MASTER

Football data analysis
an elaborate analysis on what possession loss causes and what causes possession loss

Hendriks, K.

Award date:
2016

Disclaimer
This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
Football data analysis

An elaborate analysis on what possession loss causes and what causes possession loss

Kees Hendriks

Supervisors:
- dr. N. (Natalia) Sidorova
- R. (Ruud) van Elk MSc.
- dr. C.M. (Claudia) Chituc
- Prof. H. (Heikki) Saikkonen

version 1.0

Public
Eindhoven, 31st August, 2016
Abstract

In football (soccer) it is important to analyse how a match is played, in order to keep improving the performance. Today the analysis is predominantly done manually. However, techniques to keep track of players during the match are emerging. These techniques provide a lot of data, which in turn can be analysed and then help the coach get more insights of the performance of the players.
This thesis explores how the data already collected by the club can be analysed to see what causes and happens after possession loss.
First the data is pre-processed in order to extract only the useful information. Next the critical moments are identified by looking at the features of the data. Then the data is analysed using statistical methods in order to clarify what happens to the play after possession loss and how the players react. Furthermore, multiple classification techniques and local process models are used to analyse the game before possession loss.
The automatic extraction of critical moments selects almost the same moments as a specialist. When the possession of the ball is lost the pressure put on the opponent does increase, the played on area decreases, and the players move forward. The amount of pressure put on the opponent changes due to the type of moments and the opponent. Classifiers can be used to predict when the possession of the ball changes. However, the classifiers in this thesis are not much better than random.
In conclusion, there are a lot of improvements which can be made in the analysis of football data. It is possible to extract critical moments and analyse what events occurs more or less frequent in certain circumstances. The moment possession of the ball changes is hard to identify.
# Contents

Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contents</td>
<td>v</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>ix</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2 Related Work</td>
<td>3</td>
</tr>
<tr>
<td>3 Preliminaries</td>
<td>5</td>
</tr>
<tr>
<td>4 Data</td>
<td>9</td>
</tr>
<tr>
<td>4.1 Original Data</td>
<td>9</td>
</tr>
<tr>
<td>4.2 Data Cleaning</td>
<td>12</td>
</tr>
<tr>
<td>4.3 Conclusions</td>
<td>14</td>
</tr>
<tr>
<td>5 Concepts, Models and Methods</td>
<td>15</td>
</tr>
<tr>
<td>5.1 Identification of Critical Moments</td>
<td>15</td>
</tr>
<tr>
<td>5.2 Statistical Analysis of the Critical Moments</td>
<td>17</td>
</tr>
<tr>
<td>5.3 Data aggregation</td>
<td>19</td>
</tr>
<tr>
<td>5.4 Statistical Analysis during the match</td>
<td>19</td>
</tr>
<tr>
<td>5.5 Classification</td>
<td>20</td>
</tr>
<tr>
<td>5.6 Local Process Models</td>
<td>21</td>
</tr>
<tr>
<td>5.7 Visualization</td>
<td>21</td>
</tr>
<tr>
<td>6 Results and evaluation</td>
<td>25</td>
</tr>
<tr>
<td>6.1 Identification of Critical Moments</td>
<td>25</td>
</tr>
<tr>
<td>6.2 Statistical Analysis of the Critical Moments</td>
<td>26</td>
</tr>
<tr>
<td>6.3 Statistical Analysis During the Match</td>
<td>39</td>
</tr>
<tr>
<td>6.4 Classification</td>
<td>41</td>
</tr>
<tr>
<td>6.5 Local Process Model</td>
<td>44</td>
</tr>
<tr>
<td>6.6 Striking Features and Hypotheses</td>
<td>47</td>
</tr>
<tr>
<td>6.7 Conclusions</td>
<td>50</td>
</tr>
<tr>
<td>7 Discussion</td>
<td>53</td>
</tr>
<tr>
<td>7.1 Future Improvements</td>
<td>53</td>
</tr>
<tr>
<td>7.2 Use by the club</td>
<td>54</td>
</tr>
<tr>
<td>Bibliography</td>
<td>55</td>
</tr>
<tr>
<td>Appendix</td>
<td>57</td>
</tr>
<tr>
<td>A Area Covered During All Matches</td>
<td>57</td>
</tr>
</tbody>
</table>

Football data analysis v
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Sample timeline</td>
<td>2</td>
</tr>
<tr>
<td>3.1</td>
<td>Sample decision tree</td>
<td>5</td>
</tr>
<tr>
<td>3.2</td>
<td>Sample rule induction</td>
<td>6</td>
</tr>
<tr>
<td>3.3</td>
<td>Sample local process models</td>
<td>6</td>
</tr>
<tr>
<td>4.1</td>
<td>The dimensions of the field in the tactical data</td>
<td>11</td>
</tr>
<tr>
<td>4.2</td>
<td>The dimensions of the field in the tactical data</td>
<td>11</td>
</tr>
<tr>
<td>5.1</td>
<td>Analysis process</td>
<td>15</td>
</tr>
<tr>
<td>5.2</td>
<td>A critical moment</td>
<td>16</td>
</tr>
<tr>
<td>5.3</td>
<td>Critical flowchart</td>
<td>17</td>
</tr>
<tr>
<td>5.4</td>
<td>Defenders</td>
<td>17</td>
</tr>
<tr>
<td>5.5</td>
<td>Convex Hull</td>
<td>18</td>
</tr>
<tr>
<td>5.6</td>
<td>Hull centroid</td>
<td>19</td>
</tr>
<tr>
<td>5.7</td>
<td>Player centroid</td>
<td>19</td>
</tr>
<tr>
<td>5.8</td>
<td>Used classes</td>
<td>20</td>
</tr>
<tr>
<td>5.9</td>
<td>moment viewer</td>
<td>22</td>
</tr>
<tr>
<td>5.10</td>
<td>coach report</td>
<td>24</td>
</tr>
<tr>
<td>6.1</td>
<td>Area covered after possession loss</td>
<td>27</td>
</tr>
<tr>
<td>6.2</td>
<td>Differentiated area</td>
<td>27</td>
</tr>
<tr>
<td>6.3</td>
<td>Hull centroid movement in time</td>
<td>27</td>
</tr>
<tr>
<td>6.4</td>
<td>Player centroid movement in time</td>
<td>27</td>
</tr>
<tr>
<td>6.5</td>
<td>Polygons when possession is lost</td>
<td>28</td>
</tr>
<tr>
<td>6.6</td>
<td>Polygons 5s after possession loss</td>
<td>28</td>
</tr>
<tr>
<td>6.7</td>
<td>The centroids of the polygons when possession is lost</td>
<td>29</td>
</tr>
<tr>
<td>6.8</td>
<td>well-handled moments, sample size</td>
<td>29</td>
</tr>
<tr>
<td>6.9</td>
<td>well-handled moments, absolute area</td>
<td>30</td>
</tr>
<tr>
<td>6.10</td>
<td>well-handled moments, differentiated area</td>
<td>30</td>
</tr>
<tr>
<td>6.11</td>
<td>well-handled moments, convex hull centroid change</td>
<td>31</td>
</tr>
<tr>
<td>6.12</td>
<td>well-handled moments, player centroid change</td>
<td>32</td>
</tr>
<tr>
<td>6.13</td>
<td>well-handled moments, start and end convex hulls heat map</td>
<td>33</td>
</tr>
<tr>
<td>6.14</td>
<td>poorly handled moments, sample size</td>
<td>33</td>
</tr>
<tr>
<td>6.15</td>
<td>poorly handled moments, absolute area</td>
<td>34</td>
</tr>
<tr>
<td>6.16</td>
<td>poorly handled moments, differentiated area</td>
<td>34</td>
</tr>
<tr>
<td>6.17</td>
<td>poorly handled moments, convex hull centroid change</td>
<td>35</td>
</tr>
<tr>
<td>6.18</td>
<td>poorly handled moments, player centroid change</td>
<td>36</td>
</tr>
<tr>
<td>6.19</td>
<td>poorly handled moments, start and end convex hulls heat map</td>
<td>36</td>
</tr>
<tr>
<td>6.20</td>
<td>Absolute area for different opponents</td>
<td>37</td>
</tr>
<tr>
<td>6.21</td>
<td>Differentiated area for different opponents</td>
<td>38</td>
</tr>
<tr>
<td>6.22</td>
<td>Convex hull centroid movement for different opponents</td>
<td>39</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.23</td>
<td>Area during the match</td>
<td>40</td>
</tr>
<tr>
<td>6.24</td>
<td>Area change during the match</td>
<td>40</td>
</tr>
<tr>
<td>6.25</td>
<td>Centroid during the match</td>
<td>41</td>
</tr>
<tr>
<td>6.26</td>
<td>Decision Tree</td>
<td>41</td>
</tr>
<tr>
<td>6.27</td>
<td>Rule induction</td>
<td>42</td>
</tr>
<tr>
<td>6.28</td>
<td>Petri net: pass followed by duel</td>
<td>44</td>
</tr>
<tr>
<td>6.29</td>
<td>Petri net: duel followed by pass</td>
<td>45</td>
</tr>
<tr>
<td>6.30</td>
<td>Petri net: pass and duel</td>
<td>45</td>
</tr>
<tr>
<td>6.31</td>
<td>Petri net: bring in followed by pass</td>
<td>45</td>
</tr>
<tr>
<td>6.32</td>
<td>Petri net: bring in followed by pass followed by duel</td>
<td>45</td>
</tr>
<tr>
<td>6.33</td>
<td>Petri net: opponent pass followed by opponent duel</td>
<td>45</td>
</tr>
<tr>
<td>6.34</td>
<td>Petri net: pass followed by duel filtered</td>
<td>46</td>
</tr>
<tr>
<td>6.35</td>
<td>Petri net: pass followed by duel, player B and player C</td>
<td>46</td>
</tr>
<tr>
<td>6.36</td>
<td>Petri net: pass followed by duel filtered</td>
<td>47</td>
</tr>
<tr>
<td>6.37</td>
<td>Biorhythm</td>
<td>48</td>
</tr>
<tr>
<td>6.38</td>
<td>Area and centroid of all matches</td>
<td>48</td>
</tr>
<tr>
<td>6.39</td>
<td>Centroid of both matches versus club C</td>
<td>49</td>
</tr>
<tr>
<td>6.40</td>
<td>hull centroid change when moments are handled well</td>
<td>50</td>
</tr>
<tr>
<td>A.1</td>
<td>All matches in one figure is not useful</td>
<td>57</td>
</tr>
</tbody>
</table>
List of Tables

4.1 Sample of the tactical data ............................................. 10
4.2 Sample of the positional data ......................................... 11
4.3 Sample of the combined data ......................................... 14

5.1 Added features for classifiers .......................................... 21

6.1 Critical moments identified, manually vs. automatically .............. 25
6.2 Contingency table, number of found moments .......................... 26
6.3 Moments divided in classes ............................................. 41
6.4 Rule induction: confusion matrix ...................................... 42
6.5 Rule performance ....................................................... 43
6.6 Rule performance during second half .................................. 43
6.7 First rule when clauses are removed ................................... 44
Chapter 1

Introduction

In order to improve the play of a professional football club during matches, the played matches are analysed. Currently, this analysis is done by hand. Since also in sports there is a desire to use more and more data, the club recently started gathering data about the players. Firstly, the location of the players in the field is tracked. This positional data is already used to calculate some basic metrics. E.g. the total distance covered by a player or the number of sprints of a player. Furthermore, there is also tactical data collected. This tactical data consists of key activities during the match. E.g. Player X passes the ball to player Y.

There is a lot more information hidden in the data than is currently revealed. The club would like to increase the amount of information retrieved from the data. Automatic analysis can extend the results of the analysis compared to the analysis currently done by hand. Firstly, the club wants to know if the team actually plays according to the given tactics. Especially questionable is the amount of pressure put on the opponents after possession loss. Pressure is here a combination of two factors: the covered area and the forward movement. Increasing the pressure means decreasing the area played on and moving to the goal of the opponent[4, 7]. This is the analysis of what possession loss causes.

Secondly, it is interesting to analyse what causes possession loss and see if there are recurring patterns before possession loss. A recurring pattern here could be player K passes to player X, who gives an assist to player M, who scores the goal. Only here player X is not the same player in all cases, however player K and M are the same players. Possibly there are similar patterns before possession loss. Indicating which player is most likely to cause possession loss.

The problems described before result in the following research questions:

1. Is the club putting pressure on the opponents after possession loss?

2. Is the pressure put on the opponent different after possession loss under certain circumstances?
   E.g. when the moments are handled well, or when the club plays versus different opponents.

3. Are there recurring patterns which can be extracted by local processes mining?

To answer the questions posed above, it is not possible to straightforward analyse the whole match. Since every match is very different and there are not enough similarities between the matches. For analysis this poses a problem. First the possession loss moments have to be identified. Figure 1.1 shows an example of a timeline, with some information about what happened during the match. On this timeline there are several interesting events. While it is not possible to analyse the whole match, it is possible to analyse what happens just before or after the events. These events in turn represent critical moments in the match.
Figure 1.1: An example of a timeline of a match, in this match the team got a red card and two yellow cards however the team also scored two goals. The crosses represent the moment the team loses possession.

When the critical moments are identified, the size of the covered area and the forward movement can be calculated for all snapshots within a certain timespan from the moment. E.g. calculate the covered area and the forward movement the first five seconds after each moment, every 100 ms. When the moments are filtered on certain circumstances before analysing, the differences between these circumstances can be analysed.

Instead of analysing a fixed amount of time before or after a moment, a fixed number of events before or after the moment can also be analysed. When analysing a fixed number of events, it becomes possible to find patterns of different events after each other. In order to find these patterns process mining can be used.

This thesis is organized as follows. In Chapter 2 a closer look is taken at what researchers in this field already know, and what the influence of this work is on the problems tackled in this thesis. Chapter 3 will elaborate on the basis this thesis is built upon. Chapter 4 discusses the already collected data and how this data is cleaned and pre-processed in this project. Chapter 5 explains the concepts, models and methods used in the project. Chapter 6 discusses and evaluates the results from the research proposed in Chapter 5. And finally, Chapter 7 will discuss the project as a whole, answer the research questions posed above and propose future work.
Chapter 2

Related Work

Currently there is already some research done in the football data analysis domain. The research of football analytics started with basic statistics for example the number of passes before a goal is counted. This research resulted in the conclusion that the most goals are scored out of a short passing sequence, indicating the ball should be played forward quickly even if this means losing possession of the ball and when close to the goal of the opponent, possibly, regain possession and attempt to score a goal. This results in the claim that shots with the highest probability to score a goal, are the shots with short passing sequences before the shot [9]. However, if the scored goals are normalized by the frequency of the possession length, longer possession and more passes result in a higher probability of scoring [6]. This starting point of the research is also a good indication that great care should be taken when working with statistics. Short passing sequences result in more goals, because there are a lot more short passing sequences, compared with longer passing sequences.

More innovation is done in other sports than football, mainly sports where a lot of points are scored every match. When more points are scored in a sport, it is easier to predict the result of the match, because the better team will more often actually win. In football however one lucky shot could result in winning the match by the team that actually played worse. In basketball for example there are more points scored, so a single lucky shot is less significant. For these sports models predicting what team should have won a match already exist [2]. In football it is harder to create these models because very few goals are scored each match. However there is work done on expected goals models [5]. These models predict which team should have won after analysing the match and all shots on goal. A big problem with these models however is the fact there is scored only a few times each match. When the model wrongly predicts the number of goals by only 1, this still could make a huge difference. If the model predicts a match to result in 2-1 and the match results in 3-0, this would still be considered good. However, when the same match would have resulted in 1-2, still both teams are off by only 1, but the other team did win the match. So the prediction is wrong. And the model would be considered bad. The models can be improved by adding more and more features to the model. E.g. the skill of the player could be considered. But in order to get the best model, a lot of attributes have to be added. Because the number of attributes keeps increasing, the model easily becomes overtrained. Since the available data is limited. This is why this thesis will focus on an event which will happen more, namely possession loss. In every match dozens of times possession of the ball is lost, and typically just a few goals are scored.

In [4], the authors analyse small-sided football games. These are training matches with less than eleven players on each side. This research uses two variables: the centre of a team; and the covered area of a team. The pressure is defined by the distance between the two centres. The shorter the distance between the two centres is, the higher the pressure is. Increasing the pressure results in an increase of the likelihood mistakes are made. During goal attempts the distance between the centroid of the attacking and the defending team was almost 0. In most of the time the area of the attacking team increases suddenly. In this thesis similar research is done, however not focussing...
CHAPTER 2. RELATED WORK

on the goal attempts but on possession loss.
Finally there is done a lot more research on the athlete’s body and movement of the athlete during football matches. This research, although very promising, is omitted in this thesis, since the research is in another field.
Chapter 3

Preliminaries

This chapter will introduce some techniques, which are used later in the thesis. This chapter will only give a brief introduction to these techniques and reference to sources where more elaborate explanations can be found.

Classification classifies an entry of the multi-feature input set in a class. A classifier uses values of the attributes of the input set to predict a class label. An extensively used technique for classification is a decision tree. A decision tree classifies are organized like a decision tree in which simple conditions on (usually single) attributes label the edge between an intermediate node and its children. Leaves are labeled by class label predictions [3].

<table>
<thead>
<tr>
<th>smoker</th>
<th>age</th>
<th>health risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>35</td>
<td>high</td>
</tr>
<tr>
<td>yes</td>
<td>48</td>
<td>medium</td>
</tr>
<tr>
<td>yes</td>
<td>63</td>
<td>high</td>
</tr>
<tr>
<td>no</td>
<td>51</td>
<td>medium</td>
</tr>
<tr>
<td>no</td>
<td>74</td>
<td>medium</td>
</tr>
<tr>
<td>no</td>
<td>25</td>
<td>low</td>
</tr>
</tbody>
</table>

Figure 3.1: A sample decision tree, an input set (a) and a decision tree (b) of the input

Figure 3.1 depicts an example of a decision tree. This decision tree predicts the health risks based on the data in the input set. Smoking results in a high health risk. Not smoking and an age below 50 results in a health risk. Not smoking and an age above 50 results in a medium health risk.

Closely related to decision tree solutions are rule-based solutions. A rule may be constructed by forming a conjunct of every test that occurs on a path between the root node and a leaf node of a tree. The collection of all such rules obtained by traversing every unique path from root node to leaf node is a corresponding rule-based solution for classification. However, these rule-based solutions can overcome a drawback of decision trees, namely the leaves in a decision tree are by definition mutually exclusive. It is however possible to create rules which are not mutually exclusive [1].

Figure 3.2 shows rules for the same data, used in the decision tree in figure 3.1. Using rule induction, all entries can be classified correctly, because the first and third rule are not mutually exclusive, but once an entry is classified, the other rules do not need to take it into account anymore.

Instead of classifying, it is also possible to min association rules. Given a set of entries, where
if smoker = yes and age = 48
then medium
if smoker = no and age > 45
then medium
if smoker = yes
then high
else low

Figure 3.2: Rules generated for the sample set of figure 3.1.

each entry is a set of attributes, an association rule is an expression $X \Rightarrow Y$, where $X$ and $Y$ are sets of attributes. The intuitive meaning of such a rule is that entries in the database which contain the attributes in $X$ tend to also contain the items in $Y$. The support is defined as the fraction of entries that contain both set $X$ and set $Y$. The confidence is defined as the fraction of the selected entries which are predicted correctly by this rule. And the lift is a measure of how good the prediction is measured against a random choice [10].

Petri net theory allow a system to be modeled by a Petri net. Analysis of a Petri net can reveal important information about the structure and dynamic behaviour of the system. A Petri net is composed of four parts: a set of places, a set of transactions, an input function and an output function. The input and output functions relate the transitions and places. The input function relates places to the transitions as input of the transition and the output function relates places to the transitions as output of the transition.

Furthermore, tokens can be assigned to a place. A transition can fire if all places in the input function of the transition contain a token. After the transition fired, a token is removed from all places in the input function and a token is added to all places in the output function. In order to start and end a Petri net a subset of the places are initial places, which contain tokens at the start, and another subset of the places are end places, which contain tokens at the end of the Petri net execution [8].

In this thesis places are graphically depicted by circles and transitions are graphically depicted by rectangles. Figure 3.3(b) shows an example of a Petri net.

Local process models are frequently occurring subprocess, not necessarily describing the process from start to end and they allow for choice, concurrency, loops, and sequence relations [11].

Local process models indicate a frequent pattern in the event sequences. The events however do not have to happen immediately after each other. The pattern does also not necessarily start at the start of the sequence or end at the end of sequence.

Event sequences
{A, A, C, B, A, A, C, B, B, C}
{C, A, A, B, A, A, B, C, B}
{A, D, A, B, D, C, D, A, B, C, B}
{C, A, A, B, B, B, A, D, B, C}
{B, A, B, C, C}
{D, A, A, C, B, C, A, B, C}

Figure 3.3: A sample local process model, an event sequence (a) and a local process model (b) of this event sequence
Figure 3.3 shows an example of a local process model. The identified sequence is first $A$, followed by $B$ and $C$. The order in which $B$ and $C$ occur is unimportant in this local process model. This pattern occurs 13 times in the event sequences. This is also indicated in the model by the 13. $A$ does occur 21 times in the event sequences. This is indicated in the model as well.
Chapter 4

Data

This chapter discusses the reliability of the data obtained from the club, and how this data is cleaned and pre-processed to be more useful.¹

4.1 Original Data

The first data source is the data collected by ORTEC, this data is known as tactical data. The second data source is positional data of all players on the field collected by the club. The tactical data is obtained by an individual tabbing on a tablet every time something happens. A delay is introduced due to the reaction speed of the person. The positional data is obtained by tracking every player visually on the field with three cameras. This data is less reliable when the player is on the far end of the field, and sometimes when the players get close together, the system loses which player is who because the system is only visual. This has to be corrected manually.

4.1.1 Tactical Data

The tactical data consists of one csv-file for each season. This file consists of all matches played in the league, even the matches in which the club did not participate. Each record represents an event in the match. This can be a lot of different events, for example a pass or a goal.

All entries have the following attributes:

- **Time** A timestamp in milliseconds, restarting at the start of the second half
- **Half** An integer indicating what half of the match it is. This will always be in \{1, 2\}.
- **Effectiveness** An integer indicating how effective the event was, a higher value indicates a more desired effect. The value will always be in \{1, 2, 3, 4, 5\}.
- **Category** A string indicating what kind of event this entry is.
- **Player** The player involved in this event. In duels there will be two records; one with an offensive player and one with a defensive player.
- **Team** The team the player in this event is playing for.
- **Attribute** A list containing some attributes of the event. E.g. Which foot is used
- **Definition** A list defining the event. E.g. If the pass is complete.

¹In the original documents some data is in Dutch and some in English, to be consistent I have translated everything to English.
CHAPTER 4. DATA

Match The match this event happened in.

Round An integer indicating which round of the competition this match is in.

Location X A float indicating the X position of the player, measured from 0 to 100, starting at the goal this team is defending.

Location Y A float indicating the Y position of the player, measured from 0 to 100, starting from the left of the goal this team is defending to the right of the goal.

Figure 4.1 depicts the field with the dimensions of the tactical data. A sample of the tactical data can be seen in Table 4.1. There are three things to note about this data. Firstly Attribute and Definition both can be empty. Secondly Player does contain accents on player names. Finally the location format is strange, more on this in section 4.2.

Table 4.1: A sample from the tactical data

<table>
<thead>
<tr>
<th>Time (ms)</th>
<th>Half</th>
<th>Effectiveness</th>
<th>Category</th>
<th>Player</th>
<th>Team</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>57784</td>
<td>1</td>
<td>3</td>
<td>pass</td>
<td>Player B</td>
<td>club O</td>
<td>1</td>
</tr>
<tr>
<td>192161</td>
<td>1</td>
<td>3</td>
<td>attacking action</td>
<td>Player C</td>
<td>club N</td>
<td>1</td>
</tr>
<tr>
<td>23667</td>
<td>1</td>
<td>3</td>
<td>attacking action</td>
<td>player A</td>
<td>club A</td>
<td>1</td>
</tr>
<tr>
<td>1136629</td>
<td>1</td>
<td>3</td>
<td>defending action</td>
<td>Player D</td>
<td>club R</td>
<td>1</td>
</tr>
<tr>
<td>1449000</td>
<td>2</td>
<td>3</td>
<td>foul</td>
<td>Player E</td>
<td>club N</td>
<td>1</td>
</tr>
</tbody>
</table>

... Attribute | ... Definition | ... |
... right foot | ... isPassCompleted | isPassForward | isPassLong | ... |
... head | duel touched | ... isDuelPart | isDuelWonByAttacker | isDuelAir | isAerial | ... |
... head | duel touched | ... isDuelPart | isDuelWonByAttacker | isDuelAir | isAerial | ... |
... blocked | ... isDuelPart | ... isDuelStanding | ... |

Data Reliability

The tactical data is obtained by a person registering every interaction with the ball. This will introduce some delay, due to the reaction time of this person. The locations are also obtained by a person. So this data will not be very reliable, since people make mistakes and the viewing angle of this person will not be optimal for all events.

4.1.2 Positional Data

The positional data consists of one csv-file for each match. In these files the ball, the referees and all players have a record every 100ms.

All records have the following attributes:

Timestamp A timestamp in milliseconds of the match, this timer does not reset, and even runs during half time.
CHAPTER 4. DATA

Figure 4.1: The dimensions of the field in the tactical data. All values are percentages.

Figure 4.2: the dimensions of the field used in this project. All values are in meters.

X The X location of this player. Where 0 is the middle of the field, and increases from left to right viewed from the camera.

Y The Y location of this player. Where 0 is the middle of the field, and increases from the bottom of the field to the top.

Marker Name An attribute indicating what part of the match it is. E.g. first 15 minutes

Speed The speed of this player, this will always a positive number.

Acceleration The acceleration of this player.

Name The name of the player, ‘ball’ or ‘referee’.

Figure 4.2 displays the dimensions of the field used in the positional data. A sample of the positional data can be seen in Table 4.2. There are three things to note about this data. Firstly of all there are no accents on the data. Even when the official name of the player contains accents.\(^2\) Secondly an unknown location is mapped as 99999.999. Finally the marker names changes between different matches.

Table 4.2: A sample from the positional data

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>X</th>
<th>Y</th>
<th>Marker Name</th>
<th>Speed</th>
<th>Acceleration</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>2311400</td>
<td>7.559</td>
<td>12.15</td>
<td>30</td>
<td>1.84</td>
<td>-0.66</td>
<td>Player F</td>
</tr>
<tr>
<td>5737100</td>
<td>-28.826</td>
<td>13.783</td>
<td>75</td>
<td>4.33</td>
<td>-4.49</td>
<td>Player G</td>
</tr>
<tr>
<td>23400</td>
<td>-14.919</td>
<td>28.968</td>
<td>Start1</td>
<td>1.05</td>
<td>1.08</td>
<td>Player A</td>
</tr>
<tr>
<td>1473100</td>
<td>8.788</td>
<td>0.81</td>
<td>15</td>
<td>2.67</td>
<td>-0.44</td>
<td>Player H</td>
</tr>
<tr>
<td>300</td>
<td>99999.999</td>
<td>99999.999</td>
<td>Start1</td>
<td>6.88</td>
<td>16.18</td>
<td>ball</td>
</tr>
</tbody>
</table>

Data Reliability

The positional data is obtained by three cameras at the mid-line of the field aiming at the left-side, the middle and the right-side of the field. This can be used to construct the whole field, however the far side of the field has a smaller viewing angle for the camera, so is less accurate. Also the

\(^2\)During the project when new matches were added, the accent usage changed and players did have accents.
CHAPTER 4. DATA

system which tracks the players, sometimes loses which player is who, this happens often when the players get close together. For example during a corner. This has to be corrected manually.

4.2 Data Cleaning

The needed data is separated over multiple inconsistent files, so cleaning of the data is needed.\(^3\) Data cleaning is needed in multiple ways. This section discusses how the data is cleaned.

4.2.1 Positions

First of all, the locations in the tactical data need to be cleaned. The data in the file is in one of the multiple formats and need to be cleaned in the following way. Here \( a \) to \( g \) all stand for a decimal number.

\[
\begin{align*}
  a.bcd.efg &\rightarrow ab.cdefg \\
  abc.defg &\rightarrow ab.cdefg \\
  ab.c &\rightarrow ab.c \\
  a.c &\rightarrow a.c \\
  100 &\rightarrow 100 \\
  0 &\rightarrow 0
\end{align*}
\]

Next, we can delete all the locations of the referees. These locations are not needed.

Now the positions themselves need to be cleaned. First the locations from the positional data and tactical data have to be put on the same scale. The scale used in the positional data is used throughout this thesis, because this scale is in meters, while the tactical data is in percentages of the field and later it will be more convenient when calculating the positions of other players, if these values are in meters. So a change of 1 is as far on both axes and it is easier to calculate the area.

So the positions will all be in meters with the origin of the axes in the middle of the field. As indicated in figure 4.2

For convenience during aggregation, it will be assumed all players are always defending the left goal. When in the original data the club is playing from right to left, the data needs to be swapped. Simply multiplying both \( x \) and \( y \) by \(-1\) will do.

To find whether the club is playing from left to right or from right to left, inequation 4.1 is checked. In this inequation \( c \) is a collection containing all the players with their locations in the first half. \( p \) is a collection of all the locations of a player and \( i \) is a location with \( i_x \) the X-coordinate of the location. The extreme values will always be the goal keeper and the most attacking player. However the goal keeper will have a higher absolute value. If inequation 4.1 is true, the club is playing from left to right in the first half. Otherwise the club is playing from right to left in the first half.

\[
|\min_{p\in c}(\sum_{i\in p} \frac{i_x}{|p|})| > \max_{p\in c}(\sum_{i\in p} \frac{i_x}{|p|}) \tag{4.1}
\]

So if the club is starting from left to right, the location of the opponent has to be swapped during the first half, and the location of the players has to be swapped during the second half. Otherwise both teams have to be swapped in the other half.

\(^3\)The cleaning of the data, I did together with Harm Eggels, because he needed to work with the same data from the club.
4.2.2 Combining Data

Some data of the positional data would be easy to have in the tactical data and vice versa. Because of the size of the location data it is undesirable to put all the data in one file. Creating one file for each match would also be undesirable since the tactical data would only be scarcely added to the location data. So only a few important attributes are added to both the positional and the tactical data.

When combining the tactical data with the positional data, both data sources need to be combined at the same moment. Because the timestamp of the tactical data can be any value and the timestamp of the positional data is always a multiple of 100, the timestamp of the tactical data needs to be rounded to the closest hundred.

Every entry in the tactical data is matched to an entry in the positional data by timestamp and player name, this results in only one record. As an example the third row of Table 4.1 and the third row of Table 4.2 can be matched.

The tactical data is extended with the following data from the positional data.

- **X** The $x$ location according to the positional data.
- **Y** The $y$ location according to the positional data.
- **Speed** The speed of the player.
- **Acceleration** The acceleration of the player.

The positional data is extended with the following data from the tactical data.

- **Team** The team this player plays on, this will be very convenient to have when the positions of all players of a team are needed.

The addition of the data to these files, will result in some duplicate data. However not having to calculate this every time will make the use of the data more convenient.

To every entry in the tactical data extra information is added based on the already existing data.

- **Possession** The team currently in possession of the ball. This is based on when the definition of the previous entry contains: $isPossessionLoss$.
- **Possession time** The time the current team is in possession of the ball in milliseconds.

After the combining of the data is done the data looks as in table 4.3. The third rows of table 4.2 and table 4.1 correspond with the third row in table 4.3. Note the timestamp did slightly change, this is because both data sources do not start the timers at precisely the same moment at the start of the match. Furthermore, note the coordinates are on the other side of 0. This is because this match the club started playing from right to left.
Table 4.3: A sample from the combined data

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>Round</th>
<th>Match</th>
<th>Team</th>
<th>Player</th>
<th>Category</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>4674700</td>
<td>34</td>
<td>club I - club A</td>
<td>club I</td>
<td>Player I</td>
<td>pass</td>
<td>3</td>
</tr>
<tr>
<td>2697800</td>
<td>12</td>
<td>club A - club Q</td>
<td>club A</td>
<td>Player J</td>
<td>pass</td>
<td>3</td>
</tr>
<tr>
<td>23400</td>
<td>16</td>
<td>club A - club K</td>
<td>club A</td>
<td>Player A</td>
<td>attacking action</td>
<td>3</td>
</tr>
<tr>
<td>6134200</td>
<td>28</td>
<td>club A - club C</td>
<td>club C</td>
<td>Player H</td>
<td>pass</td>
<td>3</td>
</tr>
<tr>
<td>2827200</td>
<td>33</td>
<td>club A - club U</td>
<td>club A</td>
<td>Player I</td>
<td>pass</td>
<td>4</td>
</tr>
</tbody>
</table>

... Attribute | Definition |
... ['head'] | ['isPossessionLoss', 'isPassWide', 'isPassShort', 'isAerial'] |
... ['leftfoot'] | ... |
... ['head', 'dueltouched'] | ['isDuelPart', 'isDuelWonByAttacker', 'isDuelAir', 'isAerial'] |
... ['throwkeeper'] | ['isPassCompleted', 'isPassForward'] |
... ['high', 'leftfoot'] | ['isPassCompleted', 'isPassForward', 'isPassLong'] |

<table>
<thead>
<tr>
<th>X inmotio</th>
<th>Y inmotio</th>
<th>Speed</th>
<th>Acceleration</th>
<th>Possession</th>
<th>Possession time</th>
</tr>
</thead>
<tbody>
<tr>
<td>-20.707</td>
<td>4.473</td>
<td>1.76</td>
<td>-2.13</td>
<td>club I</td>
<td>6100</td>
</tr>
<tr>
<td>-9.268</td>
<td>18.113</td>
<td>0.85</td>
<td>-0.29</td>
<td>club A</td>
<td>15800</td>
</tr>
<tr>
<td>14.919</td>
<td>-28.968</td>
<td>1.05</td>
<td>1.08</td>
<td>club A</td>
<td>2100</td>
</tr>
<tr>
<td>-48.5</td>
<td>-10.972</td>
<td>3.67</td>
<td>-0.62</td>
<td>club C</td>
<td>3200</td>
</tr>
<tr>
<td>-0.313</td>
<td>6.639</td>
<td>1.39</td>
<td>0.06</td>
<td>club U</td>
<td>62500</td>
</tr>
</tbody>
</table>

4.3 Conclusions

The used data needs to be updated regularly. This is because there will be new matches played. All of these new matches also need to be added to the data. Although there are multiple steps needed to put all the data in the correct format this could be done automatically, when both the tactical data and positional data are obtained.
Chapter 5

Concepts, Models and Methods

This chapter will introduce the concepts, models and methods. Figure 5.1 shows a model of the concepts in the research. In the first step (a) all moments, one moment for every 100 milliseconds in the match, are filtered to identify the interesting moments. In the second step (b) the analysis is done. Within step (a) the moments are filtered based on one or more criteria. Within step (b) the analysis can consist of one or more analysis methods.

![Figure 5.1: The analysis process starting with the selection of the moments in (a). And analyzing the critical moments in (b)](image)

The majority of introduced models utilizes the model of the field as defined in figure 4.2. This chapter introduces the methods used in the research. Chapter 6 explains what the results are of the execution of these methods. Section 5.1 and 5.2 will elaborate more on the criteria and analysis methods used in this thesis.

5.1 Identification of Critical Moments

Before the analysis can start the moments to analyse have to be selected. The concept of critical moments is defined based on certain criteria, which make this moment a critical moment. These criteria could be based on a certain situation, player or other criteria. The critical moments selected, will be further analysed. After the analysis, it could be possible the criteria selected were not the correct criteria. In this case the same analysis can be done, when other moments are selected based on the reviewed criteria.

In this thesis critical moments will be moments when the ball is lost. An example of such a critical moment is displayed in Figure 5.2. These can be found following one of two strategies. Identify critical moments manually from video, or identify the critical moments automatically. Obviously the second strategy is more desirable, since this involves less manual labour.

To find the critical moments, the moments when the club loses possession of the ball are extracted from the tactical data. This however includes a lot of moments which are not really
critical. E.g. when the ball is lost due to a foul; the match is briefly paused and there is less need
to quickly increase the pressure. All moments found are therefore checked further to only extract
the critical moments. Figure 5.3 depicts a flowchart showing what criteria to use, in what order.
When a moment results in ‘not critical’, it is discarded. Otherwise when it results in ‘critical’ it
becomes part of the final set of critical moments.
The following checks are executed:

1. All the moments resulting in a set piece can be excluded. E.g. throw in, free kick, etc.

2. If the number of defenders is smaller than a typically low constant, the moment needs to
be included. Here the number of defenders is the number of the club players with an X-
coordinate smaller than the X-coordinate of the ball as illustrated in figure 5.4

3. If the possession is lost in the last 16.5m of the field, the moment is excluded. This is within
the box in front of the opponents goal. Typically this is an attacking action resulting in loss
of possession.

4. When the ball moved forward half the field (52.5m) since it left possession, the moment is
excluded. Somebody kicked the ball away quickly.

5. When the opponent plays the ball back to their keeper, the moment is excluded. This
removes the critical speed of the possible counter.
5.2 Statistical Analysis of the Critical Moments

Now the critical moments are identified, it is possible to do some statistics. This section discusses the different kinds of statistics calculated on the critical moments identified in section 5.1. For these moments there is taken a more comprehensive look at the 5 seconds after the identified moment where the ball is lost. As explained in chapter 4 there is a data entry every 100 ms. So there are 50 frames in these 5 seconds. For each of these frames the statistics explained in this chapter are calculated. Then for each of these statistics it is possible to plot them over time.

5.2.1 All Moments

To get a benchmark and see what actually happens, all critical moments identified using the method in section 5.1 are analysed. In the next subsections the moments are filtered even more strictly in order to see differences between different kinds of moments.

Area

The pressure put on the opponent consists of the area played on and the forward motion. This indicates the area played on should be analysed. The area spanned by the convex hull from
CHAPTER 5. CONCEPTS, MODELS AND METHODS

Figure 5.5 is a good indication of the area played on. The area of the convex hull can be calculated by using equation 5.1. Where $V$ are all the vertices sorted counter-clock wise. With the first and last vertex being the same point. Equation 5.1 is correct as long as the polygon does not cross itself. And because it represents the convex hull, this will always be the case. This area should be compared with the total area of the field which is $105m \times 68m = 7140m^2$. However, it should be noted during all matches the club played in season 2014-2015 and 2015-2016 the average area of both teams was only $899.13m^2$.

![Figure 5.5: convex hull represented as a polygon.](image)

\[
\frac{1}{2} \sum_{i \in V} x_{i+1} \times y_i - y_{i+1} \times x_i
\]  

(5.1)

Apart from the analysis of the area itself, it is also possible to analyse the change of the area played on. Since there is a snapshot every 100ms, it is possible to calculate the change of the area every 100ms. This change of area is in this thesis referred to as the differentiated area.

Centroid

The other part of the pressure is the forward motion. The forward motion can be calculated by looking at the position of the players. The centroid of the convex hull, as depicted in figure 5.6, is an option for this centroid. This centroid is the mean average of the vertices in figure 5.6. When this centroid is analysed at the start and end of a timespan the movement can be calculated. Possibly the centroid of the convex hull is very different from the centroid of the players themselves. In order to get the similarities and differences the centroid of all players, excluding the keeper, can also be analysed. This value will be more accurate when one of the players is very far away from the others. E.g. when a player is injured on the other side of the field. Because when all players are taken into account each player has smaller weight in the calculation. Figure 5.7 visualizes the centroid of all players.

Heat Map

In order to get an indication of where on the field the ball is lost more often, a heat map of the possession loss moments can be generated. To generate this heat map, the polygons, defining the convex hull in figure 5.5, are drawn with high transparency on top of each other. A similar heat map can also be drawn some time after the possession is lost, to see whether the position on the field changes.

Instead of plotting the polygon, it is also possible to plot the centroids. This will give a more clear picture when there are lots of polygons since a centroid can be represented by a small dot.

1 clock wise works as well, however the algorithm calculating the vertices returns them already sorted counter-clockwise.
5.2.2 Subset of the Critical Moments

Instead of looking at all moments, it is also possible to even further filter the critical moments. The moments could be filtered on the effectivity of the defence. When the ball is back in possession of the club quickly the defence can be considered effective. In a similar fashion when the opponents see a change to shoot on goal quickly, the defence can be considered less effective.

While analysing the effectivity of the defence yields to some insights, it can also be interesting to select the subset of moments in different ways. One of the choices is the opponent. When a subset is taken based on the opponent, it becomes possible to analyse only the top opponents and only the bottom opponents. When both of these are analysed, the differences can be analysed.

It is interesting to see if the features explained in subsection 5.2.1 are similar or different when only these subsets of critical moments are analysed.

5.3 Data aggregation

In a lot of cases it will be enough to have some data for each moment and it is not needed to have all exact locations. To get this data without recalculating it every time, a file with some aggregate data is created for each match. This file will consist of an entry every 100ms with the following data for both the club and the opponent.

Timestamp The time since the start of this match

Count The number of players, of whom the position is accounted. Most often 10, since the goal keeper is omitted.

Area The area of the convex hull of the player positions in squared meter.

Hull centroid The centroid of the convex hull.

Player centroid The centroid of the player locations.

Hull points The locations of the players defining the convex hull, sorted counter-clockwise.

5.4 Statistical Analysis during the match

Instead of analysing only some part of the match before or after critical moments, it is also possible to see if the features change during the match. For example, players start to cut corners, and put more slowly pressure on the opponent at the end of a match.

A big problem when analysing the whole match is the tiny sample size. There are only 34 played matches in a season. So before the sample size is big enough to get statistical significant results, it will take a lot of seasons.
First the area spanned by the convex hull can be analysed. This is done by plotting the area over time during the match. When some defining events of the match are also put in the plot, it becomes possible to see if the area changes significantly after some events. This can be used to see if the team is playing different when the club is ahead in score. When also the team currently in possession of the ball is plotted, it becomes possible to see if possession changes the area significantly. Instead of analysing the area during the whole match, it is also possible to only analyse the area change after the critical moment identified in section 5.1.

Furthermore, also the centroids can be monitored during the match. When analysing the centroid during the match, it can be compared with the centroid of the opponent. And see if a difference between the centroid of the opponent and the centroid of the club increases the probability for some events. Since in the data all teams always play from left to right. When comparing the opponent and the club, one of the two needs to be transformed. In this thesis we will transform the data of the opponent, to make them play from right to left.

5.5 Classification

In order to see if the cause of possession loss is poor play or simply bad luck, a classifier can be created. When this classifier can identify the moments possession will be lost, the cause will be poor play. Then these moments can be analysed and the club can then train to avoid this play in the future.

First the classes for the classifier need to be determined. In this thesis four classes are used. Two classes where a team has possession and two classes where a team gains possession. Then these classes can be divided in which team is (just recently) in possession. Figure 5.8 illustrates these four classes. The four classes are defined as follows:

1. The opponent gains possession.
2. The club gains possession.
3. The club keeps possession.
4. The opponent keeps possession.

Different kinds of classifiers can be constructed. These classifiers can use all the available data and find the best model.

![Decision Tree](image)

Figure 5.8: A decision tree depicting how the cases are classified.
Table 5.1: The features, other than the tactical data features, the classifiers use to extract more insights. All of these features are added both for the club and for the opponent.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area</td>
<td>The area currently covered by the players as depicted in figure 5.4.</td>
</tr>
<tr>
<td>Area_change</td>
<td>The change of the area covered by the club within the last 2500 ms.</td>
</tr>
<tr>
<td>Centroid_distance</td>
<td>The distance between the hull centroid of the club and the player centroid of the club.</td>
</tr>
<tr>
<td>Max_vertex</td>
<td>The length of the longest edge of the convex hull of the club.</td>
</tr>
<tr>
<td>X_hull_centroid</td>
<td>The X-coordinate of the centroid of the convex hull.</td>
</tr>
<tr>
<td>X_max</td>
<td>The X-coordinate of the most forward player.</td>
</tr>
<tr>
<td>X_median</td>
<td>The median of all player X-coordinates.</td>
</tr>
<tr>
<td>X_min</td>
<td>The X-coordinate of the most backward player. (excluding the keeper)</td>
</tr>
</tbody>
</table>

The first classifier will be a decision tree. This tree looks at a single attribute and divides the moments in multiple branches, until there are no splits anymore increasing the model.

Next rule induction is applied in order to find a good classifier. This classifier uses multiple attributes in a conjunction. And expands by adding more attributes in the conjunction as long as the confidence keeps increasing. After the rule induction, the rules found by the rule induction are used as association rules. So the different rules can be compared to each other in a more fair way. These rules can be used as association rules, by adding each part of the and-clause to the antecedent, and the class the consequent.

For all of these classifiers we provide the algorithm with as many features as possible. All the tactical data is used as features. Then, there are some features, extracted from the positional data, added. Half of these features are depicted in table 5.1. Every feature in table 5.1 also has a counterpart with the data of the opponent, this counterpart has a suffix of ‘_opp’.

5.6 Local Process Models

Instead of analysing a fixed time before or after the possession loss, it could also be interesting to analyse a fixed amount of events instead. When this is done some concepts, models and methods described before do not provide a lot of insights. This is why local process models can be used instead. Local process models analyse what events happen in what order. And can help to identify the order of events in a process. Here the process is the football match itself and the events are the events as defined in the tactical data.

In this thesis a closer look will be taken at what events happen before possession is lost. When the models of only the moments involving one specific player then are compared to the models of the moments involving another player, there is gathered information about the contribution of this player.

5.7 Visualization

In order to give the user information which is easy to use, I have developed a graphical user interface (GUI).

A lot of times it is useful to review a specific moment. Figure 5.9 shows the GUI developed to review specific moments. In this GUI it is possible to see all the critical moments and how these moments developed within 5 seconds after possession loss. Furthermore, some Key Performance Indicators (KPIs) are given for the instance currently viewing.

In the top of the screen there are 3 button groups. These groups have the following functions.
Figure 5.9: The moment viewer, the moment select part (a), the field part (b) and the KPI part (c).
CHAPTER 5. CONCEPTS, MODELS AND METHODS

Top specifying the current match

Middle specifying the current moment

Bottom specifying seconds after the current moment

Underneath the control buttons the field with the players is displayed. The field dimensions are the same as in figure 4.2. The image of the field consists of the following parts.

white lines The lines on the field, for easier visual reference

white circle Indication of the position of the ball

red triangles the players of the club, playing left to right

black circles opponents, playing right to left

orange polygon The convex hull, the area spanned by the players

black crosses The centroid of the convex hull and the centroid of the players

black edge The longest edge of the convex hull

On the bottom of the screen some KPIs for this specific moment are displayed.

distance The area covered by the convex hull and the percentage of the total field

centroid The centroid of the players

hull centroid The centroid of the convex hull

distance The distance between both centroids

max vertex The length of the maximum vertex in the convex hull

It is possible to add even more features. Some features can still be useful although they are not needed in the statistical analysis. An example of such a feature is the player currently in possession of the ball. This can be very useful when analysing just a single moment. However it will not be useful during the statistical analysis.

Furthermore to help the coaches to read the results of a match. An extra image can be generated, showing the field with the average positions of the players at different times after possession loss. Figure 5.10 shows an image generated for the coach. The image shows the average covered area, during one match. The green polygon shows the area just after the possession is lost, the yellow area shows the average area at the most forward position of the players. And the red area shows the area after 5s. The different areas are drawn on top of each other in order to show them all.
Figure 5.10: An image of the field to be used by the coach in order to instruct the players. The green polygon is the area after possession loss, the yellow polygon is the most forward area, and the red area is the area after 5s.
Chapter 6

Results and evaluation

This chapter discusses the results of the experiments proposed in chapter 5. After the results of these experiments are presented, the results are evaluated and put in perspective. Section 6.6 will elaborate on several striking ideas in observed during the analysis. For some of these striking ideas the section introduces and discusses some hypotheses.

6.1 Identification of Critical Moments

In order to evaluate the model, first the second half of one of the matches is reviewed. During this review critical moments are reviewed manually. After the critical moments are identified manually, these critical moments are compared to the critical moments found automatically. Then the critical moments are also compared to the moments selected by a specialist of the club.

Table 6.1: The critical moments identified manual versus identified automatic.
The time values are this high, because it is the second half and the counter keeps counting during half time.

<table>
<thead>
<tr>
<th>Manual</th>
<th>Automatic</th>
<th>Specialist</th>
</tr>
</thead>
<tbody>
<tr>
<td>61:31</td>
<td>64:10</td>
<td>64:38</td>
</tr>
<tr>
<td>64:48</td>
<td>64:46</td>
<td>65:15</td>
</tr>
<tr>
<td>67:57</td>
<td>67:48</td>
<td>68:28</td>
</tr>
<tr>
<td></td>
<td>68:29</td>
<td>68:52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70:45</td>
</tr>
<tr>
<td>71:11</td>
<td>71:40</td>
<td>71:94</td>
</tr>
<tr>
<td></td>
<td>72:05</td>
<td>73:54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>74:47</td>
</tr>
<tr>
<td>80:41</td>
<td>80:37</td>
<td>77:22</td>
</tr>
<tr>
<td></td>
<td>84:48</td>
<td></td>
</tr>
<tr>
<td>85:41</td>
<td>85:29</td>
<td></td>
</tr>
<tr>
<td>94:21</td>
<td></td>
<td>95:11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>96:03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>97:03</td>
</tr>
<tr>
<td>98:21</td>
<td>98:26</td>
<td>98:17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>99:08</td>
</tr>
<tr>
<td>104:41</td>
<td>103:54</td>
<td>104:21</td>
</tr>
<tr>
<td>106:11</td>
<td>107:33</td>
<td>106:12</td>
</tr>
<tr>
<td>109:56</td>
<td>109:46</td>
<td></td>
</tr>
</tbody>
</table>

As can be seen from Table 6.1 the data moments do not translate one to one. This is because
I manually checked for critical moments, automatically there is searched for events resulting in critical moments. Furthermore, there will be a difference, because the timestamps are collected manually, the timestamp used as input for the automatic analyses is collected manually, the specialist manually identified the moments and the manual timestamp is obviously collected manually. This will result in some differences due to the reaction times.

Furthermore, the critical moments found automatically are compared to the moments identified by a specialist of the club.

Table 6.2: Contingency table displaying how many moments are found automatically and by the analyst

<table>
<thead>
<tr>
<th></th>
<th>Automatically found</th>
<th>Automatic not found</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Found by analyst</td>
<td>10</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>Not found by analyst</td>
<td>3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>7</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 6.2 shows the contingency table of the automatic identification and the identification done by the specialist. So out of the 20 moments found in total, 10 of the moments were found both automatically and by the specialist. This indicates there is still room for improvement, this is however a good start.

6.2 Statistical Analysis of the Critical Moments

This section discusses whether the models and methods introduced in section 5.2 are valid, and if the results are similar to the expected results. In order to do so, first an expected result from each method is discussed.

6.2.1 All Moments

In a similar way as in section 5.2, first all moments are discussed. This subsection discusses whether the results in all matches are as expected, and where there are interesting moments.

For the following features, all 2455 critical moments identified are used. This results in a sample size of \( n = 2455 \) for all statistics in this subsection. Some of the following graphs will have error bars. These error bars indicate the standard error.

These are the features for all moments identified using the methods described in section 5.1 for season 2014-2015 and season 2015-2016.

**Absolute Area**

The tactics the club uses implies the area will decrease after possession of the ball is lost. This is because after possession loss the club tries to put more pressure on the opponent by decreasing the area. Figure 6.1 shows the mean of the absolute area covered by the players five seconds after they lost possession of the ball. The graph in figure 6.1 displays the mean of the area over all ball possession loss moments in seasons 2014-2015 and 2015-2016. As can be seen in figure 6.1 on average the absolute area decreases within five seconds after possession loss.

Figure 6.1 depicts the area indeed decreases after possession is lost. This is what was expected.

**Differentiated Area**

The club tries to put a lot of pressure quickly. This implies the area decreases quickly after possession loss. However, this also implies the preferred (smaller) area is reached quickly. When the covered area is small enough, the area does not need to shrink anymore. So the decrease will be slower.
CHAPTER 6. RESULTS AND EVALUATION

Figure 6.1: The mean of the area 5 seconds after the possession is lost.

Figure 6.2: The differentiated area covered by the convex hull in time. The dashed line is the closest parabolic function.

Figure 6.2 shows the differentiated area. As can be seen in figure 6.2 the area decreases more shortly after the possession is lost. After 5s the decrease approaches 0, which indicates the area stays almost constant. The differentiated area behaves as expected. The area indeed decreases more quickly when the possession is just lost. After 3s the decrease of the area slows down.

**Centroid**

When reviewing the tactics the club uses the team should move forward, when the pressure is increased. Which implies the X-coordinate in the model will increase. So it is expected the X-coordinate of the hull centroid will move forward. The tactics do not indicate whether the team should move left or right. So the expectation is the Y-coordinate does not significantly change.

The centroid of all players is expected to behave similar to the centroid of the players on the vertices only, since all players are expected to move forward and so increase the pressure.

Figure 6.3 shows the movement of the centroid of the convex hull over time. The X-coordinate first increases but after about 2.5s starts decreasing. The Y-coordinate does not significantly change in time. Figure 6.4 shows the movement of the centroid of all players. The behaviour of this centroid is very similar to the centroid of the players on the vertices only.

The error increases with the time, this could be expected, when more time is elapsed, there is also more time for the players to move around.

Figure 6.3: The change of the convex hull centroid after critical moments.

Figure 6.4: The movement of the player centroid in time.
CHAPTER 6. RESULTS AND EVALUATION

The expectation of the X-coordinate seems to be partially true. The centroid does increase. However, the drop after a few seconds was not expected. The Y-coordinate is expected to stay around 0. There is actually not a lot of change in the Y-coordinate, the Y-coordinate stays close to 0.

Heat Map

It is expected most of the field is coloured in the heat map, since the possession is lost all over the field. However, since the play occurs less in the edges of the field, the corners of the field should have lower heat.

Figure 6.5 depicts all polygons the moment the possession is lost. In figure 6.5 it can be seen the centre of the field is coloured red, around the edges it is more white, this indicates the field in the centre is mostly covered during possession loss. The edges of the field are less covered overall. Figure 6.6 depicts all polygons 5 seconds after the possession is lost. Figure 6.5 and figure 6.6 are similar. However, the size of the latter is slightly smaller, which can be seen most clearly around the left 16m boxes.

Figure 6.5: Polygons when possession is lost.  Figure 6.6: Polygons 5s after possession loss.

The expectations are partially true, the fact there is less heat near the edges indicates the part near the edges is less often covered after possession loss. This could be explained, when the ball is lost on the right wing the left wing will be less covered and vice versa.

Similarities between figure 6.5 and figure 6.6 is in accordance with what is shown in figure 6.1 and figure 6.3. After 5 seconds the centroid is almost returned, the area is however decreased. Resulting in a smaller coloured area at a similar location.

The centroids are expected to be mostly around the middle of the field. This is because, the moment possession is lost, there will be players on all sides around the middle. Figure 6.7 shows the centroids the moment possession is lost. Most of the centroids swarm around the middle. Figure 6.7 looks as expected, most centroids swarm around the middle, however it should be noted there are a few centroids in the box the club is defending. This indicates the majority of the team is within the box the moment possession is lost.

Conclusions

The different graphs indicate the area is decreasing and the centroid is moving forward. Which in turn indicates the club is putting pressure on the opponents. However, it is striking to note the centroid only moves forward 2.5s after possession loss and then moves back for the next 2.5s, ending on the same height. This behaviour can be explained by the fact the opponent has possession of the ball after possession loss. And the opponent will start attacking, pushing the players back.
6.2.2 Well-Handled Moments

Now only the subset of critical moments in which the ball is returned to the possession of the club within 10s is analysed. This results in a decreasing sample size, as the moment the possession is returned to the club the moment should be removed from the analysis. Figure 6.8 shows the sample size for these critical moments. Note the sample size does not reach 0 since there are also moments in which the ball is returned to the club in the interval $[5\text{s}, 10\text{s}]$.

Now the already seen features will be discussed, comparing all moments to only the well-handled moments discussed in this subsection. In the figures, the left graph is the graph from subsection 6.2.1. The right graph will be the graph of only the well-handled moments. The standard error will increase for most features, since the sample size is only $\frac{1}{3}$ of the sample size of all moments.

Absolute Area

It is expected, when only looking at the well-handled moments, the absolute area decreases more. This is because the club uses the tactic to make the field small when possession is lost, in order to get possession back again quickly.
CHAPTER 6. RESULTS AND EVALUATION

Figure 6.9: The absolute area after critical moments, the left graph depicts all moments, the right graph depicts the well-handled moments.

Figure 6.9 depicts the absolute area. As can be seen both graphs are similar. However, the absolute area after 5s when only considering the well-handled moments is slightly bigger. Shown in figure 6.9, there is hardly any difference between all moments and the subset of moments where the ball is back in possession quickly. This is unexpected. It is expected the area will decrease more, however the decrease is similar.

Differentiated Area

The tactics the club uses indicate, the area is decreased quickly after possession loss. This is done so the ball will be back in possession quicker. When the area decreases more quickly the ball should be back in possession more quickly. So it is expected the area will decrease quicker in the subset. Which will result in a lower value in the graph of the subset.

Figure 6.10: The differentiated area after critical moments, the left graph depicts all moments, the right graph depicts the well-handled moments.

Figure 6.10 depicts the differentiated area. The differentiated area has a lot more spikes. This is
mainly because of the sample size though. Apart from the spikes, the differentiated area is more or less similar to the differentiated area of all moments.

Both graphs in figure 6.10 look similar apart from the spikes. This is not as expected, since it is expected to have a more sudden decrease when only well-handled moments are used.

**Centroid**

When the club gets the ball back quickly, it is expected there is a lot of pressure quickly, this results in a quicker higher peak in the X-coordinate of the hull centroid. So the expected hull centroid will have a higher peak quicker. The Y-coordinate is expected to stay around 0, since the tactics are not to move left or right.

The change of the player centroid is expected to be similar to the change of the hull centroid. So the well-handled moments, are expected to have a quicker and higher peak for the X-coordinate. The Y-coordinate is expected to stay at 0 as well.

![Figure 6.11: The change of the convex hull centroid after critical moments. The left graph depicts all moments; the right graph depicts the well-handled moments. The blue lines are the X-coordinates, the green lines are the Y-coordinates.](image)

Figure 6.11 depicts the movement of the convex hull centroid. The right graph depicting the well-handled moments, peaks a little earlier and lower, at about 2s instead of 2.5s in the left graph with all moments. Furthermore, the X-coordinate is negative 4.5s after a well-handled critical moment.

Figure 6.12 depicts the player centroid for both all moments and only the well-handled moments. The peak of the X-coordinate is lower when looking at the well-handled moments. The value after 5s is ± − 1. The Y-coordinate is close to 0.

Figure 6.11 is unexpected since it does not indicate a higher peak for the well-handled moments. However, it does depict a slightly earlier and lower peak. When the X-coordinate starts decreasing again, this indicates the players are moving back, which is not expected, or the moments where the players moved forward the most, are filtered out of the set as the possession is retrieved.

Figure 6.12 has a number of unexpected features. First of all, the peak is lower. This is very unexpected since it is expected the peak was actually higher and quicker. The peak also is quicker when only the well-handled moments are analysed. The value of the X-coordinate after 5s is ± − 1, which is unexpected for similar reasons as the convex hull centroid. The Y-coordinate is about 0. This is as expected.
CHAPTER 6. RESULTS AND EVALUATION

Figure 6.12: The change of the player centroid after critical moments. The left graph depicts all moments; the right graph depicts the well-handled moments. The blue lines are the X-coordinates, the green lines are the Y-coordinates.

Heat Map

It is expected the location of possession loss when the moment is handled well is similar to the location of possession loss when all moments are analysed. Since the location of possession loss is not expected to influence how possession loss is handled. It is expected the convex hulls just after possession loss and the convex hulls 5s after possession loss are similar for all moments as well as when the ball is obtained back quickly.

Figure 6.13 depicts both hulls for all moments and the well-handled moments. The coloured area is in the well-handled moments significantly smaller. The centre is in a similar place.

Figure 6.13 is a little different than expected. The only unexpected part is the size of the coloured area. This is probably because of the size of the sample size.

Conclusions

The features for the well-handled moments are almost all unexpected. The most unexpected part is the fact the centroid does not move as far forward when possession is obtained back quickly. This is strange since less forward movement indicates less pressure, but still possession of the ball is retrieved quickly.

When the ball is retrieved quickly, the pressure put on the opponent is not necessarily higher. Before the possession is back, the players are moving back. The amount of backward movement is similar when the ball is retrieved as for all moments. So surprisingly there is not a lot of difference in pressure put on the opponent when the ball is back in possession quickly.

6.2.3 Poorly Handled Moments

This subsection will elaborate about the poorly handled moments. The poorly handled moments are the moments when the opponent shoots on goal within 10s after the possession loss.

Figure 6.14 shows the sample size of the subset of moments which are handled poorly. The subset of moments analysed in this subsection is decreasing similar to the subset of the well-handled moments. It should be noted the sample size is even smaller than the sample size of the well-handled moments, this introduces even bigger errors in the features.

It should be noted the sample size is a lot smaller than the sample size of all moments. It is about $\pm \frac{1}{30}$ of the whole set of all critical moments.
CHAPTER 6. RESULTS AND EVALUATION

Figure 6.13: The heat maps of convex hulls just after critical moments and 5s after the critical moments. The left graph depicts all moments; the right graph depicts the well-handled moments.

Figure 6.14: The sample size of critical moments, when the opposition attempts to score a goal within 10s.

**Absolute Area**

When only the poorly handled moments are analysed, it is expected there is less pressure on the opponent. This will result in a smaller decrease than when all moments are analysed. Therefore,
it is expected the area decreases less when looking at the subset of poorly handled moments.

Figure 6.15: The absolute area after critical moments. The left graph depicts all moments; the right graph depicts the poorly handled moments.

Figure 6.15 depicts the area of all moments and the poorly handled moments. The obvious difference is the error which is a lot bigger. Furthermore, the area is about 20$m^2$ bigger. The area itself however still is decreasing.

The slightly bigger area when only the poorly handled moments are viewed was expected, since the pressure is lower. The bigger error is mainly explained by the smaller sample size.

**Differentiated Area**

As already explained above it is expected there is less pressure on the opponents, in the poorly handled moments. This is expected to result in a slower decrease of the area. So a higher value for the differentiated area.

Figure 6.16: The differentiated area after critical moments. The left graph depicts all moments; the right graph depicts the poorly handled moments.

Figure 6.16 depicts the differentiated area. The right graph, the poorly handled moments, has a
lot more peaks. The trend line in the right graph is about 2s lower than with all moments, this indicates a quicker decrease. After 4s the line becomes positive, this indicates an increase of the area.

The differentiated area was expected to decrease slower, however the area decreased even faster after 2s. After 4s the trend line becomes positive indicating the area even increases. This means less pressure, which is expected with the poorly handled moments.

**Centroid**

Due to the expected lack of pressure, the centroid of the convex hull is expected to not move forward as much. So the X-coordinate should peak lower.

The player centroid is expected to behave similar to the hull centroid. Therefore, the centroid is expected to move less forward.

![hull centroid movement]

Figure 6.17: The change of the convex hull centroid after critical moments. The left graph depicts all moments; the right graph depicts the poorly handled moments.

Figure 6.17 shows the movement of the convex hull’s centroid. The X-coordinate is totally different. It starts going similar to the X-coordinate over all moments, however after 1s the forward motion is already smaller. After the first second the X-coordinate plummets. The Y-coordinate stays around 0.

Figure 6.18 depicts the movement of the player centroid. Similar as with the convex hull centroid, The X-coordinate quickly starts dropping and the Y-coordinate stays more or less the same. The centroid of the convex hull was expected to peak lower. This was actually observed in the data. However, the enormous drop was not expected. This can be explained, because the opposition came in a position to attempt scoring a goal. Probably the opposition converted the gain of possession in a real attack, pushing the club back. This explains the drop.

The behaviour of the player centroid is similar to the hull centroid. The unexpected drop can be explained for similar reasons as for the hull centroid.

**Heat Map**

When only the moments resulting in a goal attempt of the opposition are analysed, it is expected the ball is lost closer to the goal the club is defending. This will make it easier for the opponent to attempt to score a goal.

Figure 6.19 depicts the start and end hulls. The poorly handled moments are lighter, this is just the number of polygons drawn, due to the lower sample size when analysing the poorly handled
CHAPTER 6. RESULTS AND EVALUATION

Figure 6.18: The change of the player centroid after critical moments. The left graph depicts all moments; the right graph depicts the poorly handled moments.

Figure 6.19: The heat maps of convex hulls just after critical moments and 5s after the critical moments. The left graph depicts all moments; the right graph depicts the poorly handled moments.

moments the field ends up less bright. Furthermore, it can be seen, the hulls are mainly on the
left part of the field, the part the club is defending. The heat maps in figure 6.19 are as expected. Possession is lost mostly on the half the club is defending. When the opponent was in a position to attempt a goal within 10s, the convex hulls even move a little more to the goal the club is defending.

Conclusions
When the opponent gets an attempt to score a goal within 10s after possession is lost, both the area does decrease less, and the centroid does not move forward as much. This can be explained because the opponent got an attempt to score, and therefore moved forward quite a bit. However just after the possession is lost, the pressure on the opponent is still increased.

6.2.4 Opponents
This subsection will elaborate on the statistics when playing versus different opponents. Each feature will be first analysed for the top teams and after this for the bottom teams. For easy reference the images from all opponents are also shown again. The left graph shows the top opponents; the middle graph shows all opponents and the right graph shows the bottom opponents. When analysing the different opponents, the sample size does not change. Since all moments versus the opponent are analysed 5 seconds after possession loss. Similar to all moments. The sample size of the top opponents is 341 and of the bottom opponents is 215. This is ±17 and ±11 of the total sample size respectively. It should be noted, there are 8 matches played against the top opponents and 6 matches against the bottom opponents in seasons 2014-2015 and 2015-2016. The average number of critical moments in each match therefore is 42.6 for the top opponents and 35.8 for the bottom opponents.

Absolute Area
When only analysing the matches versus the top opponents, it is expected there will be put more pressure on these opponents. Since the players are more focused and know they have to be stricter. This is expected to be visual by having a smaller area at most times. When playing the bottom opponents, this effect is expected to work the other way around, and so will increase the area covered by the players.

Figure 6.20: The absolute area while playing different opponents. The most left graph displays the absolute area when only the top teams are accounted for; the middle graph displays the absolute area of all matches and the right graph displays the absolute area when only the bottom opponents are accounted for.

Figure 6.20 displays all three graphs. It is interesting the graphs of both the top and bottom opponents start a little lower than the graph of all opponents. Furthermore, it should be noted the left graph, when playing top opponents, is lower after 5 seconds. And the right graph depicting the bottom opponents is very similar at the end.
CHAPTER 6. RESULTS AND EVALUATION

When playing versus top opponents the expectation is partly correct. The area when losing the possession is just a little smaller. However, after 5 seconds the area is decreased more. This probably has to do with the focus of the players. When playing versus bottom opponents, the expectation also was partly correct, the area was a bit smaller, however after 5 seconds the area was almost as big as when averaging over all opponents.

Differentiated Area

The increased focus of the players when playing versus top opponents is expected to result in a quicker decrease when playing top opponents and it is expected to decrease slower when playing a weaker opponent.

![Differentiated Area](image)

Figure 6.21: The differentiated area while playing different opponents. The most left graph displays the differentiated area when only the top teams are accounted for; the middle graph displays figure 6.2 the differentiated area of all matches and the right graph displays the differentiated area when only the bottom opponents are accounted for.

Figure 6.21 shows the differentiated area for the top opponents, all opponents and bottom opponents. The left graph depicts the top opponents. It can be seen the graph start lower than when looking at all opponents. This indicates a quicker decrease after possession loss. The right graph depicts only the bottom opponents. It can be seen the graph is higher and even positive at the end. This indicates a slower decrease and even a little increase at the end. Figure 6.21 is similar to the expectation. The left graph depicting the top opponents, is lower indicating a quick decrease just after possession loss. The negative start, indicates a more alert response of the team. The right graph is higher and more flat, indicating a slower decrease.

Centroid

When playing versus top opponents, it is expected the pressure is higher. This means that the area will be decreasing, but also the movement is forward. This leads to the expectation the team moves forward more quickly after the possession is lost when playing versus top opponents. When playing versus weaker opponents, it is expected there is put less pressure, and therefore the movement will be less and slower.

Figure 6.22 depicts the centroid movement versus different opponents. The left graph depicts the centroid versus the top opponents. The movement of the centroid of the convex hull is similar to the movement of the centroid when all moments are analysed. The right graph depicts the centroid versus the bottom opponents. The forward movement versus the bottom opponents is less and even reduces quicker than when all moments are analysed. The Y-coordinate is also positive when playing versus the bottom opponents.

The expectation versus top opponents is not met. The movement of the centroid is similar to all moments, it was expected there was a quicker movement. When playing the bottom opponents, the centroid movement is less, and the team is even pushed back more than overall in matches. It is also interesting to see the Y-coordinate is positive when playing the bottom opponents. This means the bottom opponents chooses the play more over the right wing.
CHAPTER 6. RESULTS AND EVALUATION

Conclusions

When playing versus the top opponents the pressure put on the opponent is a little quicker. Indicated by a quicker decrease of the area. However, the centroid of the players does not change significantly. So the top opponents are put under slightly more pressure.

When playing versus the bottom opponents the pressure put on the opponent is less than when playing versus the other teams. This can be seen both by the slower decreasing area and the centroid moving less forward. So the bottom opponents are put under less pressure.

6.3 Statistical Analysis During the Match

The statistical analysis can be done for the critical moments, although, the whole match could be analysed instead. This section will elaborate on the whole match. The 34 matches are a too small sample size to do statistical analysis. When these 34 matches are put in one figure it becomes too crowded, as can be seen in appendix A. Therefore, in this thesis the analysis is done for just one match. For other matches however the analysis would be similar. In this thesis the match club A-club H in the season 2015-2016 is used.

First the area spanned by the convex hull is plotted during the match. It is expected the area is very unsteady. Every time a team attacks the area is increasing and when the team again defends, the area will decrease again, resulting in a lot of movement.

Figure 6.23 depicts the area during the match of both club A and club H. The area does increase and decrease very often. However when the area decreases when one team is defending, the other team is attacking at the same time, so the lines should separate all the time. However, the lines are often close together.

Furthermore, when analysing the match itself, apart from identifying the critical moments as explained in 5.1, the moments where the area peaks could be selected as interesting moments, the moments where the area of the club and the opponent diverge could also be selected as interesting moments.

During the match the critical moments occur during different stages of the match. In order to see whether the area change decreases over time, when the players are getting tired, the change of the area in the first 2.5 seconds is also plotted.

It is expected the change of the area decreases over time, because the players will get tired and move slower. It is expected later in the match the players are getting tired and therefore the area decreases slower.
CHAPTER 6. RESULTS AND EVALUATION

Figure 6.23: The area spanned by the convex hull during the match. On the horizontal axis the important events in the match are indicated. The score is added when there is scored, the ‘y’ indicates a player got a yellow card. The colour of the font indicates of what team the player was. The dashed line indicates the critical moments identified in this match. On top the team currently in possession of the ball is indicated.

Figure 6.24: The area change within the first 2.5s after possession loss

Figure 6.24 depicts the change of the area during the first 2.5s after possession loss. The change of the area is not as expected. Often it is actually positive, indicating the area is actually increasing, while it should be decreasing. However, the change does get less extreme later in the match. Which could indicate the players are getting tired.

It is also interesting to see how the centroids behave during the match. It is expected the centroids of club A and club H stay close together. This is because when the players defend, they tend to stay close to an opponent, resulting in the centroids staying close together.

Figure 6.25 depicts both the hull centroid and the player centroid. The centroids stay very close together as expected. Both centroids, the centroid of the convex hull and the centroid of all players, are very similar. Note around the 3-0 the hull centroid is unknown. This is because there is a penalty and the tracking system loses many players. Interesting moments during the match can be identified at the moments when the centroid of the club and the centroid of the opponent diverge.

Conclusions

When analysing the whole match, the area and the centroid of both teams are very similar. However, this analysis can be used for further selection of the interesting moments. This can be done by analysing at what moment the centroids peaks or when the centroids of the club and the opponent diverge. When analysing the whole match, it becomes very hard to compare multiple matches since the behaviour is very different during the matches.
CHAPTER 6. RESULTS AND EVALUATION

6.4 Classification

In this section different kinds of classification techniques will be discussed. First a decision tree is discussed. And rule induction is utilized to obtain a classifier.

Table 6.3 displays how many times the next entry is in what class. What will happen between the current entry and the next entry. This is what we want to predict with the classifiers. Whether the ball will stay in possession of the same team, and which team has possession.

Table 6.3: The number of moments divided in the classes. The moments are in the class of the next moment, so the class it should have predicted.

<table>
<thead>
<tr>
<th>Opponent got possession</th>
<th>The club got possession</th>
<th>The club keeps possession</th>
<th>Opponent keeps possession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>Class 2</td>
<td>Class 3</td>
<td>Class 4</td>
</tr>
<tr>
<td>3825</td>
<td>3799</td>
<td>19708</td>
<td>18526</td>
</tr>
</tbody>
</table>

The decision tree does not result in a usable classifier. Figure 6.26 shows the generated decision tree. This generated decision tree is not really a decision tree, since there is not found a good differentiating decision. So all moments are put in the same node. And the most frequent class is selected as the class for this node. In this case that class is class 3.

Figure 6.26: The generated decision 'tree'.
Blue represents class 3, Cyan represents class 1, Green represents class 2 and Red represents class 4.
Next Rule induction is used to generate rules. For each rule greedily conditions are added to the rule. This procedure tries every possible value of each attribute and selects the condition with highest information gain. Then the rules are pruned to be more general. So all rules together will classify each entry in a single class.

The rules are generated on a random subset of the entries (70%), then these rules are tested on the entries not in the subset (30%). Next, the predicated classes, the values from the classifier, are compared to the actual classes. This results in the confusion matrix in table 6.4, which results in accuracy of 57.25%. However this confusion matrix has a kappa of only 0.301. The kappa value compares the accuracy with random choice. A value of 0.301 indicates, the classifier is a third between random choice and the perfect classifier.

Both the precision and recall are a lot better for classes 3 and 4. These classes both are the classes where the same team keeps possession of the ball. It is however promising the classifier can predict which team has possession, based on the attributes.

<table>
<thead>
<tr>
<th>True 1</th>
<th>True 2</th>
<th>True 3</th>
<th>True 4</th>
<th>Total</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pred. 1</td>
<td>77</td>
<td>55</td>
<td>204</td>
<td>137</td>
<td>473</td>
</tr>
<tr>
<td>Pred. 2</td>
<td>54</td>
<td>86</td>
<td>171</td>
<td>254</td>
<td>565</td>
</tr>
<tr>
<td>Pred. 3</td>
<td>625</td>
<td>329</td>
<td>3727</td>
<td>1192</td>
<td>5873</td>
</tr>
<tr>
<td>Pred. 4</td>
<td>418</td>
<td>715</td>
<td>1727</td>
<td>3986</td>
<td>6846</td>
</tr>
<tr>
<td>Total</td>
<td>1174</td>
<td>1185</td>
<td>5829</td>
<td>5569</td>
<td>13757</td>
</tr>
<tr>
<td>Recall</td>
<td>6.56%</td>
<td>7.26%</td>
<td>63.94%</td>
<td>71.57%</td>
<td></td>
</tr>
</tbody>
</table>

This however still results in too many too complex rules to be useful. The first rules are displayed in figure 6.27. After each rule the number of entries is depicted. In words the rules are as follows

When the rules are interpreted

```plaintext
if Timestamp \leq 2690800 and Possession_time > 22500 and Area_opp \leq 534.545 and X_opp_min > 16.465 then 3.00 (142 / 4 / 1529/ 7)

if Timestamp \leq 2644150 and Possession_time > 22500 and Area \leq 476.565 and X_{max} \leq -1.715 then 4.00 (5 / 87 / 3 / 1099)

if Timestamp \leq 2692550 and Possession_time > 2500 and Area \leq 1149.690 and Area_opp \leq 733.340 and Season =2015-2016 then 3.00 (78 / 11 / 787 / 17)

if Timestamp \leq 2693800 and Possession_time > 17500 and Area \leq 959.870 and Area_opp > 970.835 and X_{median} \leq -4.425 and X_{opp_max} > 14.905 then 4.00 (2 / 37 / 3 /383)

... 
```

Figure 6.27: The first four rules generated by the rule induction algorithm.

The rules in figure 6.27 can be interpreted in the following way:

- In the first half, the possession has not changed in the last 22.5s, the opponent is not covering a big area and all opponents are on their defending half, the club keeps possession of the ball.
- In the first half, the possession has not changed in the last 22.5s, the club is not covering a big area and all the club players are on the defending half,
the opponent keeps possession of the ball.

- In the first half, the possession has not changed in the last 2.5s, the club is covering a bigger area than the opponent and in the season 2015-2016, the club keeps possession of the ball.

- In the first half, the possession has not changed in the last 17.5s, the opponent is covering a bigger area, the player in the middle stands on the defending half and the most defending opponent still stands on their defending half, the opponent keeps possession of the ball.

In total there are 45858 entries to be classified. Currently there are 1682 of these entries classified by the first rule, this is 4% of all entries. 1529 entries are correctly classified, this is 91% of the classified entries.

Every rule generated by the rule induction, is generated only based on the entries still not classified by the previous rules. This makes it hard to compare the different rules.

When each rule is applied to all the entries, the rules can be compared. When each rule is applied to all entries the support, confidence and lift of each rule can be calculated.

Table 6.5: The number of entries in each class, the class this rule classifies the entries in, and the support, confidence and lift of the first rules

<table>
<thead>
<tr>
<th>Rule</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>142</td>
<td>4</td>
<td>1529</td>
<td>7</td>
<td>0.033</td>
<td>0.909</td>
<td>2.12</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>87</td>
<td>3</td>
<td>1009</td>
<td>0.024</td>
<td>0.920</td>
<td>2.28</td>
</tr>
<tr>
<td>3</td>
<td>82</td>
<td>11</td>
<td>808</td>
<td>17</td>
<td>0.017</td>
<td>0.880</td>
<td>2.05</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>37</td>
<td>3</td>
<td>385</td>
<td>0.008</td>
<td>0.902</td>
<td>2.23</td>
</tr>
</tbody>
</table>

Table 6.5 shows the number of entries classified in each class, the support, the confidence and the lift for the first rules. Note the two last rules classify slightly more entries than when every rule was applied after each other.

The first four rules start with the restriction $Timestamp$ is smaller than a number close to 2,700,000ms. This translates to the first half, since this is 45 min. The rules are automatically generated, this indicates the prediction would be different or at least less precise in the second half. In order to check this, the timestamp constraint is changed to only include the entries in the second half for the first four rules.

Table 6.6: The number of entries in each class, the class this rule classifies the entries in, and the support, confidence and lift of the first rules but in the second half

<table>
<thead>
<tr>
<th>Rule</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>85</td>
<td>66</td>
<td>895</td>
<td>538</td>
<td>0.019</td>
<td>0.564</td>
<td>1.31</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>61</td>
<td>463</td>
<td>583</td>
<td>0.013</td>
<td>0.512</td>
<td>1.27</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
<td>20</td>
<td>266</td>
<td>130</td>
<td>0.006</td>
<td>0.588</td>
<td>1.37</td>
</tr>
<tr>
<td>4</td>
<td>23</td>
<td>21</td>
<td>122</td>
<td>135</td>
<td>0.003</td>
<td>0.449</td>
<td>1.11</td>
</tr>
</tbody>
</table>

Table 6.6 shows the same values as table 6.5 but now for the second half. All the rules result in a significant lower support, confidence and lift. Indicating in the second half overall the play is significantly different.

Furthermore, in each rule each 'and'-clause can be removed. When a part of the 'and'-clause is removed, the support will increase since more entries will be selected. However the confidence and lift should decrease, otherwise the clause should not be added during the generation of the rule. However without the lose in confidence could be minimal.
CHAPTER 6. RESULTS AND EVALUATION

Table 6.7: First rule when clauses are removed

<table>
<thead>
<tr>
<th>Removed Clause</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>142</td>
<td>4</td>
<td>1529</td>
<td>7</td>
<td>0.033</td>
<td>0.909</td>
<td>2.12</td>
</tr>
<tr>
<td>Timestamp</td>
<td>228</td>
<td>71</td>
<td>2424</td>
<td>545</td>
<td>0.053</td>
<td>0.741</td>
<td>1.73</td>
</tr>
<tr>
<td>Possession_time</td>
<td>174</td>
<td>84</td>
<td>1860</td>
<td>316</td>
<td>0.041</td>
<td>0.764</td>
<td>1.78</td>
</tr>
<tr>
<td>AreaOpp</td>
<td>270</td>
<td>10</td>
<td>2365</td>
<td>20</td>
<td>0.052</td>
<td>0.887</td>
<td>2.07</td>
</tr>
<tr>
<td>Xopp_min</td>
<td>197</td>
<td>45</td>
<td>1943</td>
<td>609</td>
<td>0.042</td>
<td>0.695</td>
<td>1.62</td>
</tr>
</tbody>
</table>

Table 6.7 shows the results for the first rule, with every clause removed separately. The support increases when a clause is removed, and the confidence and lift decrease, when a clause is removed. If the AreaOpp-clause is removed, the decrease in confidence and lift is minimal. So it can be argued the rule would be better without the AreaOpp-clause. In this case, however, the gain in support is not significant enough to explain the decrease in confidence and lift.

Conclusions

Classification of snapshots during the match indicating what team will get or keep possession of the ball is hard. Interestingly enough during the second half, it is harder to predict the possession than in during the first half.

The four rules explored in this thesis all have a quite good confidence. If one entire team is on their own half, the other team will keep possession of the ball.

However it is surprising the rule only holds during the first half. And possibly even more surprising, this only holds when the area is small. This is surprising since the tactics actually imply a smaller covered area will increase possession changes. However the classifier classified it the other way around.

There still could be made improvements to the classifier. One of the improvements is to compare the results above, predicting the next class, to the current class. This will improve the results since there will be events following each other when the possession does not change, within these sequences the current class and the next class are the same.

6.5 Local Process Model

This section analyses what causes the possession loss. The six events before possession loss are analysed. So all the following Petri nets end with the club losing possession after the net. In this thesis the critical moments in the season 2015-2016 were used as defined in section 5.1. There are 1259 of these critical moments.

Figure 6.28 shows the most obvious Petri net found during the local process mining. This is the model where the club passes the ball, a player gets in a duel, and the possession is lost. So out of the 1259 moments in 603 moments the club gets in a duel after one or multiple passes.

Figure 6.29 shows a Petri net, this Petri net depicts a duel where the club did not lose the possession of the ball, this is depicted by the pass of the club after the duel, however within 4 additional actions the possession is lost.

Figure 6.30 shows a Petri net, consisting of both a duel and a pass. This is the case for both Petri nets in figure 6.28 and figure 6.29. However, when both counts are added, $603 + 257 = 860$, this
CHAPTER 6. RESULTS AND EVALUATION

Figure 6.29: A Petri net showing how often first a duel is followed by a pass, the duel is won, however after the pass the possession is still lost.

Figure 6.30: A Petri net showing how often both a duel and a pass happened before the possession is lost.

does not match the 721 in this Petri net. This means that in $860 - 721 = 139$ moments there was either a duel before the Petri net in figure 6.28 or there was a pass before the Petri net in figure 6.29.

Figure 6.31: A Petri net showing how often a bring in is followed by a pass.

Figure 6.31 shows a Petri net, where the club brings in the ball, passes the ball and loses possession. Note that in $396 - 254 = 142$ moments there is no pass after the bring in. However, all moments result in possession loss. So in these 142 moments the possession is lost after the club brings the ball back in the play without passing before. A bring in is defined as a throw in, a corner, a goal kick by the keeper, an (in)direct free kick or a penalty. For some of these actions, it is not a problem there is no pass after the bring in e.g. a penalty. For some of those it is more a problem.

Figure 6.32: A Petri net showing how often a bring in is followed by a pass which is followed by duel. After these events the possession is lost.

Figure 6.32 shows the Petri net of figure 6.31, but followed by a duel. This Petri net shows the bring in and pass are followed by a duel in 87 moments. This also indicates in $254 - 87 = 167$ moments, the possession is lost after a bring in, however not in a duel.

Figure 6.33: A Petri net showing a pass of the opponent followed by a duel of the opponent. After this the club obtain possession of the ball and loses this possession again quickly.

Figure 6.33 shows a Petri net, similar to the Petri net in figure 6.28 however, in this case both the pass and the duel are by the opponent. Although these Petri nets do look very similar. There is a very significant difference, the moments are all selected to end in possession loss by the club.
CHAPTER 6. RESULTS AND EVALUATION

This means in the Petri net in figure 6.33 first the opponent passes, engages in a duel, the club gains possession, and then the club loses possession. All of this within 6 actions. So the club gains possession and loses it again very quickly.

Now there is a basis established, it would be interesting to filter the set. When the set is filtered for a single player, it can be analysed if a player has a positive or negative impact. In this thesis the moments will be filtered, by making sure the sequence starts with a pass of player A before the possession is lost in maximum 6 actions. When the same Petri nets are taken for this subset, the numbers in the transactions change.

Figure 6.34: A Petri net showing how often a pass is followed by a duel resulting in possession loss. But the data is filtered to make sure player A passes before the possession is lost in maximum 6 actions.

Figure 6.34 depicts the same Petri net as in figure 6.28, only the numbers are different, since the data is filtered. It should be noted that 23 duels are not in this local process model. This is because these 23 duels, have another duel in the same sequence. These 23 first duels are won, but then later in the same sequence the possession is lost in a next duel. There are a total of 284 cases, where player A passed 6 or less events before possession loss. Only 110 ended in a duel. So in the other 174 cases, the possession is lost in a different way.

Figure 6.35: A Petri net showing how often a pass is followed by a duel resulting in possession loss. The data is filtered on player B (a) and player C (b).

Figure 6.35 depicts how often a duel resulting in possession loss is preceded by a pass when the data is filtered to only contain the actions where player B (a) or player C (b) is involved in. The total number of cases is 195 and 188, respectively. So a bigger percentage of the sequences starting with a pass from player C and ending in possession the loss contain a duel. This can be explained by the position of player C, as goalkeeper he often passes far, giving the teammate and opposition time to prepare to receive the ball, ending up in a duel.

The number of cases are for both players lower than for player A this can be explained by the fact player A is a midfielder and player B and player C are an attacker and the keeper. A midfielder is more involved in the game, explaining the higher number of cases.

When player A is compared to another midfielder, player D, the Petri net results in figure 6.36 Figure 6.36 depicts the Petri net when the data is filtered to start with player D. For player D there are 62 cases where player D passed before possession loss was lost.

In a similar fashion the other Petri nets can also be analysed when a payer is involved in some action. This then can be compared to other players, for example to analyse which players create the possession loss and which players prevent possession loss.
Figure 6.36: A Petri net showing how often a pass is followed by a duel resulting in possession loss. But the data is filtered to make sure Maher passes before the possession is lost in maximum 6 actions.

### 6.5.1 Remarks

Although the local process models can give useful insights, when a subset based on a player is taken not all players can be compared, since the position of the player is not accounted for; some positions will naturally result in more or less possession loss. E.g. an attacker will have more possession loss.

Furthermore, these local process models themselves do not provide a lot of information, however, they do provide useful insights and point to interesting features for further research.

### 6.6 Striking Features and Hypotheses

While working at the project some possible causalities were noticed. This section introduces and discusses these possible causalities. These causalities resulted in hypotheses. These hypotheses will be introduced in this section as well.

When the matches are analysed as a whole, it seemed there are a lot of similar movements in the matches. This could be because there is some kind of biorhythm, resulting in the players reacting in a similar fashion.

**Hypothesis 1** The matches follow a certain rhythm because the actions and reactions of the players are often similar.

Hypothesis 1 proposes the matches all follow some kind of rhythm. This is not the case as shown in figure 6.37, showing all the matches played in a season. As can be seen there is no rhythm since every match behaves very differently. This results in the rejection of hypothesis 1.

When analysing the area during a match there seems to be a causality between a sudden drop in area and when a goal is scored. This results in hypothesis 2.

**Hypothesis 2** A goal is preceded by a drop of the area.

When the video is reviewed, it becomes clear the causality is actually the other way around. Because a goal is scored, the area drops. Because the players go to one point to celebrate. So hypothesis 2 is rejected as well.

If the poorly handled moments are defined as poorly handled when the opponents reach the 16m box instead. The area and centroid are as in figure 6.38. This leads to hypothesis 3.

**Hypothesis 3** When the opponent reaches the 16m box the area is bigger than when they do not reach the box.

As can be seen in figure 6.38. This is indeed true. Since the centroid still moves in a similar way as poorly handled moments, but the area increases, this indicates the players are further apart and therefore are putting less pressure on the opponent. This is because some attackers stay around the centre line and do not go back the full field.

This makes hypothesis 3 accepted.

When analysing individual matches, it is noted the Y-coordinate of the centroids, actually are not always staying close to 0. This results in hypothesis 4.
CHAPTER 6. RESULTS AND EVALUATION

Figure 6.37: Both images display all matches played versus the top opponents defined in section 6.2.4. The area covered (a) and the centroid (b) during the match.

Figure 6.38: The covered area (a) and the centroid of the convex hull (b) for all matches
CHAPTER 6. RESULTS AND EVALUATION

Hypothesis 4 Some opponents have a preference to play over one or the other wing.

Figure 6.39 shows the Y-centroid of individual matches. In (a) the Y-coordinate is clearly negative and in (b) the Y-coordinate is clearly positive. Remember the image only shows what happens after possession loss. So after possession loss the movement is negative in figure 6.39(a) which means the club is attacked on their right flank and the opponent is attacking over their left wing. Furthermore, note both (a) and (b) are played against the same opponent. So this indicates teams sometimes play more over left, and sometimes more over right. So although an opponent can have a preference, this preference actually still can change every match. So hypothesis 4 is rejected. Since the same opponent can have a different preference during different matches when analysing what happens with the well-handled moments, maybe the players did move back after 2.5s but just all the huge forward movements are dropped, since the possession is regained, leaving only the cases with less forward movement. This results in hypothesis 5

Hypothesis 5 When looking at the well-handled moments, the moment the possession is regained, the forward movement was big.

Figure 6.40 depicts the change of the X-coordinate of the centroid when the critical moment is handled. The mean of the dots is 0.49. This indicates the ball is retrieved mostly half a meter before the X-coordinate where the possession was lost. This is less significant as expected. Which does indicate the players are moving back after some time. When the ball is not retrieved yet. Resulting in rejection of hypothesis 5
CHAPTER 6. RESULTS AND EVALUATION

Figure 6.40: The change of the convex hull centroid when the critical moment is handled well.

6.7 Conclusions

The identification of the moments is quite good. The used method does not identify all moments, but it does only identify critical moments. Furthermore, the ease of changing the criteria makes it convenient to use this method, and still do the analysis in the same way.

The behaviour of the players is as expected. The area decreases after possession of the ball is lost. The players also move forward after the possession loss. So the pressure on the opponent increases after the club loses possession. However after two to three seconds the pressure decreases again, the area decreases slower and the centroid is moving back again. Furthermore, when analysing all matches during the season, the Y-coordinate of the centroid stays around 0. This indicates the club does not prefer one wing over the other during the whole season.

When only the well-handled moments, the moments when possession is retrieved quickly, are analysed there are not many differences between the well-handled moments and all moments. This indicates most moments are handled well. Both the area and the centroid behave in a similar way.

The poorly handled moments however, are different. The area decreases less. In a number of cases the area even increases instead of decreases. This is caused by the attackers waiting at the half line, when the opponent is attacking in the 16m.

When a single match is analysed, there are noticeable differences between different matches. Even between different matches against the same opponent. In combination with the analysis over all matches, a single match can be compared to the other matches and it can be analysed why this match was good and where improvement is needed.

The analysis can also be done for the whole match. When the area during the match is analysed, the covered area can be compared to the covered area of the opponent directly. The interesting moments are when the area of the club and the opponent diverge. Since most actions are made then.

The centroids stay close together most of the time. The moments when the centroid of the club moves away from the centroid of the opponent call for a more in depth research. Furthermore, the analysis of the whole match reveals the pressure put on the opponent is less sudden later in the match. Which is probably caused by the players getting fatigued during the match.

Classification of snapshots during the match indicating what team will get or keep possession of the ball proved to be too hard. However in some cases it proved to be possible to classify the team which was in possession of the ball. Furthermore, it was surprising the second half of the match is very different, compared to the first half.

The local process models can be used to find unexpected sequences, furthermore it can be used to
compare the sequences when the data set is filtered on a particular player.
Chapter 7

Discussion

This thesis presents multiple ways the data provided by the club can be useful to extract insights from the data. Football is starting to use data collected during the match, although a good start is made, football is still a lot behind compared to other sports like basketball, where data is already used more elaborately. In football the analysis of data has started, although the connection to the coaches and eventually the players is still lacking. The models and methods in this thesis are aimed to be usable by the coaches, and do not require a lot of technical knowledge to be used.

There is still a need for more research in this field. This thesis aims to clarify what happens after the possession of the ball is lost. The same approach can be used to analyse different moments. E.g. what happens before the possession of the ball is gained or what happens after a free kick is taken on the own half.

The introduction poses three research questions. After the research done in this thesis these questions can be answered.

Is the club putting pressure on the opponents after possession loss?
The club is putting pressure on the opponents after possession loss. This is indicated by the decreasing covered area and the forward movement of the centroid. Although pressure is put on the opponent, it is striking to note the forward movement is pushed back to the original position after 5 seconds. The area however does stay smaller, indicating a backwards motion of the whole team.

Is the pressure put on the opponent different after possession loss under certain circumstances?
The pressure put on the opponent after possession loss changes under different circumstances. When the possession loss is handled poorly, defined by a goal attempt of the opponent, the forward movement is less, and quickly changes to pulling back. Also the area decreases less quickly. Well handled moments however behave very similar to the average of all moments when possession is lost.

Furthermore, when a single match is analysed, there can be observed different behaviours. One match the left wing is preferred while another match the right wing is preferred.

Are there recurring patterns which can be extracted by local processes mining?
There are recurring patterns in the data, these patterns themselves are not very interesting. It does become interesting when it is tested on different players, and these players are compared with each other. This indicates which is the better player.

7.1 Future Improvements

Although this thesis provides useful information about what happens in the matches of the club, there still could be made improvements to increase the number of insights from the data.
CHAPTER 7. DISCUSSION

Firstly when there is more data collected, there can be retrieved more insights and the confidence of the insights will increase. Although it should be noted too old data cannot be used for comparison anymore, the entire team consists of different players.

The insights could be even more useful when data from other matches the club did not participate in is also analysed. This way all matches in the season could be used not only the matches the club played themselves. This will increase the number of matches and therefore the accuracy of the insights. Furthermore the teams of the other clubs then can be analysed, and eventually the play style can be changed when playing versus different opponents. It will also be interesting to see whether the opponent plays differently versus the club compared to their normal play style. Or whether the club is considered a normal opponent.

There can also be done further experimenting with the critical moments. Totally different criteria could be used as a basis of the analysis. It could for example be analysed what leads to a yellow or even a red card. This can also be done on moments where the covered area of the club and the covered area of the opponent diverge. These are moments when often something interesting happens, and not all players are close to their direct opponent anymore.

The poorly handled moments could be defined by players having a higher probability of scoring from the position, instead of only the goal attempt. This way the research in this thesis could be merged with the probability models.

Also the statistical indicators can be varied. In this thesis the mean of the moments is used, the insights could change when the median, minimum or maximum is analysed. Also there still could be experimented with different kinds of, more complex, statistical indicators.

A Voronoi diagram could be used to define the covered area of a team. This diagram could be used to define the number of defenders as well.

All players currently have the same weight when the area and centroids are defined. Experimenting with the weight of the players could lead to interesting findings. The attackers or the centers could have a higher weight and therefore influence the centroid more.

Finally the classification indicated there are significant changes between the first and second half. So the analysis could be done again, but separating both halves. This could lead to even more insights.

7.2 Use by the club

This section will elaborate on how the techniques explored in this thesis can be used by the club in future seasons. And what should be noted when using these techniques in the future.

Firstly, the different analysis techniques explained can be used to compare different matches. This can be used to evaluate lost matches and won matches. If a match is lost but the handling of possession loss was good, the focus can shift to other areas. Furthermore the analysis can be used to compare the start of the season to the end of the season. To see if the players play different when the season is further.

In this thesis the local process models are used to see how often a player passes resulting in possession loss. This is a comparison the club could still be making in the future. In order to see how the situation after a pass is, instead of how good the pass is. This skill of the player can be used to compare players playing in the first team. However it can also be used during scouting and training of youth players.
Bibliography


Appendix A

Area Covered During All Matches

This appendix shows why it is infeasible to analyse all matches at the same time. Because this will result in images as in figure A.1. These images are too crowded to get the information out of them.

Figure A.1: When all information is put in one figure this figure becomes unreadable. This figure displays the covered area during the match for all matches.