers had to try to score as quickly as possible, whereas the defenders had to prevent them from doing so. The tests took place on a regular training pitch with field dimensions of 40 by 25 metres. Furthermore, the game was played according to the regular rules of football.

2.2 Valuing passes in football using ball event data

In addition to Van Maarseveen's observation-based model in an experimental setting (3 vs. 2 + goal keeper), the players can also be evaluated with event data. For the creation of the automated system in this thesis, the KNVB has provided event data from Amisco. Hence, it is interesting to get a better understanding of how to implement these event data into a notational model. One such model that is closely related to the final product this thesis attempts to deliver is described by Bransen [Bra17].

The first approach to value passes that Bransen has thought of is called the zone-oriented pass value (ZPV), which divides the pitch into a number of equal zones. Using historical data, each zone is given a value for possessing the ball in that zone. A value for each zone is then calculated in two different ways. The first way is by determining the expected decimal value that a goal is scored from that zone within fifteen seconds. The second way is by implementing the expected goals model, as described in section 1.2. In other words: for each zone, the probability of scoring is determined if a shot were to be taken from there. This approach already resulted in better results than by looking at whether or not a pass was successful.

The second approach is the pass-oriented pass value (PPV), which relies on measuring the similarity of passes. The algorithm used for this passing value consists of five different steps. Firstly, a distance measure is determined to measure the similarity of passes. Thereafter, for all the passes of the training set, the outcome of the possession sequence it belongs to is specified. The next step is to divide the passes of the training set over different pre-clusters by only taking into account their origin and destination. Finally, for each of the passes of the test set, its pre-cluster is deduced and the k-nearest neighbours in the pre-cluster, based on the introduced distance measure, is found. The pass is assigned the average value of the outcomes of the neighbour's possession sequences. The results show that the PPV approach gives a better valuation of the passes, the passing skills of the players and the passing skills of the teams than the other approaches.

The third approach is called the sequence-oriented pass value (SPV). In this approach, each pass is valued by the influence it has on the expected outcome of the possession sequence. In order to achieve this, for each possession sequence a certain value is required. Next to that, we value the sub-sequences of the sequences such that we can measure the influence of the pass on the sequence's outcome. Although the approach seemed promising, the results were inferior to the PPV method.

2.3 Passing Decisions in Football: Introducing an Empirical Approach to Estimating the Effects of Perceptual Information and Associative Knowledge

Bransen used event data for research and Steiner [Ste18], also focuses on moments that have been captured at the moment a pass was given. The aim of Steiner's article is to bridge the gap between perceptual and knowledge-based information and present an approach to estimating the effects of perceptual information and associative knowledge on passing decisions. This approach is shown in the paper by using scenario-based data.

Scenario-based data that Steiner showed to the respondents consisted of 40 different game scenarios of which every player of the team was the ball carrier in four of them. The goal keeper was excluded from this analysis, due to the goalkeeper being considered a qualitative outlier. For each scenario, the respondents had to answer who they would pass the ball to. In two of the scenarios, most players decided that they probably would attempt a shot at goal rather than pass the ball to a teammate. Therefore, these scenarios had been excluded from the research. The other 38 scenarios were used for validation of Steiner's four different components. These components are explained and displayed in section 3.1.

2.4 Comparing Van Maarseveen and Bransen to Steiner

Steiner's model was used as the core of the automated system to rate the off-the-ball performance of players in this thesis. The paragraphs below will explain the disadvantages of Van Maarseveen's and Bransen's models as opposed to Steiner's model, where after the benefits of Steiner's model will be explained in the final paragraph of this section.

As can be read in section 2.1, Van Maarseveen created a system for notational analysis in small-sided football games. Although this system can be considered as a first step in the right direction, it still lacks clear definitions of components that the system is based on. For example, a player is awarded points for his off-the-ball positioning based on him being either open or marked. However, Van Maarseveen does not offer a clear description about which requirements a player has to meet in order to be considered open. Moreover, Van Maarseveen only slightly distinguishes certain positions on the field from each other. Namely, by stating whether a player is positioned on either his own or the opponent's half and whether the player is positioned at the left, in the middle, or on the right side of the field. Therefore, this system is thought to be too imprecise to be implemented by a computer scientist, due to too many variables within the system not being clearly defined. Without a clear definition, it is impossible to create an automated system that provides the correct scores from spatiotemporal tracking data. Within Steiner's model, the components are clearly displayed and the formula to calculate the normalised ratings for each individual attribute could be derived relatively easy. Hence, although it does offer insights on how to rate players objectively, Van Maarseveen's model was not used as the core for the automated notational system created in this thesis.

In Bransen's paper, three different methods are described to value passes in football by using event data. The first method that is described in this paper is the zone-oriented pass value (ZPV). This method is very interesting for the creation of the model in this thesis, because it compares the zone of the player in possession to the zone of the player who could potentially receive the ball. This could be one of the components to base the positioning of a player on. If a player is positioned in a zone from which the chance of scoring a goal is larger, then the player is probably an interesting passing option. The potential increase in the threat that a player can impose compared to that of the player in possession is therefore a valuable measure and will be kept in mind during the process of creating the objective system to rate a player's positioning in this thesis. However, as can be read in section 2.3, Steiner has created an entire model based on the positions of other players compared to that of the ball carrier. This, in fact, is also a form of a zone-oriented analysis, hereby comparing the zone of the ball carrier to that of a potential ball receiver. Although Bransen achieves this zone-oriented analysis in a different way, Steiner's method was preferred to it, due to its clarity and simpilicity.

Bransen's second approach is the so-called pass-oriented pass value (PPV). This second approach focuses on the sheer quality of passes rather than the positioning of the players receiving them. In Steiner's model, which is covered in section 2.3, a passing decision is evaluated by examining the way players are positioned concerning them being able to receive the ball. Within the model, the components are all based on the positioning of players compared to the positions of all the other players on the field. Thus, Steiner's method is linked more closely to spatiotemporal positioning data of players than that of Bransen's pass-oriented pass value (PPV), making it more suitable for answering the research question.

Finally, Bransen gives a detailed description of her sequence-oriented pass value (SPV). This third approach is very interesting and recognising passing sequences is definitely something that can and probably should be done. Unfortunately, this method is too time consuming to be implemented in this thesis due to the large amount of preliminary work that has to be done before actually being able to build upon this feature. This is not the case with Steiner's model, which can be applied directly onto data. Moreover, Bransen's sequence-oriented approach requires a large set of matches to be used for preliminary analysis. Due to the Royal Dutch Football Association (KNVB) having provided thirty matches with properly configured event data, this number of matches is not considered enough if half of the matches were to be used for a preliminary analysis, due to this causing the number of experimental matches to be too small to do a proper data analysis. Therefore, Steiner's method is preferred to this final approach by Bransen.

In conclusion, Steiner's model was preferred to the methods described in Van Maarseveen's and Bransen's papers. The features in Steiner's paper, as opposed to the features described in Bransen's and Van Maarseveen's, are clearly defined. Therefore, the features can be recognised and obtained from football data by a computer. Furthermore, the correct formula can be used to normalise the features. Thus, the knowledge in the paper can be transferred to a computer science environment. Lastly, the advantage of Steiner as opposed to Bransen's model is that it is not necessary to recognise patterns in data. Hence, no precomputing of football data is required, meaning that the model can be implemented directly. Due to the above-mentioned reasons, the features that are listed in the Steiner's paper will function as the core components of the system that is created in this thesis. Minor alterations that are thought to improve the system will be implemented and described in section 3.3, "Changes to Steiner's model".

Chapter 3

Method

As was mentioned section 2.3, the core components of the automated system that is created in this thesis are obtained from Steiner's "Passing Decisions in Football: Introducing an Empirical Approach to Estimating the Effects of Perceptual Information and Associative Knowledge" [Ste18]. The components of this system are displayed in Figure 3.1.

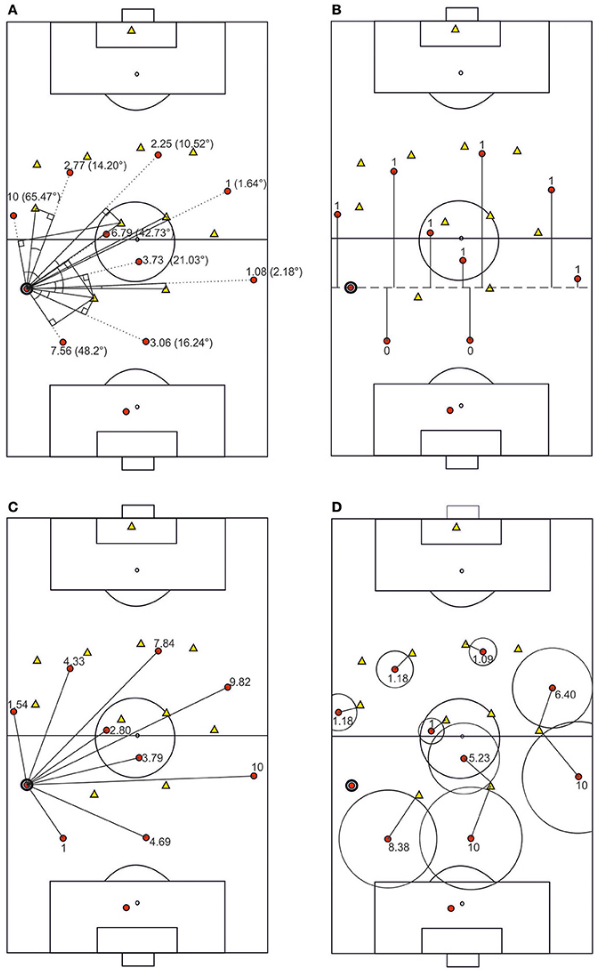


Figure 3.1: The variables from Steiner's model to value passing decisions in football [Ste18]

3.1 Explaining the components of Steiner's model

Before diving into the way Steiner's model will be used to implement the automated system to rate players' off the ball positioning during attack, it is important to get a clear understanding of the four components that Steiner's model consists of. Therefore, this section will give a detailed explanation of these components.

The variable at the top left of Figure 3.1, variable A, represents the openness of the passing lane from the ball carrier to each team member. The way this has been measured is by taking the angle between the passing line (the straight between ball carrier and receiver) and the straight to the opponent standing closest to this passing lane. The distance of an opponent to the passing lane was measured by drawing the perpendicular to the passing lane.

Variable B in Steiner's model covers a teammate's relative closeness to the opponent's goal. The aim of a game of football is to score more goals than the opponent. Therefore, it seems credible to state that a pass towards the opponent's rear line would be considered a good one. This variable used to rate this component is binary. Each player receives a 0 if the player is positioned further from the opposition's rear line than the ball carrier and a 1 if the player is positioned closer.

The third component, component C, is the Euclidian distance between the ball carrier and a teammate. A team is more unpredictable and has a larger chance of creating space in the rear line of an opponent if the area of play is shifted rapidly. Therefore, players who are positioned the furthest from the ball carrier get high scores for this variable, whereas players who are positioned near the player in possession get low scores.

The final component D of Steiner's model, is the defensive coverage of team members by opponents. This variable was operationalised as the shortest distance between any of the opposing players and each team member. The thought behind this component is that players who are covered tightly have less time to make a decision after receiving the ball. This makes them more likely to make an error, often causing the team to lose possession. Therefore, players who have a large distance to their nearest opponent get a high score for this component.

3.2 Normalisation: awarding scores to the components

In this thesis we attempt to create a system to rate players' off the ball performance. Hence, it is essential that players are rated for each of the components mentioned in section 3.1 above. Steiner achieves this by awarding each teammate of the ball carrier, besides the goalkeeper, a score between 1 and 10 for each individual component. Steiner does this by using a formula to normalise the scores of each variable, which is displayed below:

$$ComponentRating = \frac{CurrentPlayerValue - Minimumvalue}{MaximumValue - MinimumValue} * 9 + 1 \tag{3.1}$$

The *CurrentPlayerValue* is the value of the variable for the current player whom a score is being calculated for. The minimum value is the player with the lowest value for a component. The maximum value is the highest value that a player has achieved for a component at the moment the pass was given. After having properly applied the formula, at least one player will have a score of 1 and the other player will have a score of 10 for a certain component. The remaining players will have values between these extremes for this variable.

After the scores were calculated for each individual component, we decided to create a target variable which covers the overall positioning scores of the players based on the four components. Due to this not being covered in Steiner's model, more information about this implementation can be read in the final paragraph of section 3.3, "Changes to Steiner's model".

3.3 Changes to Steiner's model

For the automated system that is created in this thesis, Steiner's variables function as the core components. However, some components were adjusted slightly. The underlying reasoning is that these adjustments were thought to be beneficial for the quality of the final system.

One of the variables in Steiner's model is a player's relative distance to the opponent's rear line compared to that of the ball carrier. Based on this criterium, players are awarded a binary score for this variable. A player gets a 1 if the player is positioned closer to the rear line than the ball carrier and a 0 if the player is positioned further from it. The first adaptation that was made to this variable is that the distance to the rear line will no longer be measured. In this thesis, the distance from a player to the opponent's goal will be measured instead. The main objective within football is to score goals and the smaller the distance to the goal becomes, the higher the probability of scoring a goal is believed to be. Furthermore, due to all the other variables being ranked from 1 to 10, players that are positioned closer to the goal will get a 10 and players that are positioned further from the goal will be awarded a 1.

Another variable that has been altered is the angle between the passing line (the straight between ball carrier and receiver) and the straight to the opponent standing closest to this passing lane. Sometimes, it might occur that there is no opponent that has a straight to the passing lane. In this case, the receiver will receive a 10 for this variable. In this circumstance, no opponent was positioned in such a way as to be able to intercept the ball. Hence, the passing lane is considered open, which is awarded with the maximum score.

Finally, in order to create one variable that represents the positioning of a player, the average score of all the components will be taken and will be stored in a variable called current positioning rating. This score will, just like the components that are used to describe it, have a value between 1 and 10. The overall positioning score is calculated by applying the following formula to the calculated variables:

$$PositioningRating = \frac{angleOpponentToPasslineRating + distanceToGoalRating + distanceToBallCarrierRating + distanceToOpponentRating}{4} \tag{3.2}$$

Moreover, besides the scores awarded based on Steiner's situational model, positioning scores are also awarded to players by using the population-based ranking method. More information about this method can be read in section 3.4.

3.4 Population-based method

In Steiners model, which is described in section 2.3, players are rated on the basis of their position relative to each other at the moment a pass is given during a match. However, there are times at which the score rewarded for a component does not give a fair reflection of how well the player is actually positioned. For example, if all the players have a large distance to the goal at the moment a pass is given, a player could get a 10 for this component. In other situations, the exact same position on the field might provide the player with a lower score. By using computer science, the players can be awarded points by applying a population-based method. With this method, the position of a player will be compared to the positions of all the players who have received the ball in every match of the Dutch National Team for which spatiotemporal tracking data is available. In other words, besides the situation-based method applied by Steiner, a population-based method will also be added to the pipeline to rate off the ball performance of individual players.

To achieve this, the absolute values for Steiners components for all the players who have received the ball are extracted from the available matches. After all the passes have been obtained from the database, a ranking is made for each component. These rankings have been assembled by creating equally sized bins with Python. To closely represent Steiners model, a total of 9 different ranks have been created for every feature. If a player is awarded rank 1 for a component, the players score for a component is among the best 11.11% scores that a player has received in all the matches that have been processed. As soon as all the players have received rankings for the four components that represent their positioning, an overall positioning score is calculated on the basis of the components. This component is calculated by adding the rankings of all the components up. The best positioning score is achieved if a player receives rank 1 for every component. In this case, the positioning score of a player will be 1 + 1 + 1 + 1 = 4. In the worst case, a player has received rank 9 for each component. This would result in the player getting a score of 36 for his positioning at that moment. Therefore, the overall positioning ranks that players can get for their positioning vary from 4 to 36.

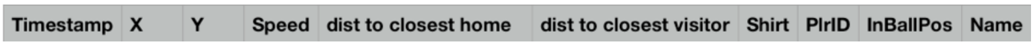
3.5 Data

Our automated system uses data that was provided by the KNVB to analyse players performance. In this section, football data that were provided are discussed. Data used to analyse the football matches in this thesis could be split into two different categories. The first type is positioning data and the second type is event data. Positioning data on the one hand describes where an actor or game object is located at a specific point in time. On the other hand, event data covers math-relevant actions and happenings during the match. More information about the data types can be found in section 3.5.1 and section 3.5.2 below.

3.5.1 Positioning Data

Positioning data used during this project was recorded with the SportVU system developed by STATS. STATS SportVU [STA18] uses a six-camera system installed in football stadiums to record spatiotemporal tracking data of both players and the ball. The system captures data at a rate of 10 Hz. This means that that the positions of players are measured 10 times per second.

Data obtained from the SportVU system are then loaded into Inmotio [Inm18]. This system has been developed by the Austrian technology company Abatec AG and sports scientists of the Dutch company TMP. After Inmotio processes these data, the program exports it to a .csv format. The csv-file that is exported by Inmotio consists of ten different columns. The timestamp is displayed in the first column, which represents the time of the match at which the positioning and further data of either a player or the ball has been recorded. The timestamp is measured with a frequency of 10 Hz. Alongside the timestamp, a binary variable called InBallPos can be found in the dataset, which displays whether the player is or is not in possession at a certain timestamp. Furthermore, positions of the players are recorded with horizontal (X) and vertical (Y) coordinates. Moreover, the speed and the distance to the closest opponent and teammate are exported for each player at each timestamp. In order to be able to recognise who these data belong to, the player's PlayerID, name and shirt number are documened in the CSV-file. The columns of Inmotio's data can be found in Figure 3.2 below.



Timestamp	X	Y	Speed	dist to closest home	dist to closest visitor	Shirt	PlrID	InBallPos	Name

Figure 3.2: Columns of the positioning data file (.csv) exported from Inmotio

3.5.2 Event data

The events during a match are processed and exported by the SportVU system in XML-format. For each event, the coordinates of the ball are recorded, alongside the time at which the event took place. The most important events that are exported by the SportVU system are ball events, match events and event results. The events that are exported by the system can be categorised into three different groups. The first category is called "ball events" and contains all the actions that can take place after a player has touched the ball. Examples of ball events are clearances, catches by the goalkeeper, passes, receptions and shots. The second category is "match events". The events listed in this section are events that indicate an interruption of the game. For example, the end of a half, the ball being out of play and substitutions. The final category covers the results of events. Results of the events can be ground duels, corner kicks, throw-ins and air duels. For this thesis, especially event data about passes will be useful to determine the quality of a player's positioning during build-up play.

3.6 Pipeline

To process spatiotemporal tracking data in this thesis, a full pipeline is used to fully exploit the Data Mining possibilities on football data. This pipeline has partially been constructed by one of the mentors of this thesis, L.A. Meerhoff. He has been doing research to find the structural differences and similarities in playing styles between football played in Brazil and in the Netherlands. The data processed in Meerhoff's project is similar to the data that was provided by the Royal Dutch Football Association (KNVB) for this thesis. Therefore, the automated system to rate off the ball performance of players during an attack that has been created in this thesis, uses the same framework.

The framework that is used was programmed in Python and consists of 7 different components. These components are the storage (A), clean-up (B), spatial-aggregation (C), event computation (D), temporal aggregation (E), match statistics (F) and control panel (G). However, for this thesis, the most relevant processes for analysing the spatiotemporal tracking data of individual football players to rate their off the ball performance takes place in the processing section of the pipeline (C, D, E).

In order to separate the code that had been written during the Meerhoff's project from the code written for this thesis, student modules were created and implemented in the framework. In these modules, the required code to answer the research question of this thesis is written. The student modules consist of different files, each dealing with a different component at a different stage in the pipeline. Due to the spatial aggregation, event computation and temporal aggregation phases being the most important altered parts of the pipeline, they will be discussed into further detail in sections 3.6.1, 3.6.2 and 3.6.3 respectively.

3.6.1 Spatial aggregation

In the spatial aggregation section, the position of a player at a certain timestamp is summarised into a single measure. Each spatial aggregate is a time series describing the development of that specific measure over time. The components of Steiner were implemented in the spatial aggregation module. For each timestamp, the absolute values for each individual component of Steiner's model (e.g. the distance to the closest visitor) are measured. Thereafter, the normalised scores for each individual attribute are calculated for all the timestamps on which a pass occurred. All the passes given during a match, alongside the timestamp on which they occurred, can be retrieved from event data provided by Amisco. The formula used to calculate the normalised ratings for each individual component is described in section 3.2.

3.6.2 Event computation

During the event computation phase of the pipeline, events are recognised and recorded on the basis of predefined criteria. For the validation of Steiner's method, the ratings of individual players who received the ball during a build-up are linked to the outcome of a build-up. To perform this validation, all the build-ups that occurred during a match are recognised and recorded during the compute events phase of the pipeline. This is done by listing the start and end time (t_{start} and t_{end}) of the event that took place, alongside the team that was in possession during the attack ($refteam$). More information about the definition of a build-up can be read in section 3.7.2.

3.6.3 Temporal aggregation

In the temporal aggregation phase of the program, the process takes place that deals with the temporal component of spatiotemporal tracking data by summarising the spatial aggregates around the computed events with varying time windows. In this case, each individual component is examined over the course of a build-up. For each build-up, among other measurements, the average, minimum and maximum values for each variable are measured. The temporally aggregated data is exported into a tabular format with the multiple occurrences of the event of interest in the rows (the build-ups) and a broad range of aggregated features describing the events in the columns. Data obtained during this phase is essential for the validation of the system. More information about the validation process can be read in section 3.7.

3.7 Validation

To validate the model, the positioning scores of players will be analysed during build-ups. The model can be considered valid if a correlation with the position scores and the attack outcome exists.

3.7.1 Hypothesis

In principle, one would expect an attack to have a higher chance of succeeding if the players who receive the ball during the build-up are positioned well. The validation of the models will therefore be based on this hypothesis. Moreover, it is to be expected that differences in the positioning scores of players will mainly be visible when solely looking at the final pass of a build-up. If an attack ends without a shot having occurred, this means that possession was lost at an earlier stage. This could mean that the player who received when the final pass of the build-up was played was properly marked and therefore lost possession. A low overall score for the positioning of the player receiving the ball is therefore expected if possession is lost without a shot having taken place. If an attack resulted in a shot, this usually means that the final pass was a good one, due to it having created a situation which according to the ball carrier was promising enough to shoot. Therefore, a relatively high score for the positioning of the player is expected to be obtained from our system's output data. In order to implement the validation method, all the build-ups have to be obtained from the data. To extract them from the dataset, the definition of a build-up described in section 3.7.2 is used.

3.7.2 Definition build-up

A build-up begins if the following three requirements are met. Firstly, the ball should be on the field. Secondly, the speed of the ball should not be equal to zero. Finally, it is important that the previous build-up ended before a new attack is allowed to start.

As soon as the beginning of the attack has been defined, the ending of the attack has to be determined. An attack ends if either of the following occurs. As soon as a player of the opposition is in possession, the attack has ended, due to the ball having been intercepted. Another way to determine the end of the attack is by checking whether the attacking team has not possessed the ball for at least 5 seconds. If this is the case, the attack has ended. The same applies if either the ball has been out of the field for at least 0.5 seconds or the ball has either not or barely moved in a timespan of 5 seconds. If the latter occurs, the referee has probably stopped play to (for example) award a free kick to a team. Finally, an attack ends if the referee blows for either half time or full time.

After all the build-ups that meet the above-mentioned requirements have been extracted from the dataset, one final measure is used to filter out the build-ups that are not considered valuable for the validation of the system. The build-ups that are used for the validation should consist of at least one pass. Passes to the goalkeeper and throw-ins will not be counted when checking if this measure is met.

3.7.3 Labelling outcome build-up

To determine the outcome of a build-up, an ordinal target value has been awarded to each attack. This variable can take five different values. These values are no final third, final third, shot not on target, shot on target and goal. If an attack gets the label no final third, the player who received the final pass of the build-up was not positioned in the final third of the field at the moment the pass was given.

3.7.4 Data Mining

After having labelled the results within the pipeline, it is time to visualise and examine the results. The visualization will firstly be done by creating box plots and histograms of the ranking methods. Furthermore, either ANOVA or Kruskal Wallis will be applied to determine whether the obtained samples for each analysed ranking method are significantly different. ANOVA allows analysing a model in more detail. Therefore, ANOVA is strongly preferred to Kruskal Wallis. However, ANOVA is a parametric method and is therefore based on the data being normally distributed. Therefore, if-and-only-if the data is not normally distributed, the weaker Kruskal Wallis test is used. This means that before either of these methods can be applied, the normality of the data has to be tested. This is done by applying the D'Agostino-Pearson omnibus test. More information about ANOVA and Kruskal Wallis can be read in section 3.7.4.1 and 3.7.4.2 respectively. After having examined the data, subgroup discovery will be used to dive into the individual components exported by the pipeline with more depth. More information about this subgroup discovery can be read in section 3.7.4.3.

3.7.4.1 Analysis of Variance (ANOVA)

ANOVA (Analysis of Variance) is a statistical technique for testing if $3(+)$ population means are all equal. During this thesis, one-way ANOVA will be used to determine if the means of all different implemented methods to rate players off the ball performance are the same when linked to the ordinal target variable (the five different outcomes of a build-up). [Sta19]

In ANOVA, the null-hypothesis is always that all population means are equal. Therefore, the null-hypothesis of ANOVA is the following (see equation 3.3):

$$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_k \tag{3.3}$$

In this formula, $\mu =$ group mean and $k =$ number of groups. The probability of obtaining an outcome if the sample means were to be the same should not be less than a predefined significance value (α). During this thesis, $\alpha = 0.05$ is used. It is important to note that one-way ANOVA is an omnibus test statistic and does not tell which groups were statistically significantly different from each other, only that at least two groups were. To determine which specific groups differed from each other, a post hoc test has to be used. In ANOVA, the null-hypothesis is always that all population means are equal. If this holds, then the sample means will probably differ a bit. However, the probability of obtaining an outcome if the sample means were to be the same should not be less than a predefined significance value (α). During this thesis, $\alpha = 0.05$ is used.

In order to indicate precisely how different the sample means are, the variance among the samples is calculated. This is done by calculating the sum of the squared deviations between the three sample means and the overall mean. The outcome is known as the sum of squares between. The sums of squares between expresses the total amount of dispersion among the sample means.

If the means of all the five ordinal target variables (the possible outcomes of a build-up) describing off the ball performance of individual football players on the basis of Steiner's (altered) model are the same according to ANOVA, then it would mean that Steiner's model does not have a predictive value for the outcome of a build-up. Therefore, ANOVA is a suitable statistical test to check whether Steiner's model has a distinctive value for the ordinal target variable, which forms the core of the validation of the model.

3.7.4.2 Kruskal Wallis

The Kruskal Wallis test is the non-parametric alternative to one way ANOVA. Non-parametric means that the test does not assume that the data comes from a particular distribution. Kruskal Wallis is used when the assumptions for ANOVA are not met (like the assumptions of normality). [Ste16]

In Kruskal Wallis, the test is based on determining whether the medians of two or more groups are different. Just like for one way ANOVA, the hypotheses for the test are:

H0: population medians are equal
H1: population medians are not equal

As is the case with ANOVA, Kruskal Wallis will only tell whether there is a significant difference between the samples. It will not tell which sample(s) cause(s) the probability of obtaining the sample distributions to be less than α. During this thesis, $\alpha = 0.05$ will be used.

3.7.4.3 Subgroup Discovery

Subgroup discovery has been established as a general and broadly applicable technique for descriptive and exploratory data mining. It aims at identifying descriptions of subsets of a dataset that show an interesting behaviour regarding certain criteria, formalised by a quality function. [Atz15] Within this technique, we distinguish two different types of data. Nominal data, for which a classification setting, and numeric data, for which a regression setting is used. Before diving into the benefits of discovering subgroups, it is essential to have a clear understanding of what a subgroup is. A subgroup is a subset of the data that is described by a certain rule. This rule can be formulated using a selected language. For example, the distance within X meters of the goal. When looking at the entire dataset, which consists of all the attacks that occurred during the 30 analysed matches of the Dutch National Team, only 0.96% of the attacks resulted in a goal. However, the percentage of attacks resulting in a goal increases if we look at attacks that ended within a X meters from the goal. Therefore, for this particular target (attacks resulting in goals), this subgroup has a distribution that stands out compared to that of the entire dataset. To express to which extent this is the case, the quality measure is used to determine when exactly a distribution of the target variable in a subgroup is significantly different from that of the rest of the dataset. More about this quality measure can be read in section 3.7.4.3.

Cortana
Cortana is a data mining tool for discovering local patterns in data. Cortana features a generic subgroup discovery algorithm that can be configured in many ways, in order to implement various forms of local pattern discovery. The tool can deal with a range of data types, both for the input attributes as well as the target attributes, including nominal, numeric and binary. A unique feature of Cortana is its ability to deal with a range of subgroup discovery settings, determined by the type and number of target attributes. Where regular SD algorithms only consider a single target attribute, nominal or sometimes numeric, Cortana is able to deal with targets consisting of multiple attributes, in a setting called Exceptional Model Mining. For the Subgroup Discovery part of this thesis, Cortana will be used to search for and examine the subgroups. [LIA19]

Quality measure
A quality measure is needed to quantify the quality of the subgroup considered. Cortana, which is the program that is used to discovery the subgroups in this thesis, has a set of quality measures implemented in it. During the subgroup discovery in this thesis, the quality measure that is used is called Weighted Relative Accuracy

(WRacc). This is the default quality measure in Cortana and is a suitable quality measure to determine how various qualities of subgroups are combined and weighted in a final metric, which is what is attempted to achieve during this analysis step. Within weighted relative accuracy, a balance between coverage and unexpectedness of a subgroup is taken into consideration. The algorithm is described as follows:

$$WRacc(S,T) = P(ST) - P(S) * P(T) \tag{3.4}$$

In WRacc, for which the formula is mentioned in equation 3.4 above, the quality of a subgroup is quantified as the product of the subgroup size and the distributional deviation within the subgroup compared to that of the overall distribution of the target.

Search Depth

Another setting that can be adjusted in Cortana is the search depth used when exploring data. The search depth d determines the maximum number of conditions (and hence the number of attributes involved) in the subgroup. During the process of subgroup discovery, a tree is created with various different calculations performed on all the different features in the dataset. At each level in the tree, an extra variable is added to the subgroup's criteria. For example, at depth 1, the subgroups that had been created on the basis of one variable. At level 3, the subgroups were based on three different variables, etc.

Search method

For the search method, there are two types of methods that could be chosen. The first option is a heuristic approach, which at each step chooses the (seemingly) best option at the moment the decision has to be made, and an exhaustive method, which functions by systematically enumerating all possible candidates for the solution and checking whether each candidate satisfies the problem's statement. Due to the large number of examples in the dataset, the heuristic method is the preferred option. The heuristic method that has been selected within Cortana for the analysis in this thesis is called beam search. Within beam search, at each search level within the tree, only the top w subgroups are considered for refinement. This w is the search width and can be altered in the settings window of Cortana. For our research, a search width of $w = 100$ was used.

Swap randomisation

An essential step within subgroup discovery is to compute a quality threshold that subgroups need to meet in order to be considered significant. Within Cortana, this can be done by using so-called swap-randomisation. The significance of a subgroup is based on the likelihood that a subgroup is found with the same quality in a copy of the dataset where the target is randomised by swapping between rows. First, the dataset is swap-randomised, which means that any possible relationships between the attributes and the target are severed. Then, subgroups can be discovered on this randomised dataset. By repeating this process x times, a distribution of subgroups that were falsely thought to be interesting can be filtered from the results. For this thesis, $x = 100$ was used. From these filtered out subgroups, a threshold can be computed that draws a line between statistically significant results and accidental findings. By doing this, the results that are found during the subgroup discovery of the actual data are all better than subgroups found accidentally.

Chapter 4

Results

4.1 Box plots

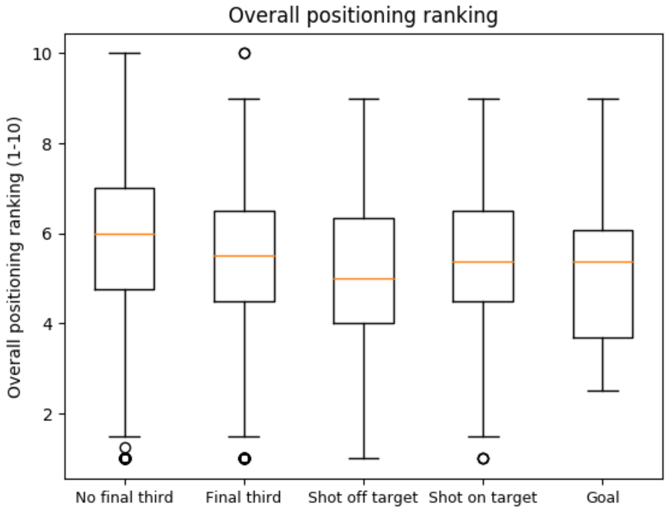


Figure 4.1: Box plot of the average overall positioning ranking of the players who received the ball during a build-up

Firstly, the data is visualised by creating box plots to get a better view on the results that are being analysed in this chapter. An example of a box plot that is analysed in this thesis can be read in Figure 4.1 above. In this box plot, it becomes clear that the spread of the data is very large and that the boxes have a large overlap. Moreover, the means mostly seem to be very much alike. Finally, the sizes of the boxes seem to be approximately the same when looking at this Figure. This is not only the case for the average overall positioning ranking, but also for the other ranking methods. The box plots of the other ranking methods can be found in the appendix of this thesis. To analyse whether the samples are indeed as identical as they are though to be when viewing the box plots, either ANOVA or Kruskal Wallis will be applied on the data. However, to decide which test to pick, section 4.2 below will cover the normality test that was done.

4.2 Normality test

To determine the normality of the six samples that are examined, the D'Agostino-Pearson omnibus test is used. In this test, firstly the skewness and kurtosis are computed to quantify how far from Gaussian the distribution is in terms of asymmetry and shape. It then calculates how far each of these values differs from the value expected with a Gaussian distribution, and computes a single P value from the sum of these discrepancies. It is a versatile and powerful normality test, and is therefore used to test the normality of the samples in this thesis. The tests results of the D'Agestino-Pearson omnibus test can be found below in Table 4.1.

Ranking method (average)	**Skewness**	**Kurtosis**	**P-value normality test**
Population-based overall positioning ranking	0.202	0.481	1.234e-12
Overall positioning ranking	-0.438	0.004	1.284e-27
Overall positioning rating	0.278	0.937	1.459e-28
Population-based overall positioning ranking (last pass)	-0.046	-0.284	6.038e-05
Overall positioning ranking (last pass)	-0.181	-1.003	9.259e-216
Overall positioning rating (last pass)	0.175	-0.132	4.518e-06

Table 4.1: The results of the D'Agestino-Pearson normality test for each of the analysed ranking methods

During the test, $\alpha = 0.01$ is used to test the significance of the datasets. Within the D'Agestino-Pearson normality test, the null hypothesis is that the data is distributed normally. Therefore, if $P < 0.05$, this means that the null hypothesis is rejected and the data is considered not to be normally distributed. As can be seen in the table above, all the P-values are less than 0.05. Thus, the data were not normally distributed and the non-parametric method, Kruskal Wallis, was used to examine the means of the ordinal target values within the six samples. The results of Kruskal Wallis and the post-hoc analysis (the Wilcoxon-Mann-Whitney test) can be read in section 4.3 and section 4.4 respectively. Moreover, in the appendix, histograms of the analysed ranking methods can be found in the appendix of this thesis to visualise the distribution of the ranking methods.

4.3 Kruskal Wallis

As has been mentioned in the Chapter Related Works (Chapter 2), Kruskal Wallis is a collection of statistical models and their associated estimation procedures (such as the variation among and between groups) used to analyse differences between groups. In this thesis, Kruskal Wallis will be used to check whether the data points that were assigned to the ordinal target values (the outcomes of a build-up) have equal sample means. If this is the case, then it means that there is no clear distinction between the samples. Thus, the ranking method applied was not able to predict the outcome of a build-up.

In Kruskal Wallis, the null hypothesis (H_0) is always that the means of the groups are the same. This will be tested with a significance level of $0,05$ ($\alpha = 0,05$). This means that if the probability (the p-value) is less than $0,05$ for obtaining the results, the null hypothesis will be rejected. In this case, this would mean that the samples are not the same and that to some extent, the samples offer valuable. The results of the Kruskal Wallis analyses can be found in Table 4.2.

Ranking Method (average)	**H**	**df**	**P-value**
Population-based overall positioning ranking	132.682	4	1.04E-27
Overall positioning ranking	35.671	4	3.38E-7
Overall positioning rating	7.793	4	0.099
Population-based overall positioning ranking (last pass)	118.763	4	9.81E-25
Overall positioning ranking (last pass)	36.814	4	1.97E-7
Overall positioning rating (last pass)	16.091	4	0.002899

Table 4.2: Results of Kruskal Wallis applied on the average scores the players who received the ball during build-ups were awarded

The results from the Kruskal Wallis tests show that, apart from the results for the average overall positioning rating, the p-values for the means being the same are less than 0.05, which is the significance value used for the Kruskall Wallis test. To determine which groups have significantly different means within the ranking methods, a post-hoc analysis has to be applied on the ordinal target values within the other five analysed ranking methods. For Kruskal Wallis, an appropriate post-hoc analysis test is the Wilcoxon-Mann-Whitney test, for which the results can be read in section 4.4 below.

4.4 Wilcoxon-Mann-Whitney

The post-hoc analysis method used to determine which samples caused Kruskal Wallis' null hypothesis to be rejected is the Wilcoxon-Mann-Whitney test. Within this test, the null hypothesis is that the means of the samples are the same. Therefore, if $p \geq \alpha$, the samples are considered to be the same. The results of the Wilcoxon-Mann-Whitney test can be found in Table 4.3.

Ranking method (average)	**Groups compared with WilcoxonMannWhitney test (p-value)**									
	1 vs 2	**1 vs 3**	**1 vs 4**	**1 vs 5**	**2 vs 3**	**2 vs 4**	**2 vs 5**	**3 vs 4**	**3 vs 5**	**4 vs 5**
Population-based overall positioning ranking	*0*	*0*	*0*	*0*	0.192	0.254	0.328	0.119	0.482	0.264
Overall positioning ranking	*0*	*0*	*0.005*	*0.022*	*0.009*	0.399	0.160	0.071	0.448	0.227
Population-based overall positioning ranking (last pass)	*0*	*0*	*0*	*0.006*	0.440	*0.044*	0.268	0.096	0.299	0.424
Overall positioning ranking (last pass)	*0*	*0*	*0.002*	0.101	0.116	0.383	0.494	0.298	0.324	0.426
Overall positioning rating (last pass)	0.349	*0.029*	*0.003*	0.204	*0.016*	*0.001*	0.192	0.125	0.464	0.327

Table 4.3: The results of the Wilcoxon-Mann-Whitney test applied on the ranking methods that had significantly different distributions according to Kruskal Wallis

Interestingly, when viewing the results in Table 4.3 above, the group that mostly causes Kruskal Wallis to find significant differences between the distributions of the samples is group 1. Group 1 covers the attacks that ended outside of the final third. For an attack ending outside of the final third, it is given that the distance to the goal is (mostly) larger than that of attacks ending inside of the final third. Due to most shots being expected to have been taken within the final third, the distance to the goal for group 1 is expected to also (mostly) be larger than that of the other groups. Therefore, the distance to the goal might have a predictive value for the outcome of an attack. Whether this is the case, will be examined in section 4.5, Subgroup Discovery.

4.5 Subgroup Discovery

During subgroup discovery, the computed variables were linked to one of the five possible values of the ordinal target attribute, which is the outcome of a build-up. To do this, five separate analyses were executed on the obtained data, each focusing on one outcome of a build-up as opposed to the retrieved data belonging to the other outcomes of build-ups. In other words, attacks resulting in the loss of ball possession before the final third were compared to all the other possible outcomes of an attack (possession lost in final third, shot not on target, shot on target and goal). For all the possible outcomes, a refinement depth of both one and two were used. The results are stated in section 4.5.1 and section 4.5.2 below. The output for all the subgroup discovery tests can be found in the appendix of this thesis. Moreover, defintions for the features are given in the appendix (Chapter C - Definitions features).

4.5.1 Refinement Depth 1

Firstly, the focus will be on refinement depth 1, which means that the subgroups examined in this section are based on a condition on a single attribute. To give an understanding of what the output of Cortana looks like after having analysed a certain test setting, the table retrieved after analysing target 1 (build-ups resulting in possession being lost outside of the final third) for refinement depth 1 can be found in Table 4.4.

Nr.	Depth	Coverage	Quality	Probability	Positives	Conditions
1	1	2106	0.211	0.962	2025.0	finalDistanceToGoal_avg >= 43.0366
2	1	2632	0.198	0.854	2249.0	finalDistanceToGoal_avg >= 35.917
3	1	1580	0.168	0.986	1558.0	finalDistanceToGoal_avg >= 51.163
4	1	2187	0.167	0.860	1880.0	lastPassPopDistanceToGoalRank_avg >= 4.0
5	1	2106	0.165	0.868	1829.0	lastPassDistanceToGoal_avg >= 48.507
6	1	2632	0.161	0.796	2094.0	lastPassDistanceToGoal_avg >= 42.391
7	1	2677	0.158	0.787	2106.0	lastPassPopDistanceToGoalRank_avg >= 3.0
8	1	2106	0.156	0.849	1788.0	passDistanceToGoal_min >= 45.414
9	1	2457	0.154	0.801	1968.0	passPopDistanceToGoalRank_min >= 3.0
10	1	1926	0.152	0.870	1676.0	passPopDistanceToGoalRank_min >= 4.0

Table 4.4: Example of output from Cortana for refinement depth 1 showing the first 10 subgroups of all the build-ups resulting in possession being lost outside of the final third. Each subgroup is ranked according to the quality measure (WRacc).

At $d = 1$, it becomes clear that the most important variable to predict the outcome of an attack is the distance a player has to the goal. For target variables 1 and 2, stating whether the attack ended either inside or outside of the final third, this is no surprise, due to the distance to the centre of the goal having a very strong correlation with the vertical distance from a player to the opposition's backline. Therefore, the 10 most significant subgroups for both target values consist solely of variables dealing with the distance to goal rank.

However, also for the other target variables the 10 best subgroups were all based on the distance to the goal. This means that, for every type of shot, only the distance a player had to the goal during the build-up offered the most interesting subgroups. Interestingly enough, none of the normalised values that Steiner awarded to the four components was found in the top 50 qualitatively best subgroups according to the subgroups' weighted relative accuracy. This means that as a matter of fact, the other components that Steiner used to create the combined overall positioning measure actually turned out to have a negative effect on being able to predict the outcome of an attack. Perhaps refinement depth 2 will offer more interesting results.

4.5.2 Refinement depth 2

At depth 2, most of the top-ranked subgroups were extensions of the best-ranked subgroups at $d = 1$. This, bearing the results of section 4.5.1 in mind, means that the subgroups found for depth 2 are (partially) based on the distance to the goal. When looking at the results, this assumption turns out to be correct. For target 2, the results are displayed in Table 4.5 below.

Nr.	Depth	Coverage	Quality	Probability	Positives	Conditions
1	1	2106	0.170	0.706	1487	finalDistanceToGoal_avg,<= 43.03657
2	2	2094	0.170	0.708	1482	finalDistanceToGoal_avg,<= 43.03657 AND eventOverlap = 'o'
3	2	2104	0.170	0.706	1485	finalDistanceToGoal_avg,<= 43.03657 AND lastPassDistToPosRank_avg <= 9.0
4	2	2092	0.169	0.707	1480	finalDistanceToGoal_avg <= 43.03657 AND distanceToOpponentRank_min <= 1.0
5	2	2081	0.169	0.709	1475	finalDistanceToGoal_avg <= 43.03657 AND distanceToOpponentGoalRating_min <= 1.0
6	2	1975	0.169	0.727	1436	finalDistanceToGoal_avg <= 51.162743 AND finalDistanceToGoal_avg <= 41.013836
7	2	1975	0.169	0.727	1436	finalDistanceToGoal_avg,<= 60.2957 AND finalDistanceToGoal_avg <= 41.013836
8	2	2093	0.169	0.727	1479	finalDistanceToGoal_avg,<= 43.03657 AND distanceToOpponentGoalRating_max >= 10.0
9	2	2083	0.168	0.707	1469	finalDistanceToGoal_avg,<= 43.03657 AND passPopDistToOpRank_min >= 1.0
10	2	2069	0.167	0.705	1460	finalDistanceToGoal_avg,<= 43.03657 AND curPositioningRank_min <= 1.0

Table 4.5: Example of output from Cortana for refinement depth 2 showing the first 10 subgroups of all the build-ups resulting in possession being lost within the final third. Each subgroup is ranked according to the quality measure (WRacc).

In Table 4.5 and the other Tables with refinement depth 2 that can be found in the appendix, it becomes clear that even for refinement depth 2 the distance to the goal offers the qualitative best subgroups. Furthermore, for refinement depth 2, it becomes visible that most highly ranked subgroups are based on conditions for either the final pass of a build-up or the final distance of a build-up. This could mean that the outcome of an attack is largely based on the last pass given during it, rather than the entire sequence preceding it. However, more about this can be read in the discussion (chapter 5). Interestingly enough, again, none of the top-rated subgroups consist of a condition performed on the overall positioning rating based on Steiner's normalised model. This means that for distinguishing the outcomes of a build-up, variables that Steiner's model was based on, turned out to have a more predictive value for the outcome of an attack than the model itself. This is in line with the results from Kruskal Wallis' post-hoc analysis (Wilcoxon-Mann-Whitney), in which the differences between the samples mostly only were significant when comparing samples to target group 1, in which it was given that the distance to the goal was larger than that of the other four target values. Therefore, based on this analysis, Steiner's method is not thought to be an optimal method to predict the outcome of an attack. However, more about this can be read in the discussion and the conclusion of this thesis.

Chapter 5

Discussion

As can be read in chapter 4, the results of the experiments were not in line with the expectations and therefore, we were not able to validate our automated system to rate off-the-ball performance of individual football players with our proposed validation method. Several data mining methods have been applied. Here after, the exported data from the pipeline have again been examined by using box plots, Kruskall Wallis, post-hoc analysis and subgroup discovery. Hence, many data mining techniques were applied to use the information within the data to its full extent. The idea behind the steps that were taken in the process of analysing the data were all well thought over and should probably all be used for the creation of automated systems to analyse football data in the near future. However, the system is not yet perfect and improvements can still be made to obtain better results. The sections below will cover aspects of the system that could and probably should be improved upon in the future.

5.1 Definition of ball possession

The player in possession was extracted from the data provided by Inmotio for each timestamp. To determine the player in possession, Inmotio used a definition for ball possession which can be read in appendix B. However, the problem with this definition is that it does not always deliver proper results. This becomes clear when comparing the player in possession according to Inmotio to the player in possession when viewing video footage of matches.

At times, Inmotio unfairly thought that a player was in possession of the ball. This occurs when a player passes the ball closely along an opponent, causing Inmotio to think that the opponent is in possession of the ball. Because of this, actual long build-ups consisting of many passes are at times split into multiple shorter build-up plays. Due to build-ups only being taken into consideration if they consist of at least one pass (according to the definition of a build-up in section 3.7.2), this means that at times, among other things, goals are lost from the output data. Hence, the poor definition of ball possession influenced the validation of our system. In the newest edition of the pipeline, a new definition of ball possession was added. Unfortunately, this method was added after the analysis for this thesis was done. In the future, this new definition of ball possession should probably be used to obtain better results.

5.2 Validating according to entire build-ups

For the validation of our automated system, the scores players received for their positioning components during build-up play were linked to the outcomes of these build-ups. However, perhaps this method was not the most suitable option to validate the system. In section 5.2.1 and section 5.2.2, two cons of the implemented validation method are described.

5.2.1 Good vs. bad passes

Perhaps Steiner's model is useful to judge whether a pass will reach its target as opposed to judging what the pass' influence is expected to be on the remainder of the attack. If the former is the case, then the applied validation method offers a skewed division between good and bad passes. Perhaps this part of Steiner's model should be validated by implementing another validation method in the future.

5.2.2 Quality of individual players

During the validation process, currently the qualities of the players receiving the ball are neglected. However, the qualities of players could have a large influence on whether the pass either does or does not reach its target or whether a shot either does or does not result in a goal. If, for example, the final pass of an attack was played to a player who was positioned well, the player could be unable to control the ball due to a personal mistake nonetheless.

Furthermore, the pass could be played poorly by the ball carrier. If this occurs, although the player on the receiving end was awarded high scores for his positioning, possession could still be lost. Personal mistakes are a common factor in football and due to our automated system not taking into consideration the individual qualities of the players receiving the ball, validating the system based on the outcomes of an attack was probably not ideal.

5.3 Mistakes in the implementation

Although the components of Steiner's model were all checked by comparing the outcomes to actual video footage, there is always a chance that in specific situations, the system does not compute results correctly. If passes are not valued correctly, then the validation will be done on incorrect exported results, causing the validation to be unreliable. However, as was mentioned in this section's first sentence, no faulty situations were found when checking the system's outputted results.

5.4 Static snapshots

Steiner's model is based on static snapshots that were taken from matches. These snapshots are not a fair representation of an actual situation in a football match, due to various dynamic aspects of the game not being taken into consideration. For example, the distance of a player to the closest defender is not necessarily the only important part to describe how well a player is marked. The speed at which this opponent is running towards the player receiving the ball is also decisive to define the pressure that is exerted by the opposing player. For example, if a player is running towards the player receiving the ball at full speed, the player needs less time to cover the distance between himself and the potential new ball carrier. Other important aspects are a player's relative speed to that of the ball and the relative position of the defender to that of the potential ball receiver. Perhaps if Steiner's model is expanded by adding these dynamic features to the already existing components, the model would function better in an actual match environment. Therefore, in the future, the automated model should be extended with these features to improve it.

Chapter 6

Conclusion

The research question of this thesis reads *"How can off-the-ball performance of individual football players be analysed using spatiotemporal tracking data, in a way that generates valuable information for the coach?"*. In this section, the answer to this research question will be deduced on the basis of conclusions drawn from the results obtained in chapter 4, Results.

In conclusion, the process of creating the automated system was an informative and well thought out process. A full pipeline was used in which a a set of data mining techniques and data analysis tools were applied to the data. It turns this has not yet led to valid results. However, this does not mean that the overall applied method is not at least a step in the right direction with regards to analysing football data.

Steiner's model looked promising, but did not quite offer the valuable information to the coach that was hoped for. Perhaps, in the future, dynamical aspects of the game should be added to the model, such as the acceleration and speed of players and the ball and the heart rate of players during the match. The reason being that for example, the distance a player may have to be considered a good passing option is largely influenced by the speed at which the opponent is moving towards this player. In the current model, this information is unfortunately neglected.

All in all, this thesis covers an advanced methodology to create a fully functioning system to rate off-the-ball performance of football players using spatiotemporal tracking data. Although improvements can be made to the system, the methods that were applied are revolutionary and give new insights on ways football data can be analysed in the future. The main issue is that at this moment not the used methodology, but that it is still difficult to describe proper criteria to indicate when a player is well positioned. As soon as more information about positioning and other aspects of football are better defined, data science is expected to have a large influence on the way football is be played.

Bibliography

[Atz15] Martin Atzmueller. Subgroup discovery. *Wiley Int. Rev. Data Min. and Knowl. Disc.*, 5(1):35–49, January 2015.

[Bra17] L. Bransen. Valuing passes in football using ball event data. `http://hdl.handle.net/2105/41346`, 2017. [Online; accessed 26-January-2019].

[Dat18] Anusuya Datta. GPS in soccer: How teams are using wearables in the Russia World Cup — Geospatial World. `https://www.geospatialworld.net/blogs/gps-in-soccer-wearables/`, 2018. [Online; accessed 15-November-2018].

[FIF18a] FIFA. FIFA - Official Website. `https://www.fifa.com/`, 2018. [Online; accessed 15-November-2018].

[FIF18b] FIFA. FIFA World Cup - Official Website. `https://www.fifa.com/worldcup/`, 2018. [Online; accessed 15-November-2018].

[Inm18] Inmotio. About us — Inmotio. `http://www.inmotio.eu/en-GB/8/about-us.html`, 2018. [Online; accessed 15-November-2018].

[LIA19] Data Mining LIACS. Cortana Subgroup Discovery. `http://datamining.liacs.nl/cortana.html`, 2019. [Online; accessed 9-January-2019].

[SJS$^+$17] Manuel Stein, Halldr Janetzko, Daniel Seebacher, Alexander Jger, Manuel Nagel, Jrgen Hlsch, Sven Kosub, Tobias Schreck, Daniel A. Keim, and Michael Grossniklaus. How to make sense of team sport data: From acquisition to data modeling and research aspects. *Data*, 2(1), 2017.

[Spo18a] Opta Sports. Advanced Metrics — Opta Sports. `https://www.optasports.com/services/analytics/advanced-metrics/`, 2018. [Online; accessed 15-November-2018].

[Spo18b] Opta Sports. Home — Opta Sports. `https://www.optasports.com/`, 2018. [Online; accessed 15-November-2018].

[Spo19] Opta Sports. Our History — Opta Sports. `https://www.optasports.com/about/our-history/`, 2019. [Online; accessed 9-January-2019].

[STA18] STATS. STATS SportVU football player tracking. `https://www.stats.com/sportvu-football/`, 2018. [Online; accessed 15-November-2018].

[Sta19] Laerd Statistics. One-way ANOVA. https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php, 2019. [Online; accessed 9-January-2019].

[Ste16] Stephanie. Kruskal Wallis H Test: Definition, Examples Assumptions. https://www.statisticshowto.datasciencecentral.com/kruskal-wallis, 2016. [Online; accessed 9-January-2019].

[Ste18] Silvan Steiner. Passing decisions in football: Introducing an empirical approach to estimating the effects of perceptual information and associative knowledge. *Frontiers in Psychology*, 9:361, 2018.

[vMOS17] Mariette van Maarseveen, Raul Oudejans, and Geert Savelsbergh. System for notational analysis in small-sided soccer games. *International Journal of Sports Science Coaching*, 12, 03 2017.

Appendices

Appendix A

Box plots

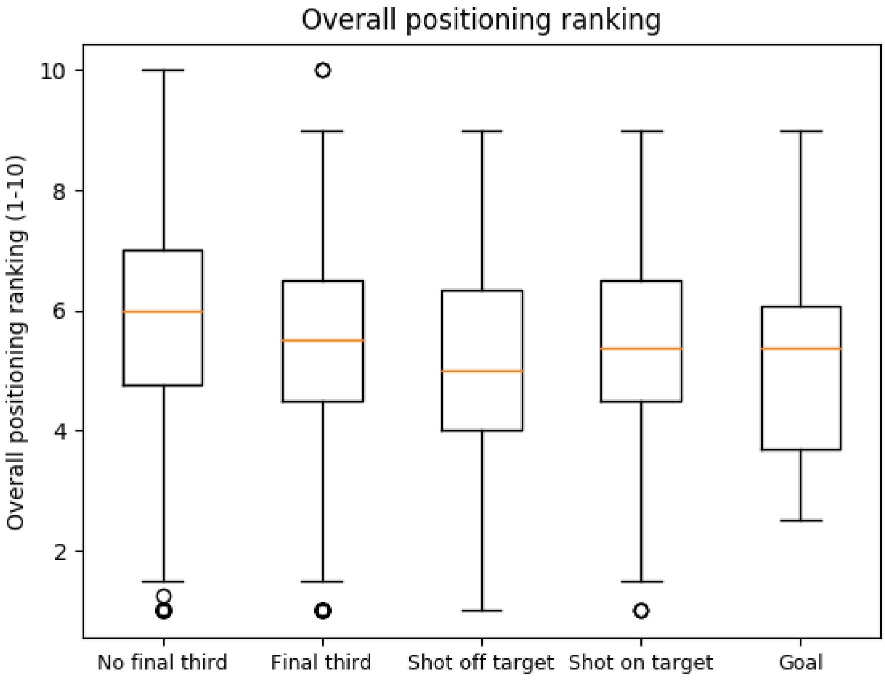


Figure A.1: Box plot of the average overall positioning ranking of the players who received the ball during a build-up

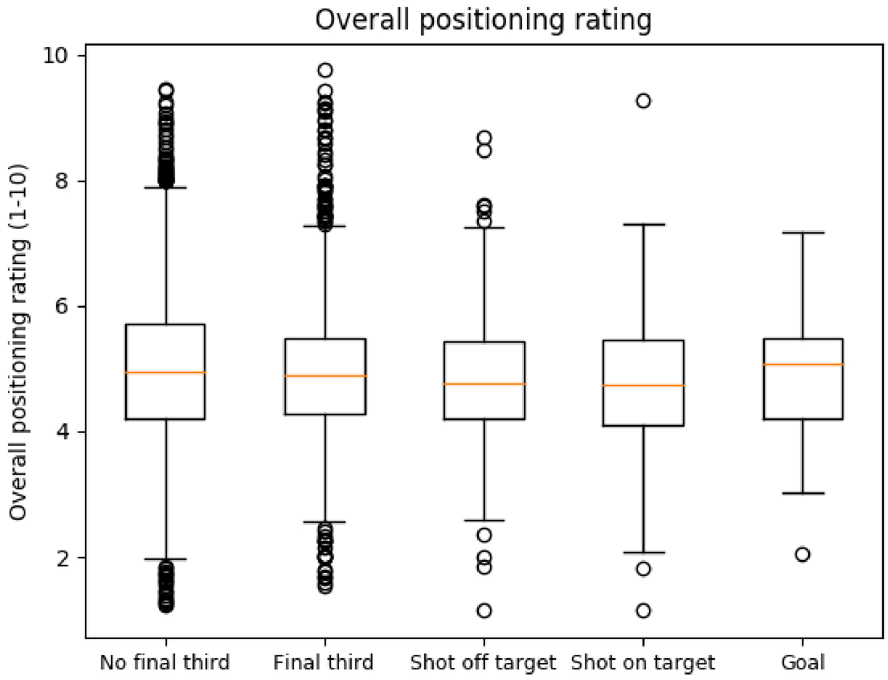


Figure A.2: Box plot of the average overall positioning ratings of the players who received the ball during a build-up

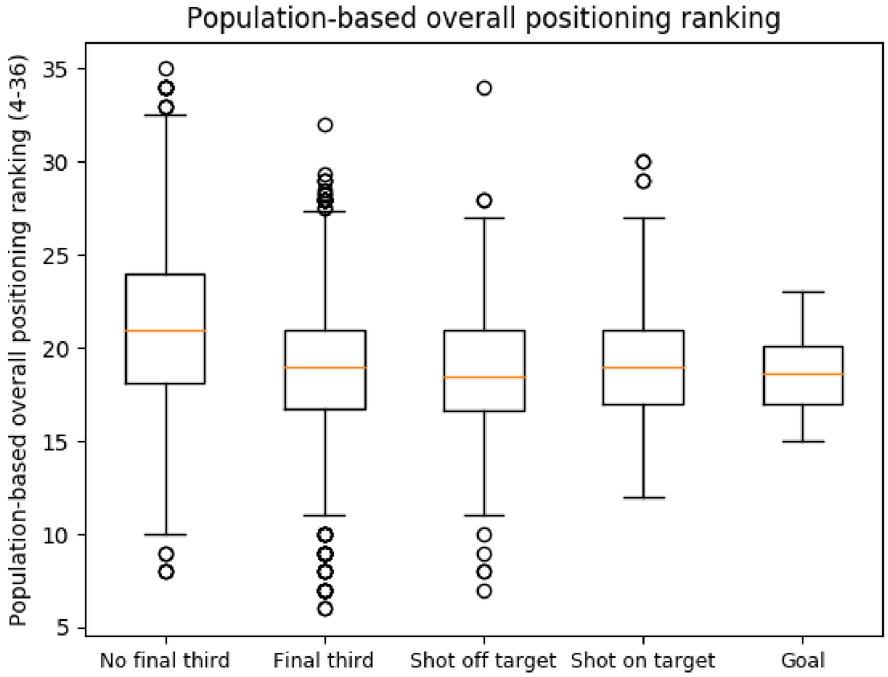


Figure A.3: Box plot of the average population-based overall positioning ranking of the players who received the ball during a build-up

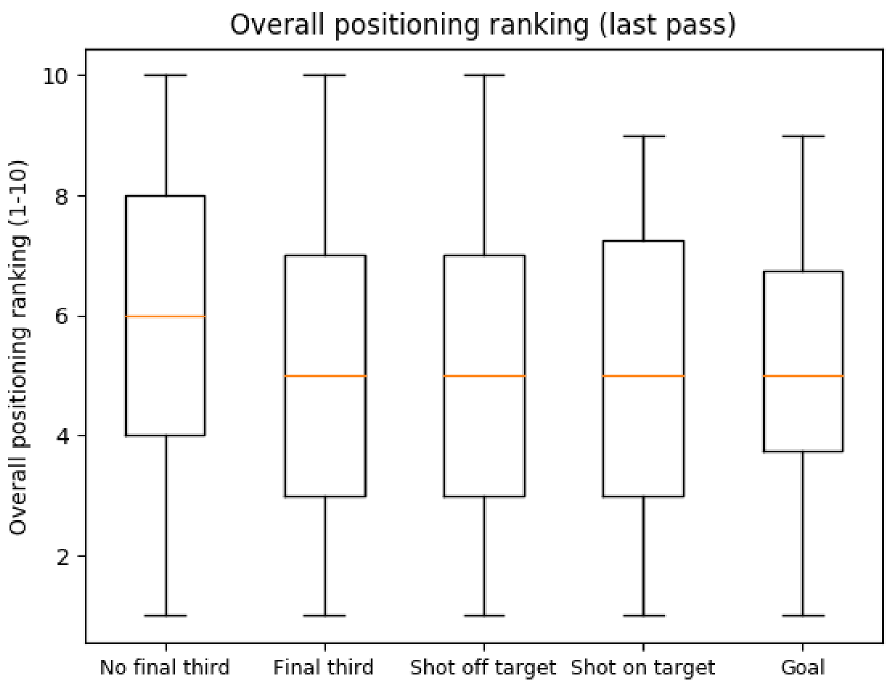


Figure A.4: Box plot of the average overall positioning ranking of the players who received the final pass during a build-up

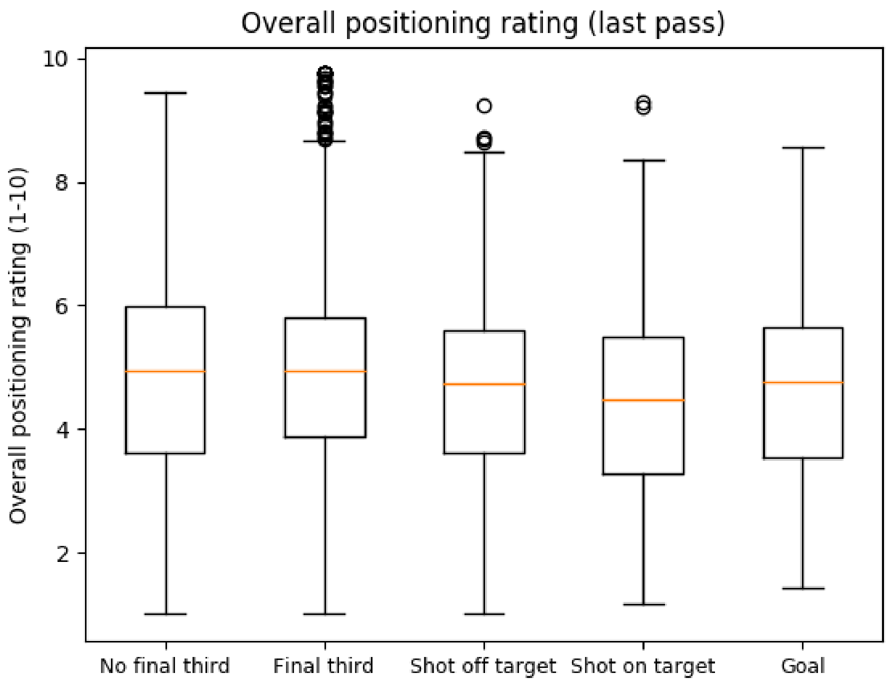


Figure A.5: Box plot of the average overall positioning rating of the players who received the final pass during a build-up

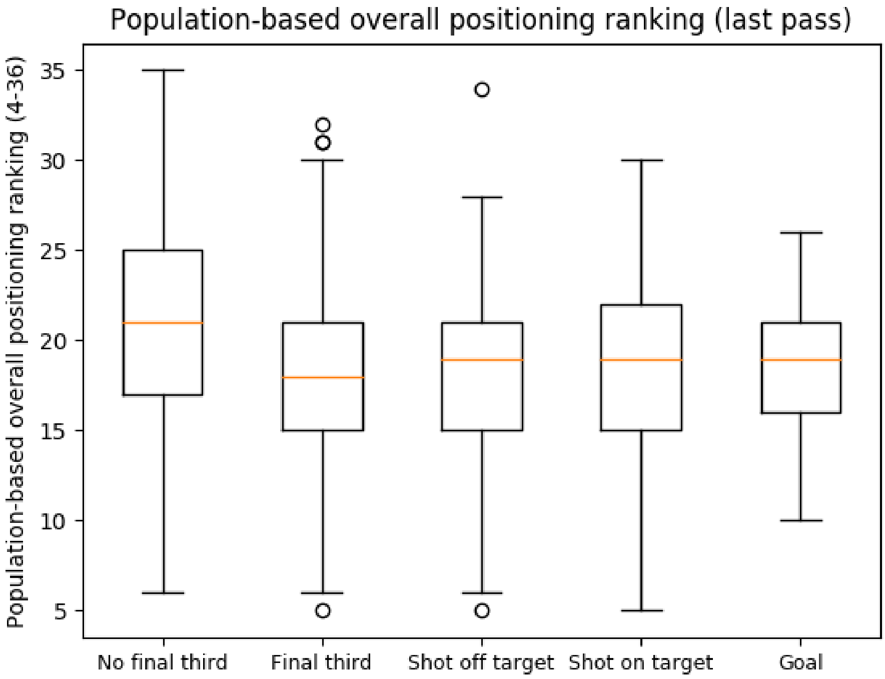


Figure A.6: Box plot of the average overall positioning ranking of the players who received the final pass during a build-up

Appendix B

Results Subgroup Discovery

Nr.	Depth	Coverage	Quality	Probability	Positives	Conditions
1	1	2106	0.21188104152679443	0.9615384615384616	2025.0	finalDistanceToGoal_avg >= 43.03657
2	1	2632	0.1978883445262909	0.854483282674772	2249.0	finalDistanceToGoal_avg >= 35.917004
3	1	1580	0.16816775500774384	0.9860759493670886	1558.0	finalDistanceToGoal_avg >= 51.162743
4	1	2187	0.1671012043952942	0.8596250571559213	1880.0	lastPassPopDistanceToGoalRank_avg >= 4.0
5	1	2106	0.16533628106117249	0.8684710351377019	1829.0	lastPassDistanceToGoal_avg >= 48.50716
6	1	2632	0.161080002784729	0.7955927051671733	2094.0	lastPassDistanceToGoal_avg >= 42.39138
7	1	2677	0.15818175673484802	0.786701531565185	2106.0	lastPassPopDistanceToGoalRank_avg >= 3.0
8	1	2106	0.15559989213943481	0.8490028490028491	1788.0	passDistanceToGoal_min >= 45.41478
9	1	2457	0.15351134538650513	0.800976800976801	1968.0	passPopDistanceToGoalRank_min >= 3.0
10	1	1926	0.1519945114850998	0.8701973001038421	1676.0	passPopDistanceToGoalRank_min >= 4.0

Table B.1: The first 10 subgroups found by Cortana for target 1 refinement depth 1

Nr.	Depth	Coverage	Quality	Probability	Positives	Conditions
1	2	2329	0.21213126182556152	0.9214255045083727	2146.0	passPopDistanceToGoalRank_avg >= 2.0 AND finalDistanceToGoal_avg >= 39.65273
2	2	2303	0.21212762594223022	0.9257490230134607	2132.0	finalDistanceToGoal_avg >= 35.917004 AND finalDistanceToGoal_avg >= 40.145752
3	2	2304	0.21199989318847656	0.9253472222222222	2132.0	finalDistanceToGoal_avg >= 22.04955 AND finalDistanceToGoal_avg >= 40.126244
4	2	2304	0.21199989318847656	0.9253472222222222	2132.0	lastPassDistanceToGoal_avg >= 30.143642 AND finalDistanceToGoal_avg >= 40.010143
5	1	2106	0.21188104152679443	0.9615384615384616	2025.0	finalDistanceToGoal_avg >= 43.03657
6	2	2304	0.2117624282836914	0.9249131944444444	2131.0	passDistanceToGoal_avg >= 36.261517 AND finalDistanceToGoal_avg >= 39.929535
7	2	2304	0.2117624282836914	0.9249131944444444	2131.0	passDistanceToGoal_max >= 39.683144 AND finalDistanceToGoal_avg >= 39.83972
8	2	2104	0.21166157722473145	0.9615019011406845	2023.0	finalDistanceToGoal_avg >= 43.03657 AND popDistToPosRank_min <= 1.0
9	2	2370	0.21164372563362122	0.9139240506329114	2166.0	finalDistanceToGoal_avg >= 29.43188 AND finalDistanceToGoal_avg >= 39.090866
10	2	2101	0.21156981587409973	0.9619228938600667	2021.0	finalDistanceToGoal_avg >= 43.03657 AND distanceToOpponentGoalRating_max >= 10.0

Table B.2: The first 10 subgroups found by Cortana for target 1 refinement depth 2

Nr.	Depth	Coverage	Quality	Probability	Positives	Conditions
1	1	2106	0.1698686182498932	0.7060778727445394	1487.0	finalDistanceToGoal_avg <= 43.03657
2	1	1580	0.1477213054895401	0.7601265822784811	1201.0	finalDistanceToGoal_avg <= 35.917004
3	1	2632	0.13715964555740356	0.5858662613981763	1542.0	finalDistanceToGoal_avg <= 51.162743
4	1	2024	0.1287968009710312	0.6343873517786561	1284.0	lastPassPopDistanceToGoalRank_avg <= 3.0
5	1	2106	0.12783583998680115	0.6220322886989553	1310.0	lastPassDistanceToGoal_avg <= 48.50716
6	1	1580	0.12088681757450104	0.6886075949367089	1088.0	lastPassDistanceToGoal_avg <= 42.39138
7	1	2106	0.11928680539131165	0.6049382716049383	1274.0	passDistanceToGoal_min <= 45.41478
8	1	2429	0.11920273303985596	0.5730753396459448	1392.0	lastPassPopDistanceToGoalRank_avg <= 4.0
9	1	1534	0.1184777319431305	0.6916558018252934	1061.0	lastPassPopDistanceToGoalRank_avg <= 2.0
10	1	2285	0.11748455464839935	0.5829321663019693	1332.0	passPopDistanceToGoalRank_min <= 3.0

Table B.3: The first 10 subgroups found by Cortana for target 2 refinement depth 1

Nr.	Depth	Coverage	Quality	Probability	Positives	Conditions
1	1	2106	,0.1698686182498932	,0.7060778727445394	1487	,finalDistanceToGoal_avg,<= 43.03657
2	2	2094	,0.1697254478931427	,0.7077363896848138	1482	,finalDistanceToGoal_avg,<= 43.03657 AND eventOverlap = 'o'
3	2	2104	,0.16956771910190582	,0.7057984790874525	1485	,finalDistanceToGoal_avg,<= 43.03657 AND lastPassDistToPosRank_avg <= 9.0
4	2	2092	,0.16942451894283295	,0.7074569789674953	1480	finalDistanceToGoal_avg <= 43.03657 AND distanceToOpponentRank_min <= 1.0
5	2	2081	,0.16919434070587158	,0.7087938491110043	1475	finalDistanceToGoal_avg <= 43.03657 AND distanceToOpponentGoalRating_min <= 1.0
6	2	1975	,0.1691564917564392	,0.7270886075949367	1436	finalDistanceToGoal_avg <= 51.162743 AND finalDistanceToGoal_avg <= 41.013836
7	2	1975	,0.1691564917564392	,0.7270886075949367	1436	,finalDistanceToGoal_avg,<= 60.2957 AND finalDistanceToGoal_avg <= 41.013836
8	2	2093	,0.16910004615783691	,0.7270886075949367	1479	,finalDistanceToGoal_avg,<= 43.03657 AND distanceToOpponentGoalRating_max >= 10.0
9	2	2083	,0.16759544610977173	,0.7066411849020545	1469	,finalDistanceToGoal_avg,<= 43.03657 AND passPopDistToOpRank_min >= 1.0
10	2	2069	0.16667641699314117	,0.7052328372539607	1460	,finalDistanceToGoal_avg,<= 43.03657 AND curPositioningRank_min <= 1.0

Table B.4: The first 10 subgroups found by Cortana for target 2 refinement depth 2

Nr.	Depth	Coverage	Quality	Probability	Positives	Conditions
1	1	1053	0.03128387778997421	0.1794871794871795	189.0	finalDistanceToGoal_avg <= 29.43188
2	1	1580	0.028277704492211342	0.12974683544303797	205.0	finalDistanceToGoal_avg <= 35.917004
3	1	1580	0.024240659549832344	0.1189873417721519	188.0	lastPassDistanceToGoal_avg <= 42.39138
4	1	527	0.02359083667397499	0.2428842504743833	128.0	finalDistanceToGoal_avg <= 22.04955
5	1	1534	0.02340986765921116	0.11864406779661017	182.0	lastPassPopDistanceToGoalRank_avg <= 2.0
6	1	2106	0.023384662345051765	0.10113960113960115	213.0	finalDistanceToGoal_avg <= 43.03657
7	1	1580	0.022340873256325722	0.11392405063291139	180.0	distanceToOpponentGoal_min <= 14.965799
8	1	1053	0.022259894758462906	0.14339981006647673	151.0	lastPassDistanceToGoal_avg <= 36.738216
9	1	1754	0.022231074050068855	0.10775370581527936	189.0	passPopDistanceToGoalRank_min <= 2.0
10	1	2024	0.022068887948989868	0.10029644268774704	203.0	lastPassPopDistanceToGoalRank_avg <= 3.0

Table B.5: The first 10 subgroups found by Cortana for target 3 refinement depth 1

Nr.	Depth	Coverage	Quality	Probability	Positives	Conditions
1	2	853	0.03172944113612175	0.21101992966002345	180.0	lastPassPopDistanceToGoalRank_avg <= 7.0 AND finalDistanceToGoal_avg <= 27.176487
2	2	988	0.03164834901690483	0.18927125506072875	187.0	distanceToOpponentGoal_min <= 21.272202 AND finalDistanceToGoal_avg <= 29.546146
3	2	922	0.03155078738927841	0.1984815618221258	183.0	finalDistanceToGoal_avg <= 29.43188 AND distanceToOwnGoal_max >= 89.667114
4	2	835	0.03148694708943367	0.21317365269461078	178.0	passPopDistanceToGoalRank_min <= 6.0 AND finalDistanceToGoal_avg <= 26.957848
5	2	988	0.03141087293624878	0.1882591093117409	186.0	distanceToOwnGoal_max >= 87.40847 AND finalDistanceToGoal_avg <= 29.36972
6	2	1047	0.03136136010289192	0.18051575931232092	189.0	finalDistanceToGoal_avg <= 29.43188 AND distanceToOpponentGoalRating_min <= 1.0
7	2	1052	0.03129678964614868	0.1796577946768061	189.0	finalDistanceToGoal_avg <= 29.43188 AND lastPassDistToPosRank_avg <= 9.0
8	1	1053	0.03128387778997421	0.1794871794871795	189.0	finalDistanceToGoal_avg <= 29.43188
9	2	1054	0.031270962208509445	0.1793168880455408	189.0	finalDistanceToGoal_avg <= 43.03657 AND finalDistanceToGoal_avg <= 29.44404
10	2	790	0.03111819177865982	0.22025316455696203	174.0	distanceToOwnGoal_min <= 6.3409877 AND finalDistanceToGoal_avg <= 26.621393

Table B.6: The first 10 subgroups found by Cortana for target 3 refinement depth 2

Nr.	Depth	Coverage	Quality	Probability	Positives	Conditions
1	1	1053	0.019055213779211044	0.11016144349477683	116.0	finalDistanceToGoal_avg <= 29.43188
2	1	1580	0.017655018717050552	0.0810126582278481	128.0	finalDistanceToGoal_avg <= 35.917004
3	1	527	0.015222935006022453	0.1555977229601518	82.0	finalDistanceToGoal_avg <= 22.04955
4	1	2106	0.015075521543622017	0.0641025641025641	135.0	finalDistanceToGoal_avg <= 43.03657
5	1	1534	0.01303903665393591	0.0697522816166884	107.0	lastPassPopDistanceToGoalRank_avg <= 2.0
6	1	1754	0.01292720902711153	0.06499429874572406	114.0	passPopDistanceToGoalRank_min <= 2.0
7	1	1580	0.012905553914606571	0.06835443037974684	108.0	lastPassDistanceToGoal_avg <= 42.39138
8	1	2024	0.012887110933661461	0.060770750988142296	123.0	lastPassPopDistanceToGoalRank_avg <= 3.0
9	1	2106	0.01270078681409359	0.059354226020892686	125.0	passDistanceToGoal_min <= 45.41478
10	1	2106	0.012463314458727837	0.05887939221272555	124.0	lastPassDistanceToGoal_avg <= 48.50716

Table B.7: The first 10 subgroups found by Cortana for target 4 refinement depth 1

Nr.	Depth	Coverage	Quality	Probability	Positives	Conditions
1	2	922	0.01916174218058586	0.12147505422993492	112.0	finalDistanceToGoal_avg <= 29.43188 AND passDistToPos_avg >= 8.945137
2	2	1048	0.019095536321401596	0.11068702290076336	116.0	finalDistanceToGoal_avg <= 29.43188 AND distanceToOpponentRank_min <= 1.0
3	2	1020	0.019083861261606216	0.11274509803921569	115.0	finalDistanceToGoal_avg <= 29.43188 AND passDistToPosRank_max <= 9.0
4	2	1052	0.019063279032707214	0.11026615969581749	116.0	finalDistanceToGoal_avg <= 29.43188 AND lastPassDistToPosRank_avg <= 9.0
5	2	964	0.019060514867305756	0.11721991701244813	113.0	finalDistanceToGoal_avg <= 29.43188 AND passPopDistToOpRank_avg <= 8.0
6	1	1053	0.019055213779211044	0.11016144349477683	116.0	finalDistanceToGoal_avg <= 29.43188
7	2	1054	0.019047152251005173	0.11005692599620494	116.0	finalDistanceToGoal_avg <= 43.03657 AND finalDistanceToGoal_avg <= 29.44404
8	2	970	0.019012130796909332	0.11649484536082474	113.0	finalDistanceToGoal_avg <= 29.43188 AND lastPassPopAngleRank_avg >= 2.0
9	2	1031	0.01899515464901924	0.11154219204655674	115.0	finalDistanceToGoal_avg <= 29.43188 AND curPositioningRank_min <= 1.0
10	2	1185	0.018940623849630356	0.10126582278481013	120.0	finalDistanceToGoal_avg <= 60.2957 AND finalDistanceToGoal_avg <= 31.083696

Table B.8: The first 10 subgroups found by Cortana for target 4 refinement depth 2

Nr.	Depth	Coverage	Quality	Probability	Positives	Conditions
1	1	1053	0.004808397032320499	0.026590693257359924	28.0	finalDistanceToGoal_avg <= 29.43188
2	1	527	0.004540584050118923	0.04364326375711575	23.0	finalDistanceToGoal_avg <= 22.04955
3	1	1580	0.004362042061984539	0.0189873417721519	30.0	finalDistanceToGoal_avg <= 35.917004
4	1	1053	0.0038585038855671883	0.022792022792022793	24.0	distanceToOpponentGoal_min <= 12.319344
5	1	527	0.00382816419005394	0.03795066413662239	20.0	distanceToOpponentGoal_min <= 9.308121
6	1	2106	0.003679961897432804	0.014719848053181387	31.0	finalDistanceToGoal_avg <= 43.03657
7	1	1580	0.0036496222019195557	0.01708860759493671	27.0	distanceToOpponentGoal_min <= 14.965799
8	1	2106	0.0034424886107444763	0.014245014245014245	30.0	lastPassDistanceToGoal_avg <= 48.50716
9	1	1580	0.003412148915231228	0.016455696202531647	26.0	passDistanceToGoal_min <= 40.077274
10	1	2024	0.003348367754369974	0.01432806324110672	29.0	lastPassPopDistanceToGoalRank_avg <= 3.0

Table B.9: The first 10 subgroups found by Cortana for target 5 refinement depth 1

Nr.	Depth	Coverage	Quality	Probability	Positives	Conditions
1	2	659	0.005259714554995298	0.0409711684370258	27.0	lastPassDistanceToGoal_avg <= 55.82319 AND finalDistanceToGoal_avg <= 24.63637
2	2	693	0.005200275685638189	0.03896103896103896	27.0	lastPassPopDistanceToGoalRank_avg <= 5.0 AND finalDistanceToGoal_avg <= 25.060026
3	2	922	0.005037411116063595	0.03036876355748373	28.0	finalDistanceToGoal_avg <= 29.43188 AND dtPrevEvent <= 62.6
4	2	922	0.005037411116063595	0.03036876355748373	28.0	finalDistanceToGoal_avg <= 29.43188 AND passDistToPos_avg <= 21.607086
5	2	922	0.005037411116063595	0.03036876355748373	28.0	finalDistanceToGoal_avg <= 29.43188 AND passPopRank_avg >= 14.4
6	2	922	0.005037411116063595	0.03036876355748373	28.0	finalDistanceToGoal_avg <= 29.43188 AND popDistToPosRank_avg >= 2.7789326
7	2	922	0.005037411116063595	0.03036876355748373	28.0	finalDistanceToGoal_avg <= 29.43188 AND angleOpponentToPassline_max >= 47.91058
8	2	922	0.005037411116063595	0.03036876355748373	28.0	finalDistanceToGoal_avg <= 29.43188 AND dist to closest home_avg <= 10.345883
9	2	922	0.005037411116063595	0.03036876355748373	28.0	finalDistanceToGoal_avg <= 29.43188 AND dist to closest home_max >= 31.337
10	2	922	0.005037411116063595	0.03036876355748373	28.0	finalDistanceToGoal_avg <= 29.43188 AND dist to closest visitor_min <= 0.592

Table B.10: The first 10 subgroups found by Cortana for target 5 refinement depth 2

Appendix C

Definitions features

Feature / extension	Description
(feature)_avg	Average value a player received for the feature during a build-up
(feature)_max	Maximum value a player received for the feature during a build-up
(feature)_min	Minimum value a player received for the feature during a build-up
angleOpponentToPassline	Angle from opponent to the passline
distanceToBall	Distance from a player to the ball
distanceToOpponentGoal	Distance to opponent′s goal′
distanceToOwnGoal	Distance to own goal′
distanceToOpponentRank	Rank(1-9) for the distance a player has to his opponent
angleToPasslineRank	Rank(1-9) for the angle an opponent has to the passline
distanceToPossessionRank	Rank(1-9) for the distance a player has to his teammate in possession
distanceToOpponentGoalRank	Rank(1-9) for distance that a player is closter to the goal to than the player in possession
curPositioningRank	Rank(1-9) for the positioning of a player at a certain timestamp
distanceToOpponentRating	Rating (1-10) for the distance a player has to his opponent
angleToPasslineRating	Rating (1-10) for the angle an opponent has to the passline
distanceToPossessionRating	Rating (1-10) for the distance a player has to his teammate in possession
distanceToOpponentGoalRating	Rating (1-10) for distance that a player is closter to the goal to than the player in possession′
curPositioningRating	Rating (1-10) for the positioning of a player at a certain timestamp
ballPassedTo	Binary value (0/1) that indicates whether the player is passed the ball to′
passRating	The positioning rating of the player who received the pass
passDistanceToGoalRating	The distance rating of the player who received the pass
passAngleRating	The pass angle rating of the player who received the pass
passDistToOpRating	The distance to opponent rating of the player who received the pass
passDistToPosRating	The distance to the player in possession rating of the player who received the pass
passDistanceToGoal	The distance rating of the player who received the pass
passAngle	The pass angle rating of the player who received the pass
passDistToOp	The distance to opponent rating of the player who received the pass
passDistToPos	The distance to the player in possession rating of the player who received the pass
passRank	The positioning ranking of the player who received the pass′
passDistanceToGoalRank	The distance ranking of the player who received the pass′
passAngleRank	The pass angle ranking of the player who received the pass′
passDistToOpRank	The distance to opponent ranking of the player who received the pass′
passDistToPosRank	The distance to the player in possession ranking of the player who received the pass′
passPopRank	The population based positioning ranking of the player who received the pass′
passPopDistanceToGoalRank	The population based distance ranking of the player who received the pass′
passPopAngleRank	The population based pass angle ranking of the player who received the pass′
passPopDistToOpRank	The population based distance to opponent ranking of the player who received the pass′
passPopDistToPosRank	The population based distance to the player in possession ranking of the player who received the pass′
passDistanceToGoal	The distance ranking of the player who received the pass′
passAngle	The pass angle ranking of the player who received the pass′
passDistToOp	The distance to opponent ranking of the player who received the pass′
passDistToPos	The distance to the player in possession ranking of the player who received the pass′
positioningRating	Rating (1-10) for the positioning of a player over the course of a match
lastPassRating	The positioning rating of the player who received the pass
lastPassDistanceToGoalRating	The distance rating of the player who received the pass
lastPassAngleRating	The pass angle rating of the player who received the pass
lastPassDistToOpRating	The distance rating to opponent rating of the player who received the pass
lastPassDistToPosRating	The distance rating to the player in possession rating of the player who received the pass

Feature / extension	**Description**
lastPassDistanceToGoal	The distance rating of the player who received the pass
lastPassAngle	The pass angle rating of the player who received the pass
lastPassDistToOp	The distance to opponent rating of the player who received the pass
lastPassDistToPos	The distance to the player in possession rating of the player who received the pass
finalDistanceToGoal	The player's distance to the goal at the end of the builUp'
lastPassRank	The positioning rank of the player who received the pass'
lastPassDistanceToGoalRank	The distance rank of the player who received the pass'
lastPassAngleRank	The pass angle rank of the player who received the pass'
lastPassDistToOpRank	The distance rank to opponent rating of the player who received the pass'
lastPassDistToPosRank	The distance rank to the player in possession rating of the player who received the pass'
lastPassPopRank	The distance rank of the player who received the pass'
lastPassPopDistanceToGoalRank	The population-based pass angle rank of the player who received the pass'
lastPassPopAngleRank	The population-based distance to opponent rank of the player who received the pass'
lastPassPopDistToOpRank	The population-based distance to the player in possession rank of the player who received the pass'
lastPassPopDistToPosRank	The population-based player's distance to the goal rank at the end of the builUp'
popDistanceToGoalRank	The population based rating for the angle from an opponent to the passline
popAngleRank	The population based ranking for the distance to the goal of a player
popDistToOpRank	The population based ranking for a player to the player in possession
popDistToPosRank	The population based ranking for the disting from a player to his opponent
popPositioningRank	The population based ranking for the overall positioning of a player

Appendix D

Histograms ranking methods

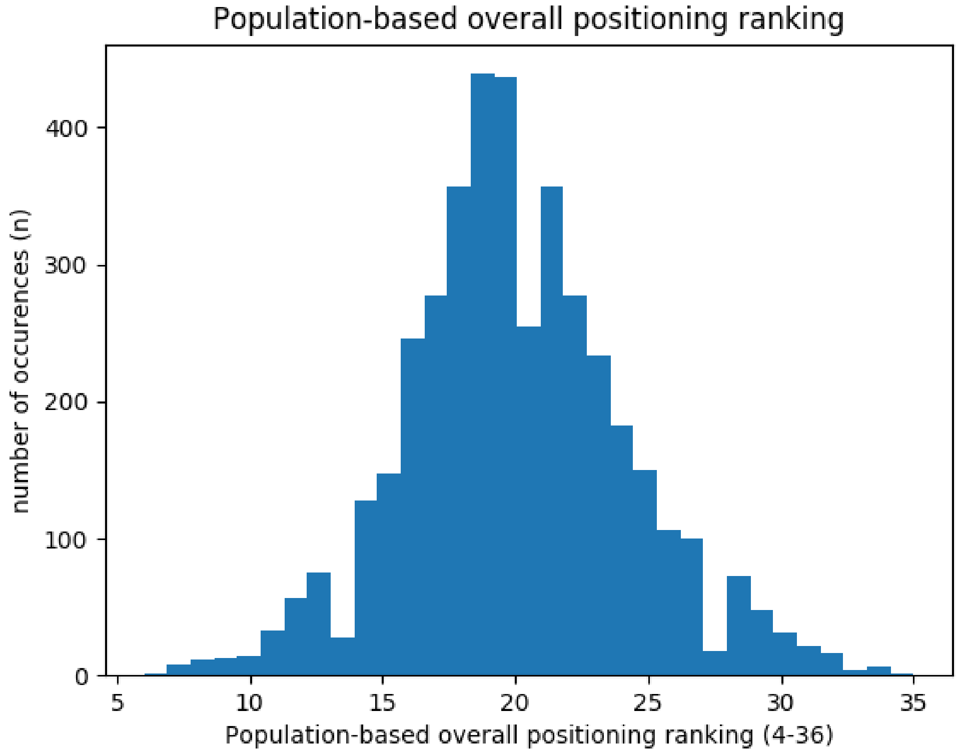


Figure D.1: Histogram population-based overall positioning ranking

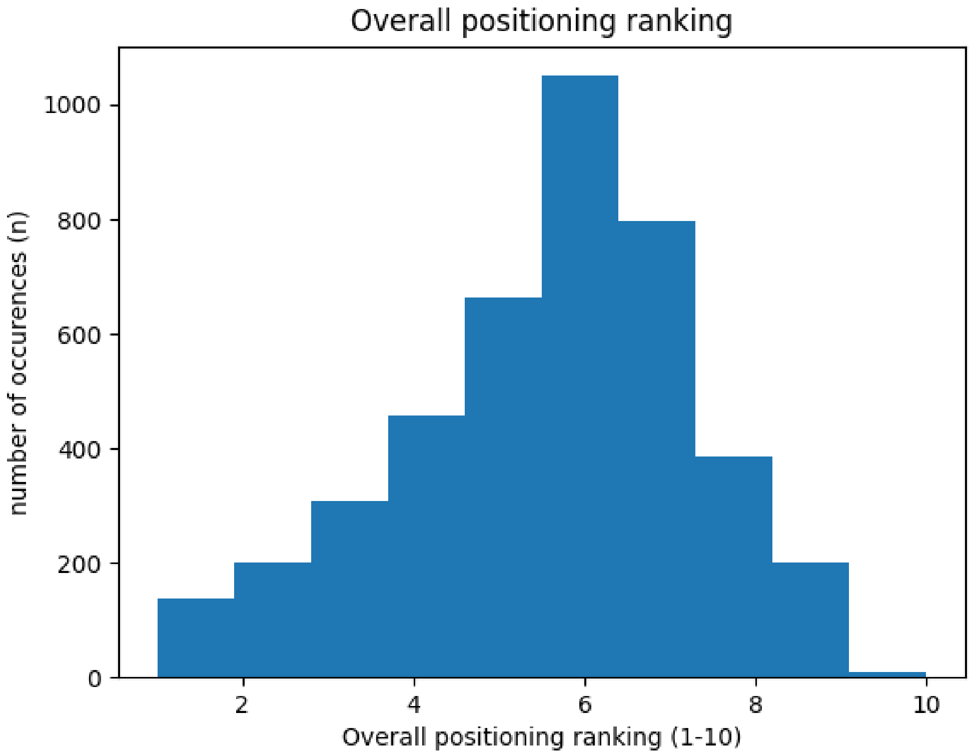


Figure D.2: Histogram overall positioning ranking

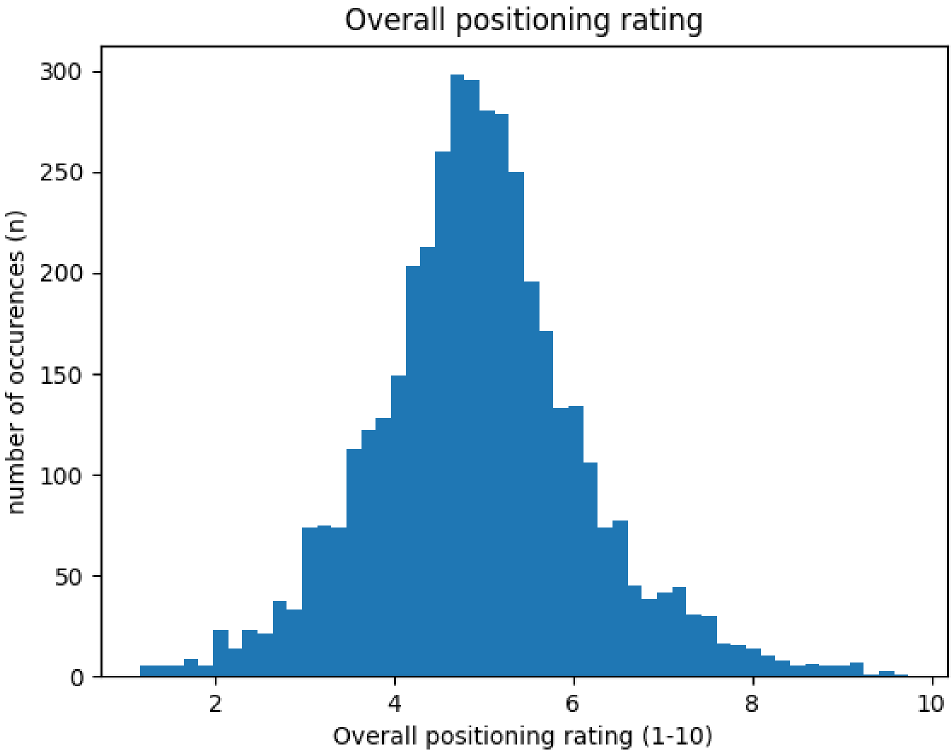


Figure D.3: Histogram overall positioning rating

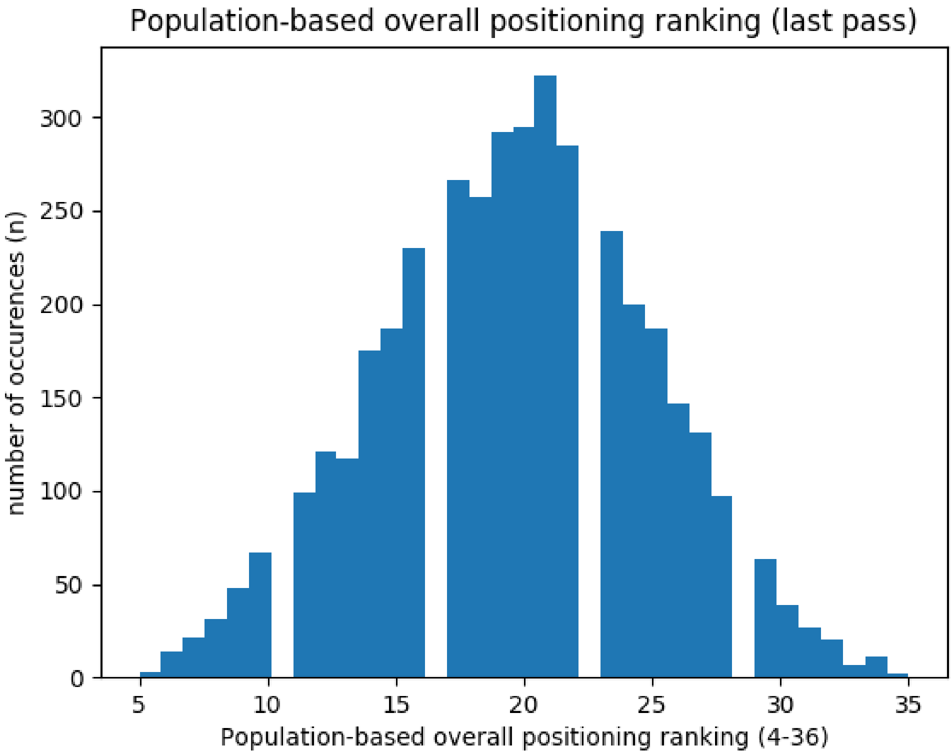


Figure D.4: Histogram population-based overall positioning ranking (last pass)

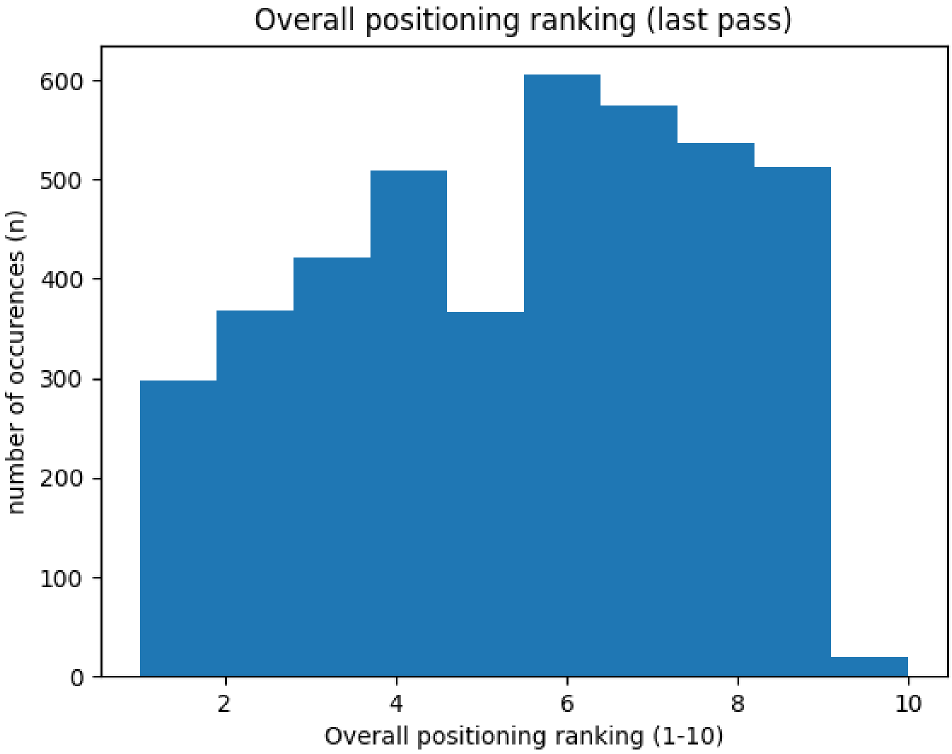


Figure D.5: Histogram overall positioning ranking (last pass)

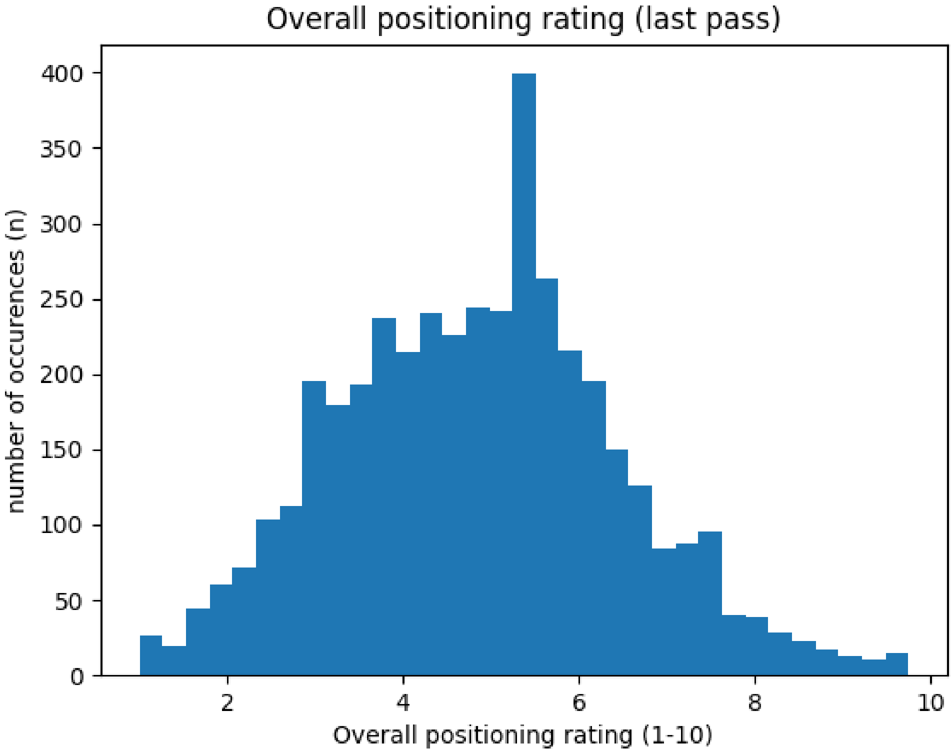
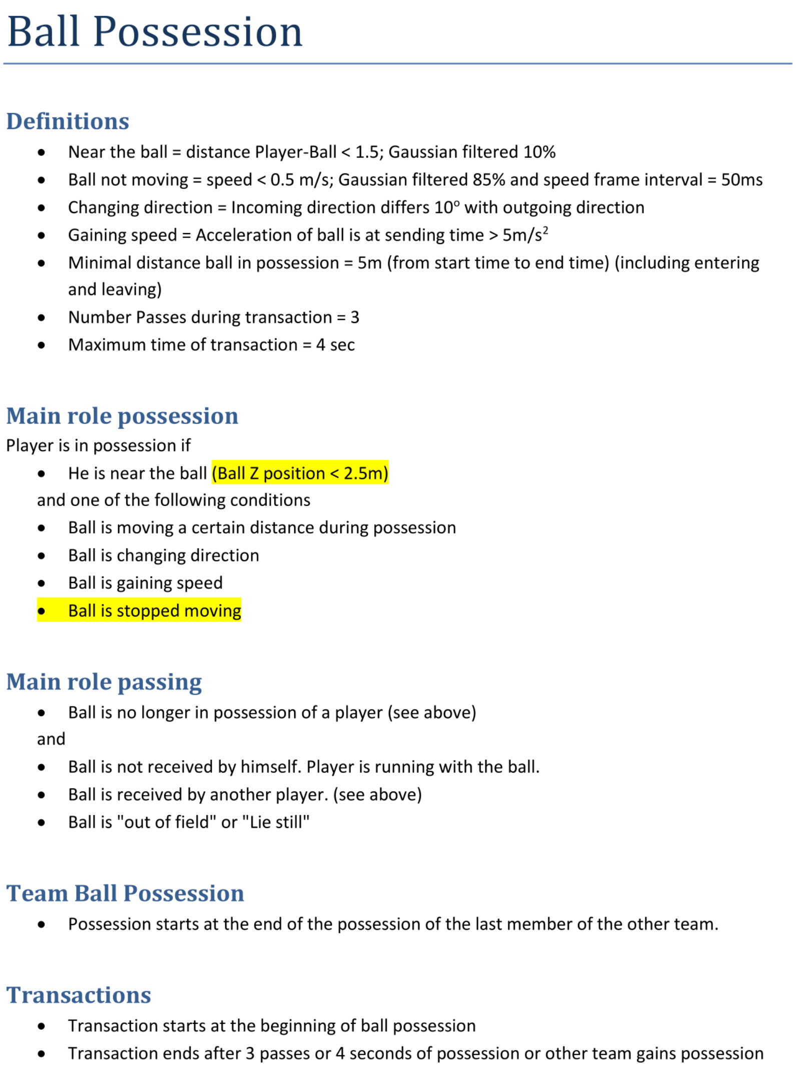


Figure D.6: Histogram overall positioning rating (last pass)

Appendix E

Definition of ball possession in Inmotio



Ball Possession

Definitions

- Near the ball = distance Player-Ball < 1.5; Gaussian filtered 10%
- Ball not moving = speed < 0.5 m/s; Gaussian filtered 85% and speed frame interval = 50ms
- Changing direction = Incoming direction differs 10° with outgoing direction
- Gaining speed = Acceleration of ball is at sending time > 5m/s^2
- Minimal distance ball in possession = 5m (from start time to end time) (including entering and leaving)
- Number Passes during transaction = 3
- Maximum time of transaction = 4 sec

Main role possession

Player is in possession if

- He is near the ball (Ball Z position < 2.5m)

and one of the following conditions

- Ball is moving a certain distance during possession
- Ball is changing direction
- Ball is gaining speed
- Ball is stopped moving

Main role passing

- Ball is no longer in possession of a player (see above)

and

- Ball is not received by himself. Player is running with the ball.
- Ball is received by another player. (see above)
- Ball is "out of field" or "Lie still"

Team Ball Possession

- Possession starts at the end of the possession of the last member of the other team.

Transactions

- Transaction starts at the beginning of ball possession
- Transaction ends after 3 passes or 4 seconds of possession or other team gains possession