Tristan Kristoffer Cook
Eivind Keil
Inge Dovre Sætherø

Modelling Off-Ball Decision Making in Swedish Top Division Football

Master's thesis in Industrial Economics and Technology Management
Supervisor: Magnus Stålhane, Lars Magnus Hvattum
July 2020

Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Economics and Management
Dept. of Industrial Economics and Technology Management

NTNU
Kunnskap for en bedre verden

NTNU
Norwegian University of
Science and Technology

Modelling Off-Ball Decision Making in Swedish Top Division Football

Tristan Kristoffer Cook
Eivind Keil
Inge Dovre Sætherø

DEPARTMENT OF INDUSTRIAL ECONOMICS AND TECHNOLOGY MANAGEMENT
NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Supervisors:
Professor Magnus Stålhane
Professor Lars Magnus Hvattum

Preface

This master's thesis was written for the Department of Industrial Economics and Technology Management at the Norwegian University of Science and Technology, finalizing the authors' Master of Science. It was written during the spring of 2020 and is to some extent a continuation of the authors' project thesis from the autumn of 2019. All authors are specializing in financial engineering, with two having a background from computer science and artificial intelligence, and one from energy and environmental engineering.

The work done in the thesis concerns the use of analytical methods on tracking data from association football. After being impressed by the ongoing work with tracking data at Hammarby IF, the authors reached out to the Swedish top division team to propose a cooperation, and work on developing their existing models. Hence, this thesis is the result of a cooperative initiative between the authors and Hammarby IF.

Abstract

The published analytical research on association football is growing, but several interesting areas of the game are still to be explored by academic research. This master's thesis seeks to expand on the existing literature by analyzing off-ball decision making by players and teams from the 2019 season of the Swedish top division, Allsvenskan.

The data used is positional tracking data provided by Signality, tracking all 22 players, the ball and the referee, while also providing some descriptions of on-ball events. Three existing models used for calculating pitch control, pitch impact and a combination of the two, provided by the Swedish professional club Hammarby IF, serve as three alternative metrics to evaluate the success of off-ball movement. Players and teams are evaluated in relation to how well they perform compared to an optimal performance identified for each metric. To identify and account for situational dependencies, two types of prediction models, generalized additive models and feed-forward neural networks, are developed to analyse performance and behaviour, and to create a situation adjusted rating in relation to the metrics. The top ten performers on both actual and situation adjusted ratings are presented. The ratings are also compared with existing ratings for off-ball movement, provided by a professional scouting network used in the video game Football Manager 2020. Results show a moderate positive correlation between some of the ratings presented in this thesis and the ratings from Football Manager 2020.

Furthermore, role specific differences in positional strategies are investigated. Findings suggest that differences exist in positional priorities between different player roles, with attacking players seeming to focus more on the included metrics than defensive players. An analysis of the relations between included metrics and goals scored is also conducted, with the most notable finding being that the highest scoring teams seem to divide positional responsibility more than other teams.

A generalized additive model and a feed-forward neural network are also developed to predict player positions over a one second time interval. Findings show that the neural network is better at describing the dynamics behind player movement and decision making than the generalized additive model and other alternative benchmarks. Results also show that player movement is harder to predict for players deviating a lot from initial direction and velocity.

Sammendrag

Den publiserte analytiske forskningen innen fotball er voksende, men fortsatt finnes flere interessante aspekter ved idretten som ikke er belyst gjennom akademisk forskning. Denne masteroppgaven søker å supplementere eksisterende litteraturen ved å analysere hvordan spillere og lag fra øverste divisjon i svensk fotball, Allsvenskan, i sesongen 2019 foretar beslutninger uten ball.

Denne oppgaven bruker posisjonsdata levert av Signality. Dataen inneholder bevegelsene til alle 22 spillere, ballen og dommeren, samt noen beskrivelser av hendelser som omhandler ballen. Tre eksisterende modeller, levert av den svenske klubben Hammarby IF, for å beregne kontroll av rom på banen, posisjonsinnvirkning og en kombinasjon av disse, brukes for å definere suksess tilknyttet bevegelse uten ball. Spillere og lag evalueres i forhold til hvordan de presterer sammenlignet med optimale verdier beregnet for de nevnte modellene. To typer prediksjonsmodeller, generaliserte additive modeller og feed-forward nevrale nettverk utvikles for å redegjøre og justere for situasjonsavhengige faktorer når prestasjoner og adferd uten ball evalueres. En rangering av spillerne basert på prestasjon i både faktiske og situasjonsjusterte observasjoner foretas hvor de ti beste spillerne ifølge rangeringene blir presentert. Rangeringene blir sammenlignet med eksisterende rangeringer av relevante attributter, satt av dataspillet Football Manager 2020 sitt profesjonelle speidernettverk. Resultatene viser en moderat positiv korrelasjon mellom rangeringene fra denne oppgaven og rangeringene fra Football Manager 2020.

Videre utforskes rollespesifikke forskjeller i posisjonelle strategier. Resultatene i denne seksjonen antyder at angrepspillere fokuserer mer på å oppnå høye verdier for kontroll og innvirkning enn forsvarspillere. Sammenhengen mellom antall scorede mål og prestasjoner uten ball blir analysert, med resultater som tydet at lag som scorer mange mål fordeler posisjonelt ansvar mer enn lag som scorer færre mål.

En generalisert additiv modell og et feed-forward nevralt nett er utviklet for å forutsi spillerposisjoner over et tidsintervall på ett sekund. Det nevrale nettet viser seg å være bedre egnet til å beskrive dynamikken i spillerbevegelser og beslutningstaking enn den generaliserte additive modellen, samt et utvalg alternative målestokker introdusert for å evaluere modellene. Resultatene viser også at spillerbevegelse er vanskeligere å predikere for spillere som avviker mye fra initiell retning og hastighet.

Acknowledgements

First, we want to thank our supervisors, Professor Magnus Stålhane and Professor Lars Magnus Hvattum. They have been of great help in guiding us through the process of creating this thesis, giving excellent feedback and guidance.

Furthermore, we want to express our gratitude to the people at Hammarby IF, with special thanks to David Sumpter, Professor of Applied Mathematics at the University of Uppsala, and Francisco José Peralta Alguacil, Data Scientist at Hammarby IF. Their willingness to share their models and data have been crucial for this thesis.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

With 265 million active participants (FIFA, 2007) and with over half of the world's population watching the 2018 World Cup (FIFA, 2018), association football, in this thesis referred to as football, is considered to be the most popular sport in the world. This amount of participation and attention has created an economic landscape where small improvements in results can lead to significant financial gains for professional clubs. In a sport where the difference between one spot on the league table can have such considerable implications on revenue, it is not surprising that many clubs are looking into every opportunity that potentially can enhance performance. In recent years, the field of sports analytics has been on the rise, with an increasing number of clubs seeing the possibility of utilising analytical methods to both improve performance on the field and assist in the process of scouting and acquisitions of new players. Technological advancements have made large amounts of data available for analysis, with comprehensive data sources on both game events and player movement available to those willing to invest in the right equipment and technology. Using these tools, teams can now develop new perceptions on the game, showcase elements not visible to the naked eye, and create models to describe a wide range of performance factors.

1.1 Background and Motivation

A common way of analysing and evaluating football players is to collect aggregated statistics based on goals, shots, passes and tackles. The common denominator for these statistics is that they all measure individual players' on-ball involvements. However, football is a complex game of collective movement, where off-ball involvements also can have an impact. With 22 players and only one ball on the pitch, players spend far more time making off-ball decisions than they do on-ball. When not in possession of the ball, players continuously have to adjust their position on the pitch to either create or defend against threats. The legendary Dutch player and coach Johan Cruyff once famously said: "When you play a match, it is statistically proven that players actually have the ball 3 minutes on average (...) So, the most important thing is: what do you do during those 87 minutes when you do not have the ball. That is what determines whether you're a good player or not." Evaluating players solely on their on-ball actions will, therefore, only capture parts of the players quality and contribution to the team. When not in possession of the ball, players continuously have to adjust their position on the pitch to either create or defend against threats.

With recent advances in player tracking technology, a new and rich data source, often referred to as tracking data, has become available. Tracking players' every move, this data opens up the possibility of evaluating off-ball movements as the available data now contains information on these actions.

An important aspect of evaluating off-ball actions is to establish measures for what constitutes off-ball success. When on-ball actions are evaluated, these measures are usually quite intuitive; a goal is better than a missed attempt, and passing the ball to a teammate is better than passing it to an opponent. For off-ball actions, defining success is far more challenging and complicated, requiring the development of measures of success, using complex models. This thesis uses models developed by the Swedish top division club Hammarby IF to create measures of success for off-ball movements made by players during the 2019 season of Allsvenskan. Currently, these models are used by Hammarby IF to coach players and analyse players' decisions' and movement in individual situations. These models are modified to determine the amount of the maximum and minimum obtainable control and impact a player achieves. An objective of this thesis is, therefore, to see if these models can be applied to create a framework for analysing a large number of off-ball situations, evaluating players movement and decision making abilities.

To aid in the work of comparing and evaluating positional decisions made in different situations, models are developed to predict the outcome, using a number of features designed to describe the situation accurately. The two modeling techniques used in this thesis are generalised additive models and feed-forward neural networks, the first being a statistical model, and the latter coming from the field of artificial intelligence. Both models are introduced as they are known to deal with non-linear relations between variables. The results from the preceding project thesis (Cook et al., 2019) showed that many relationships in football are non-linear. As these models have different advantages and disadvantages, comparing the results from the different models can be used to infer which is most suited to analyse off-ball actions in football. With the generalised additive model being a statistical model, the contribution of individual features is simpler to interpret and analyse, while deep learning-based models have great abilities to deal with complex relations between variables. Further, the models are modified to predicting player movement over short time periods. Together, these approaches aim to shed light on aspects of how players make positional decisions that are hard to decipher with the naked eye and the predictability of player movement.

1.2 Research Questions

The discussion in the previous section forms the basis for the four research questions answered in this thesis. These research questions are formulated as followed, with an explanation of how this thesis seeks to answer these questions:

Research Question 1: *Can available state of the art off-ball metrics be used to evaluate off-ball movement and decision making in elite-level football?*

Metrics for defining off-ball success will be applied to a large set of observations, with the results compared to existing ratings for off-ball aspects provided by the professional scouting network of Football Manager 2020.

Research Question 2: *Are there individual differences in off-ball decision making, and do decision making vary for different player roles?*

An evaluation of individual players is presented based on different off-ball success metrics, both for their obtained score and situation adjusted score. Further, different player roles will be compared on how their score, identifying trends for different player roles.

Research Question 3: *Are players on the best attacking teams in Allsvenskan making different off-ball positional decisions than other teams?*

This question seeks to further link success on off-ball metrics up to a central objective of football; goals scored. The 16 teams from the 2019 Allsvenskan are divided into four groups, separated by how many goals they scored during the season, and then compared by their performance on off-ball metrics. Visualisation of location-specific performance will aid in interpreting the results.

Research Question 4: *How well can player movement be predicted over a short time interval, and what types of models are best suited to model this movement?*

A generalised additive model and a feed-forward neural network are developed to predict player movement one second into the future. The models are then compared on existing benchmarks and evaluated on a number of validation metrics.

1.3 Thesis Structure

First, in Chapter 2 related work is presented to give an introduction to the existing work and progress of sports analytics. Chapter 3 then introduces the theoretical foundation and concepts relevant to the analyses presented in the later chapters. This chapter is followed by Chapter 4 presenting the methodology used for defining off-ball success. The data used in this thesis will then be introduced in Chapter 5, followed by the model set-up and validation in Chapter 6. Applications of the models are then presented in Chapter 7, before extending the models to predict player movement in Chapter 8. In Chapter 9, Answers to the research questions presented in this chapter are given, and finally, Chapter 10 concludes this thesis and offers suggestions for future work.

Chapter 2

Literature Review

2.1 Introduction

Sports analytics is the use of scientific methods and analysis in sports, and the practice dates back over a century (Memmert and Raabe, 2018). One of the earliest examples of sports analytics is Hugh Fullerton's 1912 paper, where he divided a baseball field into different zones to analyse the success probabilities of balls hit into these zones. Almost 40 years later, the first systematical match analyses in football were undertaken by former Royal Air Force officer Charles Reep. In the 1950 Reep developed a notation system to analyse football matches. Around the same time, the field of operations research also made its entrance into the world of sports. Arguing that several American team sports, like baseball, American football and basketball, share similarities with warfare, Mottley (1954) proposes that the same operations research methods used in warfare can also be used to gain tactical advantages in these sports. In the late 1980s, A.H.Ali drew attacking patterns from 18 games in the Scottish top division by hand. By using an overlying grid to determine the players' movements in x- and y- coordinates, this effort is often cited as the first analysis of tracking player movement in football (Memmert and Raabe, 2018). With the development of new technology and statistical methods, sports analytics have become an increasingly important part of top-level sports with research in the field becoming increasingly advanced.

This literature review focuses on modern academic research using both event-based and spatio-temporal tracking data. Following improvements in both methods and available data sources, most of the research relevant to this thesis is relatively modern, leading to this chapter focusing on newer research. As this thesis focuses on the sport of football, most of the research presented analyses football. Some research conducted on other team sports that share similarities to football is also included when the research is relevant to the subjects covered in this thesis. The first section of this literature review focuses on research using tracking data to analyse player movement. As off-ball movement is the main focus of this thesis, previous studies on player movement are highly relevant. Furthermore, the models presented and used in this thesis uses both event-based and spatio-temporal tracking data, and the following sections, therefore, covers research using both types of data separately or combined. The structure of this chapter is constructed as to give insight into the research previously done in the areas of the game the different models used in this thesis covers. First, some research on player movement using tracking data is presented. In Section 2.3, an introduction to research conducted on modelling passing is presented, showing some of the different ways this has been approached by researchers. This section is followed by a presentation on some relevant research conducted on how players and teams control available space. As with modelling passing, there are different approaches to modelling space control and some of them are presented in Section 2.4. One of the most well known and commonly used metrics in football are the expected goals metric discussed in Section 2.5. Besides from being an interesting metric on its own, expected goals metrics are a widely used component of more extensive player evaluation models. These models seek to quantify a players impact or contribution to the team, and some of these models are discussed in Section 2.6 of this chapter.

2.2 Player Movement

Research on player movement have evolved along with the technology used to track players. This section covers research done on player movement both focusing on individual players and teams.

One way to utilise tracking data to study player movement is to aggregate positional data to analyse properties like distance covered, speed and intensity among players. Barros et al. (2007) used an automatic video tracking system to measure total distance covered, and distance covered with different intensities by 55 players from the Brazilian First Division. Relating the results to player roles the study showed that midfielders and wide defenders covered more distance than forwards who again covered more distance than central defenders. Players covered the most distances at walking or jogging speed, and the total distance dropped by 7% from the first to the second half. Also relating player roles to physical demands, Di Salvo et al. (2007) studied 300 top-class players during 20 games in the Spanish top division and 10 games in the Champions League. This study found no significant difference in the total distance covered between the two halves, and that midfielders followed by wide defenders and forwards covered the most distance while central defenders covered the shortest distance. To investigate whether high intensity movement was related to team success, Di Salvo et al. (2009) studied positional data from 563 players in the English Premier League. Teams where divided into three groups based on league position with one group consisting of the top five teams, one consisting of the middle ten teams and one group consisting of the bottom five teams. The findings suggested that teams finishing lower in the table covered more distance with higher intensity and in sprints than higher finishing teams. Another finding from this study was that midfielders and attackers seemed to cover the most distance during high intensity running and in sprints, and that central defenders covered the least distance in both of the same categories.

Using aggregated tracking data to study player movement and physical demands like intensity and distances covered, can be a valuable resource for for coaches and athletic trainers when developing training regimes for players. But these types of analyses lacks deeper insight into how players interact and cooperate during games. To study how players move collectively as a team, Moura et al. (2012) and Moura et al. (2013) used tracking data to create two metrics based on player positions. Total space covered by the team, calculated as the convex hull covered by all players, and distance between the players, calculated by the Frobenius norm of individual distances between players. (Moura et al., 2012) analysed the difference in these properties between attack and defense, finding that teams covered a smaller area and had smaller distances between players when defending than attacking. Also relating these properties to attacking and defending success, the results showed that teams were less compact in the attacks leading to a shot on goal than when they suffered tackles. In defense the study found that teams were more compact when they made tackles than when they suffered shots on goal. Evaluating the frequency of time series of the same properties, Moura et al. (2013) measured how fast teams were able to increase or decrease their compactness during play. The study concluded that teams used longer time to increase or decrease their compactness in the second half as the frequency decreased from the first to the second half.

An important part of a team's attacking structure is what types of runs and combinations of simultaneous runs are made by players. Miller and Bornn (2017) uses Bézier curves and a machine learning approach to cluster movement in basketball, and create topics of simultaneous movement to group possessions with a similar structure. Gregory (2019) used the same approach of Bézier curves to create a framework of clustering runs in football. This framework could give insight into what types of runs players make, and what combinations of these runs are done simultaneously.

2.3 Pass Probability

Passes are one of the most common and important events in football, and a central skill for every player. From playmakers searching for the key passes to unlock the opponent's defence to defenders and, increasingly important in the modern game, goalkeepers initiating attacks through forward passing, all aspects of play is significantly driven by passing the ball. This makes passing one of the most sought after skill for players and much research has therefore focused on this aspect of the game. As passes are discrete events, research based only on event-based data is possible as this data often describes many features of the pass. But as tracking data adds many important features, research based on both event and tracking data has the possibility to add additional insight.

One key aspect of evaluating a pass, whether it was completed, failed or never tried, is by the probability for success. If the probability of completing a certain pass is calculated, players and teams can be evaluated on several metrics related to that probability. Was the decision to try a certain pass a good decision, was there a better option available to the player at the time of the pass, how does a player or team's success rate compare to the calculated probabilities? These are some of the questions researchers have tried to answer through developing pass probability models. Generally, pass probability models can be separated by two different approaches, data driven and physics-driven. Data-driven models seeks to fit models to available passing data, while physics-based models uses equations of motion and probability distributions to calculate possible ball and player trajectories and interception probabilities.

Whether they use only event data or a combination of event and tracking data, data-driven models seek to fit models to the available data. Some examples of this approach is the work done by Szczepański and McHale (2016); Håland and Wiig (2018); Tovar et al. (2017) all using event-based data. Szczepański and McHale (2016) used data from the English Premier League and Håland and Wiig (2018) from the Norwegian top division, to create generalised additive mixed models (GAMM) predicting probabilities for successful passes. Using features such as the coordinates of the pass, game time and what part of the body was used for the pass, players were then evaluated on how well they performed compared to the models created. A GAMM was also used by Tovar et al. (2017) to create a similar pass probability model, using event data from Colombian league and Spanish top division (LaLiga) the purpose of this model was to use passing ability as a proxy for performance in the Colombian league and predict future performance in the Spanish LaLiga after an eventual transfer. Better predicting how players perform in new leagues can be a useful tool for clubs in evaluating potential transfer targets. In their study of how players perform under different levels of mental pressure Bransen et al. (2019) used a Gradient boosted tree model to compute the probability of different actions, among them passes, being successful. This was done to evaluate the expected contribution of choosing a certain action, and thereby evaluating the player on both the choice and execution of the action.

Implementing a logistic regressor using both event and tracking based data from the English Premier League, Power et al. (2017) quantified the quality of a pass by its risk and reward. The risk of a pass was quantified as the probability of it reaching a teammate, and pass reward as the probability that a successful pass leads to a shot within the next 10 seconds. This was used to analyse risk and reward for teams during matches and ranking players based on two metric called Passing Plus Minus (PPM) and Difficult Pass completion (DP%). PPM is a metric that weighs the completion percentage against how risky the pass was, and DP% is the percentage of high risk passes, passes in the 75th percentile of riskiest passes, a player completes. McHale and Relton (2018) implemented tracking data from the English Premier League to create a GAMM estimating the probability of a pass being successful, and the difficulty of the pass.

Another data-driven approach is to use machine learning techniques to create passing models fitted to large amounts of data. Fernández et al. (2019) included the probabilities of an action being a pass and the outcome of the pass, successful pass, or a turnover, in their Expected possession value model for football. While the outcome probability was estimated using logistic regression the action likelihood was estimated using a convolutional neural network. The data used was optical tracking data from the 2012-2013 season of the English premier league and FC Barcelona's matches during the 2017-2018 and 2018-2019 seasons of the Spanish top division LaLiga. Passing is also an important type of action in other sports than football, and one sport where passing is an integral part, considered to be the most valuable type of action (Eager and Chahrouri, 2020), is American football. Burke (2019) used a feed-forward artificial neural network on over 45 thousand pass attempts from the 2016 and 2017 seasons of the National Football League to model passing situations in American football. Included in the feature vector fed to the network the possible receivers were represented as vectors that included their position, velocity, distance and angle from the quarterback (passer), their shoulder orientation and their playing roles such as wide receiver, tight end or running back. Along with the possible receivers, the position, velocity and shoulder orientation of the two closest defenders to each receiver plus a vector of metadata such as down and distance, yard line and if the quarterback was under pressure was included in the full feature vector. The network then produced three types of output, the probability that the quarterback chooses to target each receiver, the probability that the pass would be complete, incomplete or an interception and the expected yards gained from choosing to target the different receivers. This combined results in a model that can evaluate both the choices made by the quarterback, by comparing the alternative receivers and the execution of the pass by comparing the predicted and actual outcome.

Instead of fitting models to data Gudmundsson and Wolle (2014) used equations of motion to compute a surface of possible ball interception points for all players and therefore the passable area for the ball carrier in a specific situation. Spearman et al. (2017) also used equations of motion to compute reachable areas for players along with the possible trajectory of the ball to compute surfaces of possible interceptions areas. Combining the interception surface with a probability distribution modelling the likelihood of a player being able to control the ball given the player and the ball trajectory intercepts. This model is also the inspiration behind the pass probability model in Peralta Alguacil (2019). Many of the same equations and principals are used, with the main difference being that Peralta Alguacil (2019) models player motion differently to save computational resources. Details and further explanation of this model is provided in Section 3.1.1.

2.4 Pitch Control

With the increased access to, and use of, player tracking data, the notion of space control has become a common performance metric. Space control seeks to assess how much, or to what extent, areas on the pitch are controlled by certain players. One common approach is to partition the pitch into Voronoi cells, as suggested by Taki et al. (1996). Voronoi cells partition a plane into cells where each cell represents the area that is closest to what is called the seed of the cell. In the context of team sports like football, this means that the pitch is divided into cells that represent the area controlled by a certain player since the player is closer, in Euclidean distance, to any point in that cell than any other player. A variant of the Voronoi cells approach is to include the initial velocity of players in order to partition the pitch so that the area of the cells is all the point a player is able to reach, given all initial velocities, faster than any other player.

Space control defined by Voronoi cells has been used in many different types of analyses both in football and in other sports. Because of the difficulties of attaining precise positional data from real-life games, Kim (2004) used Voronoi cells to analyse a game played using the video game FIFA Soccer 2003 by EA Sports. By calculating the area of the Voronoi cells of the virtual football players, the ratio of the total area of the two teams was used to quantify each teams dominance over the other. Rein et al. (2016); Chawla et al. (2017) used changes in space control during passing events to evaluate the quality and effectiveness of the passes. Analysing dominant regions during both successful and unsuccessful offensive performances, Ueda et al. (2014) found that narrow dominant regions were linked to successful offensive performances. Perl and Memmert (2017); Memmert and Rein (2018) used the amount of space controlled by a team, calculated using Voronoi cells, as a Key Performance Indicator for offensive success. The assumption behind using space control as a performance metric is that increased space control leads to more success. Rein et al. (2017) used a linear mixed model to analyse the effects of space control gained, using Voronoi cells, during passing events on goals scored, shots made and match success. The results showed that space control gains were significantly related to both goals scored and match result, finding that increased space control during passing events led to both more goals scored and a more positive match results. Memmert et al. (2019) used space control as a measure in the evaluation of two common but different formations 3-5-2 and 4-2-3-1. Modelling space control using Voronoi cells and assessing space control gain as the change in space control during passing events, the study found no significant difference in this measure for the two formations.

Fernandez and Bornn (2018) used a slightly different approach to measuring space control. Based on the position and velocity of all players along with the position of the ball, each player has a degree of influence, from 0 to 1, over a position on the field. This makes it possible for more than one player to have a degree of influence on a position of the field, which differs from Voronoi cells where control is discrete. Adding a machine learning approach to quantifying space value Fernandez and Bornn (2018) then created metrics for space occupation gain and space generation gain to measure the quality of player movement. The framework was applied to a Spanish first division match between F.C. Barcelona and Villareal F.C in January 2017 to evaluate the F.C. Barcelona players on how much space value they generated for themselves and their teammates. One of the notable findings was how Lionel Messi was able to generate a lot of valuable space for himself while moving at low velocities, while it also showed that Neymar jr. and Luis Suarez often created valuable space for their teammates and especially for Lionel Messi.

2.5 Expected Goals

Another challenge in analysing the game of football is to quantify the value of different actions. In other team sports like American football and baseball, many actions can be easily quantified by their direct results. In baseball, this could be if the pitcher threw a strike or a ball, or the amount of yard gained by the runner in American football. Some sports like handball and basketball share many similarities with football in this regard, but they have the advantage of a much higher number of goals, or points, scored during a game. This is an advantage in the analytical sense since it means more actions are directly quantifiable. Considering the 2018/2019 season of five major European domestic leagues, French League 1, Spanish La Liga, Italian Serie A, English Premier League and German Bundesliga, the average goals per game in the five leagues ranges from 2.56 to 3.18 (Krishna, 2019). With only two or three events per 90 minutes, an analysis only considering goals and assists will, therefore, be lacking insight into much of the complexities of the game. One way to evaluate team performance beyond just goals scored is to use the metric commonly known as expected goals or xG. The idea behind xG is that goals are relatively rare events that contain a lot of uncertainty. So in a single game or even over the course of a season, goals scored and conceded may not represent the true quality of a team. Shots are more common than goals, and an xG metric seeks to quantify the probability that a certain shot ends in a goal, in other words, the expected amount of goals the shot should result in.

In an article written for Optasports, Sam Green used xG as a metric to analyse players from the 2011-2012 season of the English Premier League. Based on Opta event data, the player's total goals and goals per shot were compared with the results from the xG metric created. Rathke (2017) examined shots from the 2012-2013 seasons of the German Bundesliga and English Premier League using an xG model based on dividing the field into zones based on distance and angle from the goal. Both teams and individual players were assessed on actual versus expected goals finding correlations between efficiency and final placings in the league, with top teams more efficient than lower placed teams. In their analysis of Leicester City's unexpected English Premier league winning 2015-2016 season with findings showing that Leicester City's defensive efficiency was part of what set them apart from other teams. Using tracking data from an elite team's home matches from the 2011-2012 to 2014-2015 seasons, Schulze et al. (2018) analysed the position of defenders on the outcome of shots. The findings suggested that shots from tight angles and shots close to the goal were affected by the defender's positions.

Yam (2019) used a Post-shot xG model to evaluate the shot-stopping qualities of goalkeepers during the 2017-2018 season of the English Premier League. Post-shot xG differs from Pre-shot xG, or what is simply called xG in this thesis, in that Post-shot xG only considers shots on target and not blocked and missed shots. The reasoning behind using Post-shot xG when evaluating goalkeepers is that shots missed or blocked by a defender leads to a positive outcome for a goalkeeper without the goalkeeper's involvement and Pre-shot xG could, therefore, bias the sample. The data used consisted of event-based data that also included the coordinates of outfield players and goalkeepers at the time of the event. An extreme gradient boosting model was used to estimate the Post-shot xG model using features both defining the shot characteristics and the positions of defenders. Comparing the post-shot xG model with actual outcomes of the shots, the study found Manchester United's David de Gea to be the best shot-stopper in the league while West Ham United's Joe Hart and Liverpool's Simon Mignolet performed the worst.

2.6 Player Evaluations

In the modern game, players are bought and sold for increasingly high transfer fees, and the clubs that can find and acquire undervalued players can achieve large economic and competitive gains. However, the complexities of the game make evaluating individual players difficult and methods to better evaluate players are an important research question in the field of football research. Analysing games from the 2015/2016 and 2016/2017 seasons of the Italian top division, Serie A, Pappalardo et al. (2017) investigated which features, some derived from event data and some contextual, affected how three major Italian sports newspapers rated player performances. The results showed that the journalists working for the newspapers tended to only focus on a small number of features when assigning their ratings and that most of these features were contextual features. Humans, therefore, seem to put a lot of importance on contextual features like the expected pre-game result and the goal difference which arguably says less about a player's actual performance than event-based features like passes, shots and tackles.

One way to evaluate a players contribution to team performance is to construct a model that values a player's actions in how they affect the probability that an attack eventually leads to a goal scored. By modelling possessions as a chain of actions each action can be evaluated on its contribution to the success of the possession. Considering only the final link in the chain, the shot, this becomes an expected goals model, and moving one link backwards it becomes an expected assist (xA) model that evaluates the probability that a certain pass becomes a direct assist (Worville, 2017). Dividing the xG of the final shot, or highest xG in the possession chain, among all players involved in the possession a simple xG-chain model can be constructed (Lawrence, 2018). An xG-chain model rewards players for being involved in the build-up but lacks the sophistication of being able to quantify each action's individual contribution, as not all actions in a possession are equally valuable to the final outcome.

To evaluate all actions in a possession Singh (2019) developed an expected threat (xT) model that assigned a threat-value to pitch location and evaluated actions on the difference in threat-value between their start and end positions. Here, threat is defined as the probability of scoring from a shot or how easy it is to move the ball to an even more threatening position. Modelling the possession chain as a Markov game is another way of evaluating actions that do not directly lead to shots. Markov chains use the probabilities of transitioning from one state to another to model how likely the different outcomes of events are. In the context of football, this means that given a game state, described by features such as ball location, the likelihood of that possession ending in a goal or a turnover can be calculated. Rudd (2011) used Markov chains to evaluate actions and players from the 2010-2011 season of the English Premier league while Nørstebø et al. (2016); Haave and Høiland (2017) used this approach on data from the Norwegian top division.

Following a similar idea but different approach, Mackay (2017) used ridged logistic regression with a sliding window to model goal probabilities for possession chains. A generalised additive model to create an xG model that was included as a feature in the possession probability model, and the data used was event-based data from five seasons of the English

Premier League. Findings suggested that during the 2016-2017 season Manchester City's Kevin De Bruyne was the player that most increased his team's goal-scoring probabilities per 90 minutes played, followed by West Ham United's Dimitri Payet and Chelsea's Eden Hazard. Decroos et al. (2019) created a framework for valuing player actions called Valuing Actions by Estimating Probabilities, or VAEP, using event-based data. The idea behind VAEP is that the value of an action made by a player on team i is the change in the probability of team i scoring, offensive value, minus the change in probability that team i concedes defensive value. The model was estimated using the CatBoost algorithm, and findings suggested that the top-performing players during the 2017-2018 season of the English Premier league were Liverpool's Philippe Coutinho and Mohamed Salah. The idea of estimating the probability of scoring or conceding in a specific situation is also the basis for the expected possession value, EPV, framework developed by Fernández et al. (2019). Using a machine learning approach the expected value of possession is defined as a number in the range of $[-1, 1]$ expressing the probability of the outcome of the possession, where 1 indicates that the possession ends with a goal scored by the attacking team and -1 indicates it ending in the defending team scoring a goal. From this framework, players and teams can then be evaluated on how their actions in three main categories, passes, shots and ball drives, change the expected value of that possession.

Another player evaluation approach is the plus-minus metric. The idea behind plus-minus is to evaluate a player on how the team performs with the player compared to how it performs when the player is not playing. Does the player have a positive, plus, or negative, minus, impact on team performance. Plus-minus is a common player evaluation in sports like ice hockey and basketball, and in its most basic form, it measures the difference in points scored or conceded with and without the player on the court. Adjusted plus-minus (Sill, 2010), or APM, is an extension of the basic plus-minus metric that uses a regression model to account for teammates and opponents in the final plus-minus rating. With a lot fewer points/goals scored and less rotation in terms of team composition during a game, football seems to be less fitting for a plus-minus rating of players. Kharrat et al. (2020) developed two plus-minus models for football, expected goals plus-minus and expected points plus-minus. Expected goals plus-minus is an APM model evaluating players on their contribution to the xG achieved while expected points plus-minus is an APM model evaluating player contribution towards the number of league table points achieved. An APM model using goals as the basis for the plus-minus rating was developed by Sæbø and Hvattum (2019) with the intention of modelling the financial contributions from players.

Another type of player evaluation models follows the approach known as wins above replacement, commonly abbreviated as WAR. Known mostly for being used in several American sports such as basketball (Basketball-Reference, 2020), baseball (Baseball-Reference, 2020), and American Football (Eager and Chahrouri, 2020) the idea behind WAR shares many similarities with PM ratings. It seeks to quantify a players contribution to the team, but while PM quantifies micro properties like points, goals or xG, WAR seeks to quantify a players contribution to the macro property of winning the game. An important aspect of WAR models is the concept of a replacement player, someone a team could bring in as an immediate replacement for a player currently in their roster. In a paper presenting their WAR model for American football, Eager and Chahrouri (2020) uses data from the NFL to create a WAR model that can be used to compute the difference in value between players and different positions.

Technical and physical skill are undeniable parts of the overall quality of a player, but football is also a mental game. The quality of a player also depends on the quality of the decisions made with and without the ball. Bransen et al. (2019) created performance metrics that evaluate players on-ball performance during different states of mental pressure. Quantifying pressure as a combination of pre-game pressure, based on such things as league standings, form and whether it is a derby game, and in-game pressure, based on such things as score and time left in the game, players were evaluated on decisions, execution and total contribution.

Chapter 3

Theoretical Foundation

This chapter introduces concepts, models, and theory that are used later in this thesis. First, three models used by Hammarby IF for analysis and coaching is presented. These models are important parts of the analyses in this thesis and the models are therefore presented individually. Following the presentation of the three models, relevant theory on generalised additive models and artificial neural networks is presented. Finally, the concepts of centroids, compactness, and Bézier curves are presented. These concepts are relevant for feature engineering and data processing.

3.1 Models for Creating Off-Ball Metrics

This section covers three different models currently used by Hammarby IF that are used as parts of the further analyses conducted in this thesis. First, the pass probability model developed by Peralta Alguacil (2019) inspired by Spearman et al. (2017) is presented, followed by the pitch control model developed by Fernandez and Bornn (2018). Last, the pitch impact model developed by Twelve is presented. The focus of the upcoming section is to give an introduction, short explanation, and overview of the three models. For a more thorough and detailed explanation, the reader is encouraged to explore the referenced research papers.

3.1.1 Pass probability model and reachable area

Modelling the probability of a successful passing event in football is useful because it allows for a detailed analysis of the passer's skill level and their decision-making process. Several approaches to modeling the probabilities of successful passes exist. One approach is to train a regression model on a large data set of passes (Szczepański and McHale, 2016; Håland and Wiig, 2018; Tovar et al., 2017) and estimate the probabilities of successful passes. The model used in this thesis employs a different approach and is a physics-based model developed by Peralta Alguacil (2019) inspired by the work done by Spearman et al. (2017). The concept is based on modelling the motion of the ball as well as the reachable area for all players on the pitch. Given the positions of teammates and opponents, a surface of the probability for a pass being hit towards a certain area being intercepted and controlled by a teammate can be created.

Equation (3.1) is used to model the motion of the ball, with $\vec{r}$ being the ball's motion vector.

$$\ddot{\vec{r}} = \begin{cases} -\frac{1}{2m}\rho C_D A \dot{r}\dot{\vec{r}} & \text{if } t \leq \frac{2t_{max}}{3} \\ -\mu g \hat{r} & \text{if } t > \frac{2t_{max}}{3} \end{cases} \tag{3.1}$$

The equation is split into two parts with the assumption that two forces are acting on the ball, aerodynamic drag force and friction between grass and ball. It is also assumed that one of these two forces is always dominant and that the other is therefore negligible. The Magnus force, caused by the rotation of the ball, is not included since including it would require data on the exact spin of the ball which is not available at this time. The first part of Equation (3.1) models ball movement influenced only by aerodynamic drag. Where m is the mass of the ball, ρ is the density of air, C_D is the drag coefficient and A is the cross-sectional area of the ball. This part, therefore, models the forces affecting the ball while the ball is moving above the pitch surface. The second part models the ball's movement influenced only by the friction between the ball and the pitch with μ representing the friction factor and g the gravitational constant. In the motion model, the dominant forces are switched at $t = 2/3t_{max}$, where t_{max} is the total time for the ball trajectory. The time t when the dominant forces switch is found by Peralta Alguacil (2019) through a trial and error experiment where $t = 2/3t_{max}$ were found to yield trajectories most similar to real-life passes.

Equation (3.2) describes the motion of the players, where $\vec{F}$ is the driving force of player motion, their legs, and $k\vec{v}$ is a drag force limiting their maximum speed.

$$m\frac{d}{dt}\vec{v} = \vec{F} - k\vec{v} \tag{3.2}$$

Equation (3.3) is the solution to the differential equation in (3.2)

$$\vec{x} - \vec{x_0} = V_{max}\left(t - \frac{1 - exp^{-\alpha t}}{\alpha}\right)\vec{e} + \frac{1 - exp^{-\alpha t}}{\alpha}\vec{v_0} \tag{3.3}$$

where $\vec{x_0}$ and $\vec{v_0}$ is the initial position and initial velocity of the player respectively. $V_{max} = F/k$ is the maximum speed a player can reach, $\alpha = k/m$ is the magnitude of the resistance force and the player's direction of acceleration is represented by the unit vector $\vec{e}$. Equation (3.3) then describes a player's reachable area as a circle with the second part of the equation determining the center of the reachable area, and the first part determining its outer bounds. This equation for a player's reachable area is not only used in the pass probability model but is an important part of the optimisation procedure explained in Chapter 4.1.

Following the equations of motion for both the ball and players, a player is deemed to have a chance of intercepting a pass if he can reach any part of the ball's trajectory with a time $\Delta t = t_{int} - T \geq 0$ where t_{int} is the time for the player, and T is the time for the ball to reach a point along the trajectory of the ball.

An assumption is made that there exists some form of uncertainty regarding interceptions. Therefore, the probability of a player being in a position where he is able to intercept the ball at time t_{int} is modelled by using a logistic distribution presented in Equation (3.4). This uncertainty around Δt is represented by σ.

$$P_{int} = \frac{1}{1 + exp^{\frac{T - t_{int}}{\sqrt{3}\sigma/\pi}}} \tag{3.4}$$

Furthermore, an assumption is made that players increase their probability to control the ball during an interception with increased time in proximity of the ball. The probability that a player in proximity to the ball for a time t is able to control the ball is then given by the exponential distribution shown in Equation (3.5)

$$P(t) = 1 - exp^{\lambda t} \tag{3.5}$$

Combining Equations (3.4) and (3.5), the probabilities for each player to receive the pass is solved by the system of differential equations shown in Equation (3.6)

$$\frac{dP_j}{dT} = \left(1 - \sum_k P_k(T)\right) P_{int,j}(T)\lambda \tag{3.6}$$

Table 3.1 shows the values used for properties in the equations used to calculate ball trajectories and the reachable areas for players. All values are equal to the ones used by Peralta Alguacil (2019). Details on the reasoning behind the exact values of all properties are not provided in this thesis, but are found in Peralta Alguacil (2019) and Spearman et al. (2017). An illustration of the pass probability surface can is shown in Figure 3.1. In

Property	Value
m	$0.42\ kg$
ρ	$1.225\ kg/m^3$
C_D	0.25
A	$0.038\ m^2$
μ	0.55
g	$9.8\ m/s^2$
V_{max}	$7.8\ m/s$
α	1.3

Table 3.1: Values for different constants used in Equations (3.1) - (3.3).

this thesis, the pass probability model will be referred to by the abbreviation PP. Further explanations on the PP model can be found in Peralta Alguacil (2019).

3.1.2 Pitch control model

The pitch control model proposed by Fernandez and Bornn (2018) is based on how much influence a team is deemed to have over a certain area on the pitch. Influence, I for player

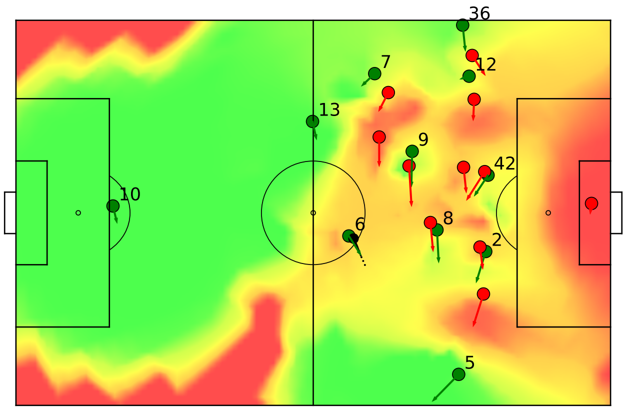


Figure 3.1: Illustration of PP in a situation during a match between Malmö FF (green) and Hammarby IF (red).

i at location p and time t, is based on both the position and velocity of the player as well as the distance between the player and the ball. Equation (3.7) shows the players influence at the position p normalised by the players position p_i.

$$I_i(p,t) = \frac{f_i(p,t)}{f_i(p_i(t),t)} \tag{3.7}$$

The function $f_i(p,t)$ is defined as the standard multivariate normal distribution probability density function as shown in Equation (3.8).

$$f_i(p,t) = \frac{1}{\sqrt{(2\pi)^2 det[COV_i(t)]}} \exp\left(-\frac{1}{2}(p - \mu_i(\vec{s_i}(t)))^T COV_i(t)^-1(p - \mu_i(t))\right) \tag{3.8}$$

From this measure of individual influence, the total pitch control of a team ,$PC(p,t)$, at position p and time t can be calculated using Equation (3.9).

$$PC(p,t) = \sigma\Big(\sum_i I_i(p,t) - \sum_j I_j(p,t)\Big) \tag{3.9}$$

Here i and j are the players on the two different teams and σ is the logistic function. Equation (3.9) thereby transforms the difference in control between the two teams at a point on the pitch to a probability range between 0 and 1. Calculated for every point on the pitch at a time t, a surface describing the degree or probability of control for the two teams is generated. Further details on the derivation of the equations and reasoning behind the pitch control models can be found in Fernandez and Bornn (2018). Figure 3.2 shows the pitch control surface during a match between Hammarby IF and Malmö FF. Areas controlled by Malmö are colored green and have values closer to 1, while areas controlled by Hammarby are red and closer to 0. PC, short for pitch control, will be used when referring to this model in this thesis.

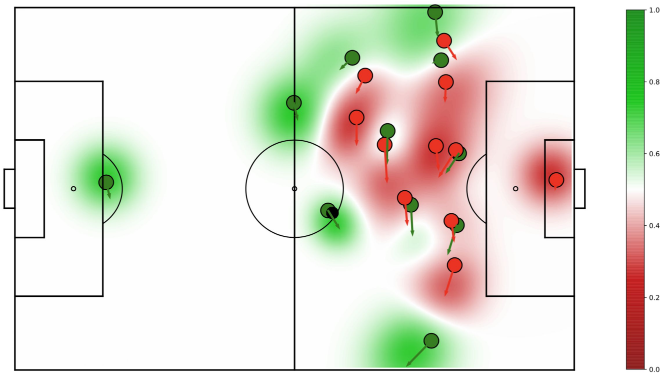


Figure 3.2: Illustration of PC in a situation during a match between Malmö FF (green) and Hammarby IF (red).

3.1.3 Pitch impact model

In general, a pitch impact model seeks to quantify the value, or impact, of possessing the ball in a certain position on the pitch. One central question these types of models seek to answer is therefore what the impact of moving the ball from one position to another is. In this thesis, the pitch impact model used is developed by the Swedish company Twelve. This model is a combination of an xG model, similar to Rathke (2017), and a model for the probability of a possession resulting in a shot. Equation (3.10) shows how the probability of a goal from a given pass is calculated.

$$P_{pass}(Goal) = P(goal|shot) \cdot P(shot) \tag{3.10}$$

$P(goal|shot)$ is the probability of a shot resulting in a goal, xG while $P(shot)$ is the probability that the possession chain the pass is a part of leads to a shot. A possession chain is a sequence of actions by a team without losing possession. Both parts of the model were fitted using logistic regression on data from three seasons of the English Premier League, Spanish LaLiga, and the UEFA Champions League. First, a regression is fitted using a value of 0 if the possession chain ends without a shot and 1 if it ends with a goal. Possession chains that lead to a shot, but not to a goal, are valued using a second logistic regression to find the probability of the shot ending in a goal (Peralta Alguacil et al., 2020).

Many possible applications of this model exist. One application is to evaluate teams and individuals on their on-ball actions in terms of how they change the impact of the current possession. This can be used to evaluate player performance or gain deeper insight, beyond just the final score or number of shots, into completed matches. Another application, and the one most useful for this thesis, is to use the results from this model to evaluate the impact of moving the ball to different locations on the pitch. Given the ball's location, a

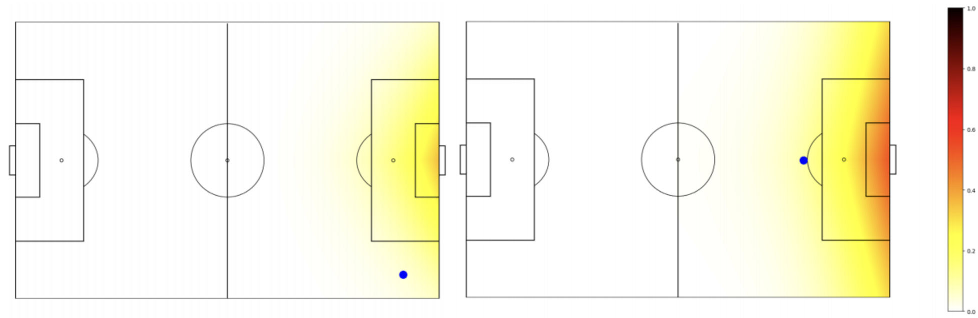


Figure 3.3: PI surface for two passes with different starting coordinates (blue dot). Color intensity gives the probability that a pass ending at that point results in a goal.

surface of the probabilities that a possession chain ends with a goal if the ball is moved to that point can be calculated. This can give an indication if a player is in a dangerous position for the opponent. As with the PP and PC models presented in the previous sections, further referrals to the pitch impact model will be done by the abbreviation PI. Figure 3.3 shows the PI surface for two different ball locations.

3.1.4 PC*PI model

As both the PC and PI models are probabilistic, another metric can be generated by multiplying the value of PC and PI, hereby referred to as PC*PI. This metric can be interpreted as a weighted PC model, by assigning a positive bias towards controllable areas with higher impact. The result of combining the PC and PI models is illustrated in Figure 3.4.

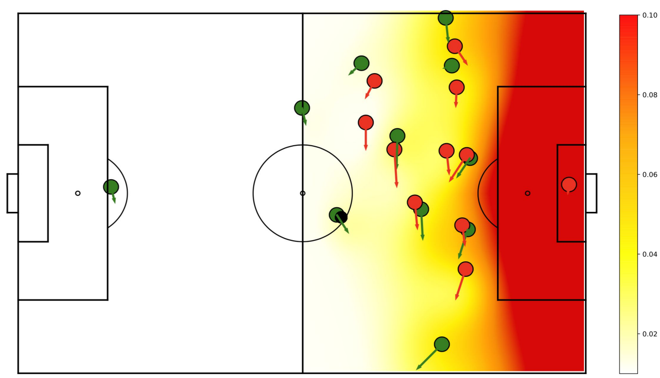


Figure 3.4: Illustration of PC*PI for the same situation as in Figure 3.2.

3.2 Bézier Curves

Gregory (2019) has previously used Bézier curves to model the runs of football players. An advantage of modelling runs as Bézier curves is that it creates a continuous curve from a set of discrete points so that a run can be interpreted at any point in time. Floater (2015) states that a Bézier curve of degree n, defined on some interval $[a, b]$, is a parametric polynomial given by

$$\mathbf{p}(t) = \sum_{i=0}^{n} \mathbf{c}_i B_i^n(u), \qquad t \in [a, b] \tag{3.11}$$

Here, u is the local variable, $u = (t - a)/(b - a)$, the points $\mathbf{c}_i \in R$ are control points of $\mathbf{p}$ and B_i^n is the Bernstein polynomial

$$B_i^n(u) = \binom{n}{i} u^i (1-u)^{n-i}, \qquad u \in [0, 1] \tag{3.12}$$

Intuitively, a Bézier curve can be interpreted as the center of a mass of a set of point masses. By allowing these point masses to vary by a parameter *t*, a curve is created.

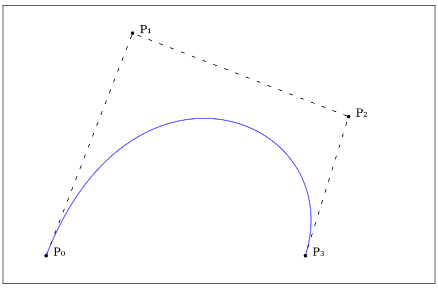


Figure 3.5: A third order Bézier curve and its corresponding four control points.

3.3 Centroid and Team Compactness

The centroid and team compactness are measures of team structure introduced in the preceding project thesis Cook et al. (2019). The centroid, or geometric center, is the arithmetic mean of all players on the pitch. At time t the centroid in x and y direction, given by $\bar{x_t}$ and $\bar{y_t}$, is defined as

$$\bar{x_t} = \frac{1}{n} \sum_{i=1}^{n} x_{it} \tag{3.13}$$

$$\bar{y_t} = \frac{1}{n} \sum_{i=1}^{n} y_{it} \tag{3.14}$$

where x_{it}, and y_{it} is the coordinates of player i at time t, and n are the number of outfield players on the pitch being tracked.

To quantify the distances between players on the same team, a definition of team compactness is introduced. Team compactness in x and y direction, CX_t and CY_t, is defined as the standard deviation of individual player positions with respect to the team centroid:

$$CX_t = \sqrt{\frac{\sum_{i=1}^{n}(x_{it} - \bar{x_t})^2}{n}} \tag{3.15}$$

$$CY_t = \sqrt{\frac{\sum_{i=1}^{n}(y_{it} - \bar{y_t})^2}{n}} \tag{3.16}$$

Furthermore, the total compactness is given by the Euclidean distance of the compactness in x- and y direction:

$$TC_t = \sqrt{CX_t^2 + CY_t^2} \tag{3.17}$$

3.4 Generalised Additive Models

A generalised additive model (GAM) is an extension of the generalised linear model (GLM) where the relationship between the dependent and independent variables are changed from the linear function $\sum_1^n \beta_i X_i$ to a more general function $\sum_1^n s_i(X_i)$. In these additive functions $s(\cdot)$ are non-parametric smooth functions estimated as part of the fitting procedure (Hastie, 2017). The result of this extension is a model that allows for the inclusion of non-linear relationships between dependent and independent variables. Equation (3.18) shows a GAM where the dependent variable depends on n smooth functions and an intercept term.

$$E(Y|X) = s_0 + \sum_{i=1}^{n} s_i(X_i) \tag{3.18}$$

A GAM with some explanatory variables represented as smooth functions and some as linear functions, shown in Equation (3.19), is an alternative when some variables are assumed to have linear relationships with the dependent variable while others are assumed to have non-linear relationships.

$$E(Y|X) = s_0 + \sum_{i=1}^{n} s_i(X_i) + \sum_{i=n+1}^{m} \beta_i X_i \tag{3.19}$$

Depending on the problem, different distributions can be set for the conditional mean $E(Y|X)$ and the link function $g(E(Y|X))$. Several distributions and link functions are possible through available frameworks, examples of distributions are, normal, logistic, gamma and beta distributions while examples of link functions are, identity, logit, and probit link functions. As the beta distribution is used later in this thesis, Equation (3.20) describes the probability density function of the beta-distribution.

$$f(x;\alpha;\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{3.20}$$

Where Γ is the gamma function, and α and β are shape parameters regulating the shape of the distribution. A GAM with smooth and linear terms with a set distribution of the conditional mean and a link function can, therefore, be expressed by Equation (3.21).

$$h(X) = g(E(Y|X)) = s_0 + \sum_{i=1}^{n} s_i(X_i) + \sum_{i=n+1}^{m} \beta_i X_i \tag{3.21}$$

3.4.1 B-Splines

One way of creating non-parametric smooth functions is to use penalised basis splines, known as B-splines. These splines consist of connected polynomial pieces that are joined at knots, represented in this thesis by x_i for knot i. The splines can be of different degrees, with B-splines of the degree 1 being two linear pieces connected at one knot, x_i, B-splines of degree 2 being three quadratic pieces connected at two knots, x_i and x_{i+1}, with splines of higher degrees following the same pattern (Eilers and Marx, 1996). Figure 3.6 is an illustration of five b-splines of degree 2. By using B-splines, a non-parametric

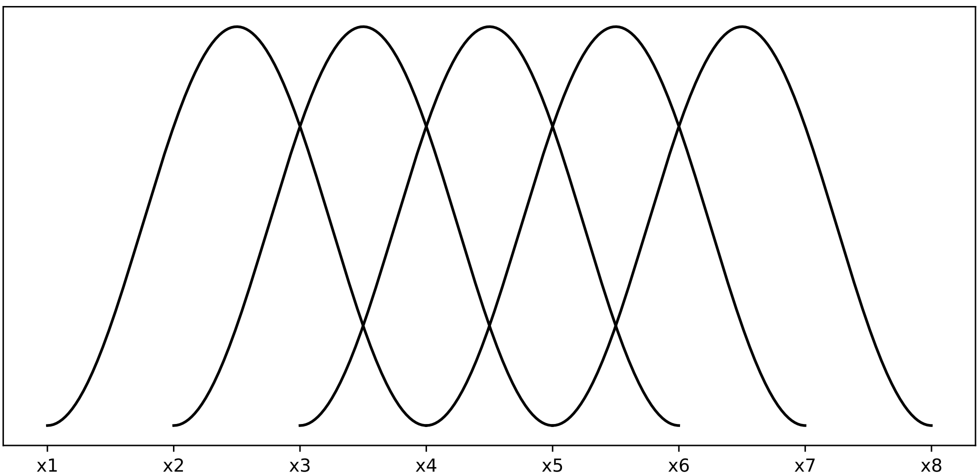


Figure 3.6: Illustration of 5 b-splines of degree 2

smooth function can be obtained as a linear combination of the splines. For a curve $\hat{y}$ fitted to the data (x_i, y_i) where the value at x for the jth B-spline of degree q is $B_j(x; q)$ Equation (3.22) represents the fitted curve. The estimated height or amplitude of spline j is represented by the coefficient $\hat{a}_j$.

$$\hat{y}(x) = \sum_{j=1}^{n} \hat{a}_j B_j(x; q) \tag{3.22}$$

An illustration of the smooth function created by five b-splines of degree 2 is shown in figure (3.7), where the black lines are the individual splines while the blue line represents the non-parametric smooth function that is the sum of the individual splines.

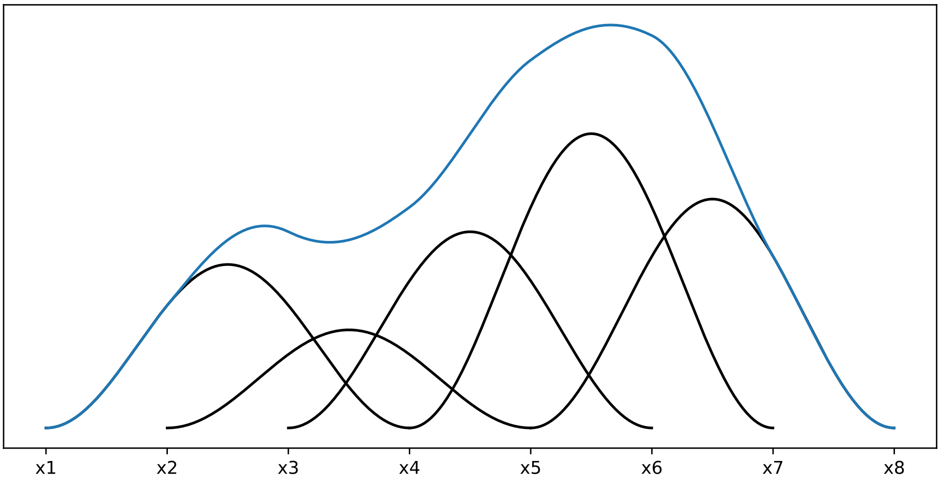


Figure 3.7: Illustration of a smooth function (in blue) created using five b-splines of degree 2

3.4.2 Model estimation

When estimating GAMs penalised likelihood maximation is used, which in practice will be achieved with penalised iterative least squares algorithm (P-IRLS) (Wood, 2006). With all splines of the same degree, q estimation of the curves is done by effectively minimising the objective function shown by Equation (3.23).

$$S = \sum_{i=1}^{m} \Big(y_i - \sum_{j=1}^{n} a_j B_j(x_i) \Big)^2 + \lambda \sum_{j=k+1}^{n} (\Delta^k a_j)^2 \tag{3.23}$$

Here, λ is the smoothing parameter, regulating the penalty on the differences between coefficients of adjacent B-splines. Figure (3.8) shows how two different values for the smoothing parameter changes the smooth function. A higher λ value penalises the difference in a values more, leading to a smoother function with less wiggliness. When building GAMs, λ has to be selected prior to fitting the model. To estimate λ, restricted maximum likelihood (REML) can be used (Patterson and Thompson, 1971). The REML approach measures the fit of the variance of the parameters by finding the mean of the likelihood over all possible values of B (Wood, 2006). Other algorithms to estimate λ exist, but the REML is preferred as it is less prone to converging towards a local minima (Wood, 2011).

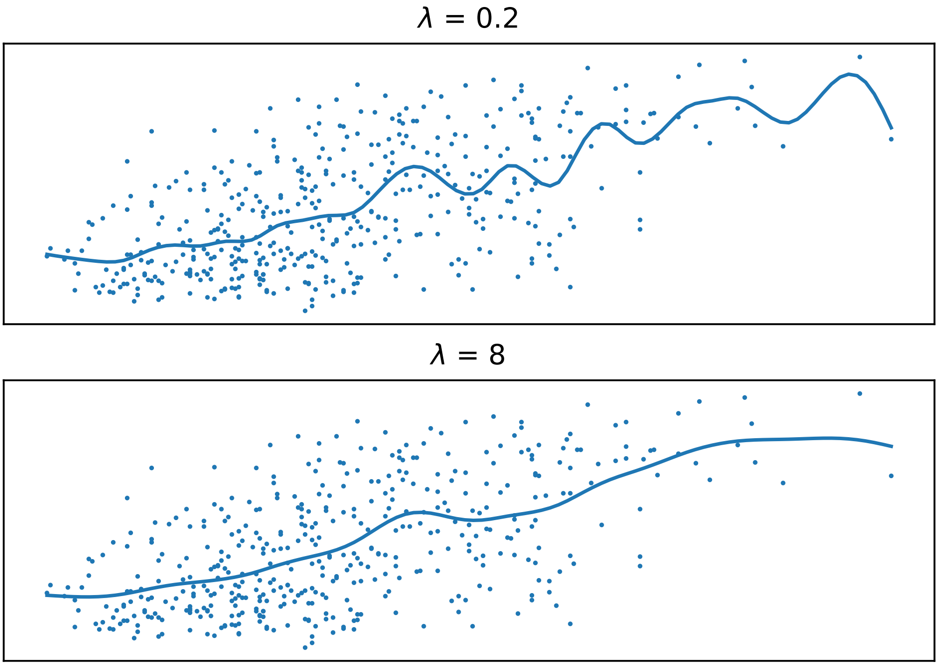


Figure 3.8: Illustration showing the effect of the smoothing parameter for $\lambda = 0.2$ and $\lambda = 8$

3.4.3 Model selection

To ensure a final model that accurately represents the problem at hand, it is desirable to penalise independent variables that are insignificant. This can effectively be achieved by introducing a shrinkage term to the smoothing penalty of the smooth functions. The shrinkage term ensures that when the smoothing penalty is large, the smooth is set to zero, which essentially means a linear term. By doing this, it is possible to perform an automatic feature selection by modifying the eigenvalues of these shrunk terms to a small positive number and penalise them out of the model. These terms will appear as horizontal lines at 0 in the final model.

3.5 Artificial Neural Networks

The field of artificial neural networks (ANNs) attempts to create structures in the spirit of neurobiology to solve computational problems of the kind that biology does effortlessly (Hopfield, 1988). This is done by mimicking the computational mechanisms of the animal brain. From a mathematical perspective, an ANN essentially functions as a non-linear statistical model (Hastie et al., 2009).

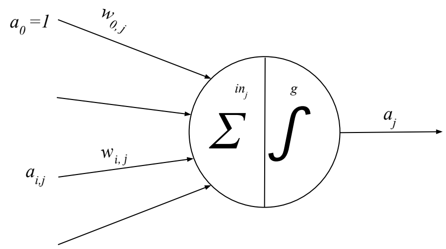


Figure 3.9: Illustration of an artificial neuron.

The core component of an ANN is the artificial neuron, hereby referred to as a node. A node is in essence a mathematical representation of the brain's core component, the neuron. To explain how the ANN functions mathematically, it is convenient to consider a single node of the network. An edge between nodes i and j propagates an activation a_i from i to j. The activation is multiplied by a weight $w_{i,j}$ that denotes the strength and sign of the input. Nodes also have a dummy input $a_0 = 1$. For every node j, the weighted sum of inputs in_j is computed and the activation function g is applied to the sum which gives the node's output.

$$a_j = g(in_j) = g\Big(\sum_{i=0}^{n} w_{i,j} a_i\Big) \tag{3.24}$$

3.5.1 Feed-forward neural network

The feed-forward neural network (FFNN) is a type of neural network, illustrated in Figure 3.10. The network is a directed network where information flows from left (input) to right (output), without any internal cycles. Nodes are organised in layers corresponding to their position in the network. The input and output layers handle input and output to the network respectively, while any layer between the two is called a hidden layer. Typically, the hidden layers and output layers have different activation functions. In the hidden layers a regularly used activation function is the rectified linear unit (ReLU) function, $r(x) = max(0, z)$. The activation function most frequently used in the output layer is the sigmoid function $f(x) = \frac{1}{1+e^{-x}}$. However, when using ANN's as a regression model, it is normal to employ a linear activation function for the output layer as the sigmoid function is bounded $x \in [0, 1]$. When using a linear activation function in the output layer, there has to be non-linear activation functions in the hidden layers for the ANN to be able to capture non-linear relationships in the input data.

3.5.2 Training of ANN's

Training of ANN's is usually conducted by using the gradient descent approach and back-propagation. Data is iteratively passed through the network, and the error between the actual target and the predicted value is calculated by a loss function, often the mean squared error (MSE). The derivatives of the loss function are calculated and propagated

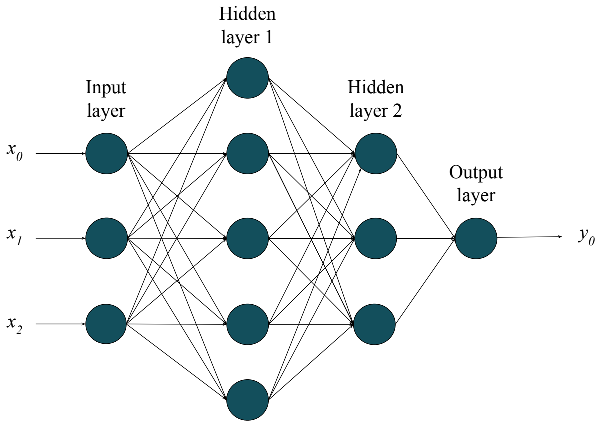


Figure 3.10: Architecture of a FFNN with two hidden layers.

back through the ANN using backpropagation. By using backpropagation, the weights of the nodes are adjusted by calculating how much they contributed to the final error in the loss function. An iteration of this entire process is referred to as an epoch.

When training ANN's, an optimiser algorithm is assigned with the task of minimising the loss function. An example is the adaptive moment estimation (ADAM) optimiser. This is an algorithm for first-order gradient-based optimisation, based on adaptive estimates of lower-order moments (Kingma and Ba, 2014). The training of the network is usually finished when the loss falls below a predetermined threshold, or after a given number of epochs have been completed.

Generalisation

One of the main goals of machine learning is to be able to create models to detect patterns and give accurate predictions on new and unseen similar data. The concept of generalisation refers to the ability to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model. The number of samples available, the complexity of the underlying data, and network architecture are all factors impacting the ability to generalise well. The process of training a network to generalise well is a demanding task with several challenges to overcome. One of the main being avoiding over- and underfitting. Figure 3.11 illustrates this problem. The blue points show the samples used to train the model, with the orange line as the underlying structure. The blue line is the fitted function. Underfitting is a problem of the model not being complex enough to capture the relationship between the features and a target variable, neither producing accurate predictions on training data nor being able to generalise to new data. Overfitting is the problem of a model being too complex relative to the complexity of the data, fitting the data well on the training set, but not being able to generalise well on new data. This is often a result of fitting to noise in the training set. The result of an overfitted model is that it performs well on the training data, but poorly on new, unseen data.

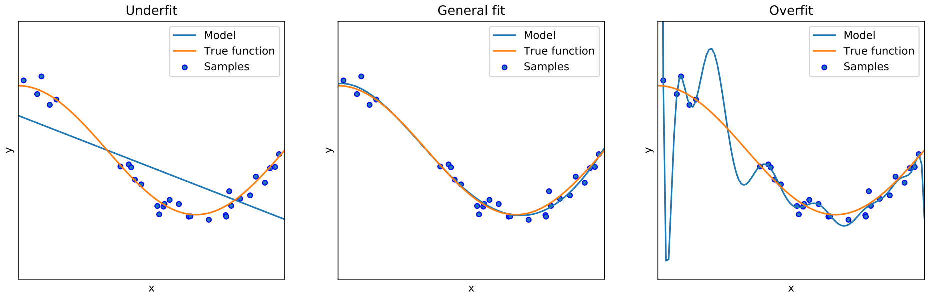


Figure 3.11: Example of three categories of model fits for ANN's.

L1 and L2 regularisation

When using a complex model architecture, regularisation techniques can be used to overcome overfitting and reduce test error. A consequence of employing regularisation is often an increased training error. L1 and L2 regularisation work by adding a norm penalty $\Omega(\theta)$ to the loss function, limiting the capacity of the ANN to fit to the data. Equations (3.25) and (3.26) show the different norm penalty in L1 and L2, with w referring to the weights.

$$\Omega(\theta)_1 = ||w||_1 \tag{3.25}$$

$$\Omega(\theta)_2 = \frac{1}{2}||w||_2^2 \tag{3.26}$$

L2 is commonly known as weight decay, driving weights closer to the origin by adding the sum of squared values of the weights to the loss function. L1 on the other hand adds the absolute value. L1 therefore has the ability to cause some of the weights to become zero, essentially working as a mechanism for feature selection. L2 regularisation forces the weights to be small, but does not make them zero and works best when all input features influence the output (Goodfellow et al., 2016, p. 227-231).

Early stopping

As large models are trained, one often observes that training errors decrease steadily over time, while the test set errors begin to rise after some training iterations. This means that the model is starting to overfit, as it is able to describe the training data very well, but loses the ability to generalise. Early stopping is a technique where the training is stopped when the error of the test set is starting to rise.

Chapter 4

Methodology

This chapter presents the methodology used to quantify and evaluate off-ball decision making. First, a technique is presented for finding optimal positions for maximising off-ball metrics. Then, a normalized scoring criterion is introduced to score individual movement decisions. Last, the method for determining and identifying relevant situations are detailed.

4.1 Defining Off-Ball Success

Quantifying off-ball involvements is a challenging task as there is no clear definition of success, and the outcome space is large. Players may have great impact on the outcome of a situation by creating space for their teammates, pressuring defenders, or positioning themselves in positions with a high probability of receiving the ball. Being at the right place at the right time is a task highly dependent on teammates and opposing players, as well as the players' assigned roles. With many ways of impacting a situation, quantifying and measuring off-ball involvements demands breaking the problem into different aspects. This means that no single optimal position may exist, but optimal positions can be found for maximising specific metrics.

A technique of identifying the optimal position for the different metrics of off-ball involvement will now be presented. This technique is developed by Peralta Alguacil (2019) and used by Hammarby IF for coaching and analysis. This will be restricted to using PC, PI, and PC*PI. As the code for computing the PP surface is not parallelised, using this metric to analyse large numbers of situations is at this moment so computationally expensive that it is considered outside the scope of this thesis. Therefore, PP is not used in this thesis, but as the reachable area for players is an important part of the methodology presented in this chapter, the part of the PP model describing player motion is used.

4.1.1 Scoring off-ball metrics

For this thesis, the off-ball metrics are restricted to the PC, PI, and PC*PI metrics, all using numerical measures to quantify off-ball involvements. All are represented as 68×105 matrices, covering each square meter of a pitch following UEFA category three or four pitch standard (UEFA, 2018). All matrices consist of probabilistic numbers, ranging from 0 to 1. In the PC-matrix a value of 0 corresponds to a position totally controlled by the opponent, 0.5 is a neutral point and 1 corresponds to total control of the position. For PI, the value corresponds to the probability of obtaining a goal if the ball is moved to that point. As the PC- and PI-matrices are probabilities, the PC*PI matrix is computed from the Hadamard product of the PC and PI matrices, also giving probabilistic numbers. The matrices are given as

$$A_m = \begin{pmatrix} a_{m,1,1} & a_{m,1,2} & \cdots & a_{m,1,105} \\ a_{m,2,1} & a_{m,2,2} & \cdots & a_{m,2,105} \\ \cdots & \cdots & \ddots & \cdots \\ a_{m,68,1} & a_{m,68,2} & \cdots & a_{m,68,105} \end{pmatrix} \qquad 0 \leq a_{m,y,x} \leq 1 \tag{4.1}$$

where y and x are the coordinates on the pitch, and m is the metric used. From the values in the matrices generated from a situation, a score can be assigned to the different metrics. In this thesis, the score player i obtains using metric m during observation j is referred to as $S_i(m, j)$. For PI, the score is computed by simply obtaining the value in the PI-matrix for the position (x, y) of the player:

$$S(PI) = a_{y,x} \tag{4.2}$$

For the PC- and PC*PI-matrices, the score is computed by summing the values in the matrices

$$S(m) = \sum_{x=x_l}^{x_u} \sum_{y=1}^{68} a_{m,y,x}, \quad m = \{\text{PC, PC*PI}\} \tag{4.3}$$

where x_l and x_u represent a lower and upper bound set on to the x axis for the area where the sum is calculated. This is done to ensure that the pitch control is restricted to an area of interest. This will have a substantial effect on the scores $S_(PC)$ and $S_(PC * PI)$, and therefore also the optimal position found. For this thesis, the lower bound x_l is set to the opponent player positioned second-most offensive, and the upper bound x_u is set at the opponent placed third-most defensive. This is the lower and upper bounds used by Hammarby IF for computing the PC score in their analyses. An illustration of this can be seen in Figure 4.1. If no such restriction is set, players will on many occasions increase the PC score the most by moving to non controlled areas, which are the areas where the PC-matrix value is 0.5. Non controlled areas can be observed as white areas on the pitch in Figure 4.1. For a defender during attacking play, this means moving backwards on the pitch, which intuitively is not optimal when attacking. On the other hand, for an attacker, this means pushing forward past opponent defenders, which is an attacking trait better assessed with the PI model, as the PI score increases when moving towards the goal. In essence, the lower and upper bounds help to cope with the non-controlling areas and makes it a tool for assessing players ability to capture areas that are controlled by the opponent.

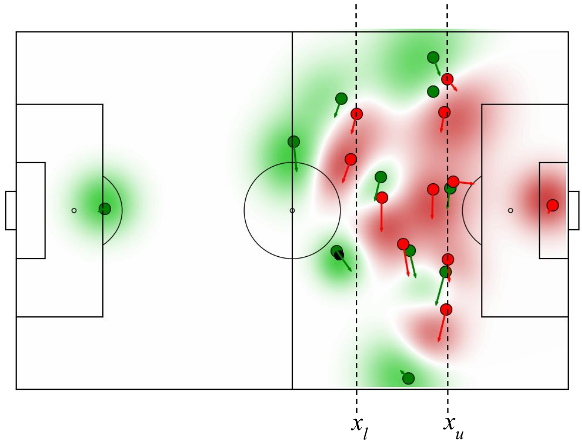


Figure 4.1: Example of lower and upper bounds, x_l and x_u, used when calculating PC and PC*PI matrices.

4.1.2 Reachable area and assumptions

By assessing the scores obtained from Equation (4.2) and (4.3), different positional alternatives can be compared by evaluating the scores obtained for each metric. The optimal position can also be found for a metric by finding the position giving the highest score. However, it would make little sense to uncritically optimise a player's position given the positions of teammates and opponents, as their position will likely be dependent on each other. It is, therefore, more reasonable to find the optimal position on an area that is reachable in the short term future, given a player's initial position and velocity. This way, the mutual dependency between the positions of players becomes less prevalent. A central assumption in this thesis regards players ability to predict future game states over short time intervals. As the optimal position in the future is dependent on the future positions of the other players, finding the optimal position is dependent on the player being able to predict other players' short term movement accurately. The assumption, therefore, becomes that predicting the movement of other players, and adjusting their own movement in response, is a skill that some players master better than others. This leads to some players being better at making positional decisions as they are able to achieve higher $S(m)$ in similar situations.

By using initial position and speed, the reachable area of a player is computed using Equation (3.3) from Section (3.1.1). Then, 200 sample points are created inside the reachable area using a random angle and radius stretched out from the reachable area centre. All sample points that lead to a player being positioned outside the pitch or in an offside position are removed. From the reachable sample points computed, the values of the different metrics are stored with the highest and lowest scoring points working as optimal and worst points for the different metrics. Figure 4.2 illustrates this process.

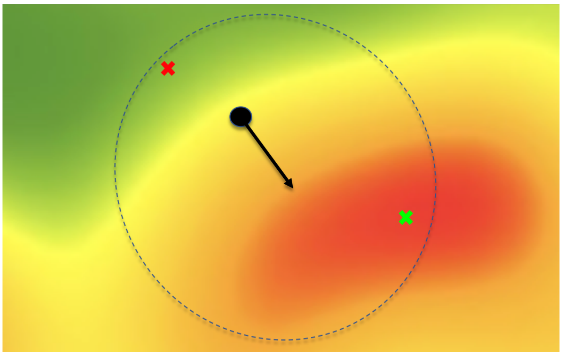


Figure 4.2: Illustration of reachable area (dashed line) generated from a player (black dot) with an initial position and speed. The coloured surface represents the value obtained from a chosen metric, with associated points scoring the optimal and worst score.

4.1.3 Normalized scoring criterion

Knowing the optimal positions for a situation, and the actual position of a player, a measure can be obtained of how close the player was from achieving the optimal score. A consequence of bounding the scoring area on movable objects, in this case, the opponents, is that the scoring criteria are changing depending on the situation. To be able to compare situations with one another, a normalized scoring criterion has to be introduced. Calculating the score $S(m, j)$ for all 200 sampled points, the maximum and minimum score are set as the optimal $S_{max}(m, j)$ and the worst $S_{min}(m, j)$ score respectively. Players' performance is therefore measured relative to their ability to influence the outcome score of the different aspects and not the actual value of the metric $S(m)$ itself. By using the players obtained score, and the maximum and minimum value, the relative score, $R(m, j)$, for observation j using metric m is defined as

$$R(m, j) = \frac{S(m, j) - S_{min}(m, j)}{S_{max}(m, j) - S_{min}(m, j)} \tag{4.4}$$

bounding the score to an interval between 0 and 1. This approach makes the score uncritical of initial position, making it possible for all players, regardless of the initial position and role, to obtain high scores on all metrics. Players are therefore evaluated on their decision making in that specific situation, and not merely the outcome of the situation as this is highly dependent on the characteristics of the situation itself.

4.1.4 Interpreting scores and strategies

From the normalized scoring criteria presented in Equation (4.4), a score close to 1 means that a player is either located at, or close, to the optimal position, or that the player has found an alternative position that achieves a score similar to the optimal position. A score close to 0 means that the player is located at a position scoring close to the worst score. The following sections introduce the fundamental trends for achieving optimal score for the three models used in this thesis.

Optimal point PC

When finding the optimal point for PC, it is necessary to recognise the effect of the lower and upper bound, x_l and x_u, when computing the score $S(PC)$. When a player is outside or close to the boundaries, the player is often encouraged to move deeper into the bounded area to increase $S(PC)$, as only values inside the boundaries contribute to this score. For a player already well inside the boundaries, the score increases if the player can capture uncontrolled areas, or is able to overtake control obtained by the opponent. Hence, no incentives to follow certain directions exists, with the individual situation deciding the optimal direction of movement.

Optimal point PI

For the PI model, the score increases with movement towards the goal line. This can be observed in Figure 3.3, as the value of a position increases closer to the goal. The optimal point for PI will ,therefore, always be towards the goal, but not past the offside line. This can therefore be considered as an attacking feature, scoring the pressure a player chooses to put on the opponent. It is important to recognise the level of intensity that is required to score high on this metric, as the player will have to move with maximum speed towards the goal to reach the optimal point. For attackers positioned close to the offside line, however, this point may be easier to reach as the point cannot move behind the offside line. A consequence of this is that players are encouraged to move behind the offside line, as this score exceeds the optimal score in the reachable area, giving them a $R_(PI, j) > 100\%$. This problem is addressed and handled in Section 4.2.2.

Optimal point PC*PI

As PC*PI is a combination of PC and PI, scoring high on this metric is the ability to increase them both. As with PC, this metric is scored by summing the values of the model within the lower and upper boundaries. The clue is therefore not how much pressure the player achieves in his position, but how much his position puts control over an area of pressure, in this case, closer to the opponent's goal. An optimal score can, therefore, be obtained if the player is able to capture control from the opponent higher up the pitch, encouraging moving towards the opponent's goal, while obtaining control.

4.1.5 Identifying strategies and evaluating performance

Different player roles and situations impact the attractiveness of achieving high scores on individual metrics. A defender may have a higher desire to focus on an aspect associated with defending, while an attacker may choose to focus on an attacking aspect in the same situation. From the collected scores, the aspects a player chooses to focus on can be identified. The overall performance on the different metrics can be found by averaging the scores from the collected data. By restricting the data set only to include situations meeting certain conditions on game state, one can identify what aspects a player chooses to focus on for a specific situation. As situations differ by many factors, a prediction model can be created for situations to give an indication of how the player is expected to perform in a particular game state. Performance can, therefore, be evaluated by how the player performed compared to the expected performance from the prediction model, indicating if a player performs better or worse than what is to be expected for the different metrics. Creating such a prediction model demands a detailed description of the game state, a process presented in Chapter 6.

4.2 Situations of Interest

Building accurate prediction models demands collecting a training set that covers situations and outcomes that are similar to the situations chosen to model. Football is a sport of continuous states, meaning that there exists an infinite number of unique situations. Finding two identical situations are, therefore, unlikely. Collecting many observations is, therefore, a necessity for creating general models, covering situations that are as similar as possible to the chosen situations. The more observations collected, the probability of the models having seen a similar situation increases. However, an increased number of observations demands more computational power. Computing optimal position for a player is a computational heavy process. This is especially the case for the PC model, where Equation 3.9 has to be calculated for all 105*68 square meters of the pitch, for each of the 200 points sampled in the reachable area. By using an external GPU[1], and parallel computing, this process is sped up. Still, collecting one observation takes approximately one second. With a total of 1,296,000 seconds played in the 2019 Allsvenskan, and 22 players on the pitch, an uncritical computation of optimal position for every second would take approximately 250 days with the equipment used for this thesis, collecting 28,512,000 observations. To overcome this problem, the problem is narrowed down to a situation based approach; only collecting observations from situations that fit a certain description.

4.2.1 Selecting situations

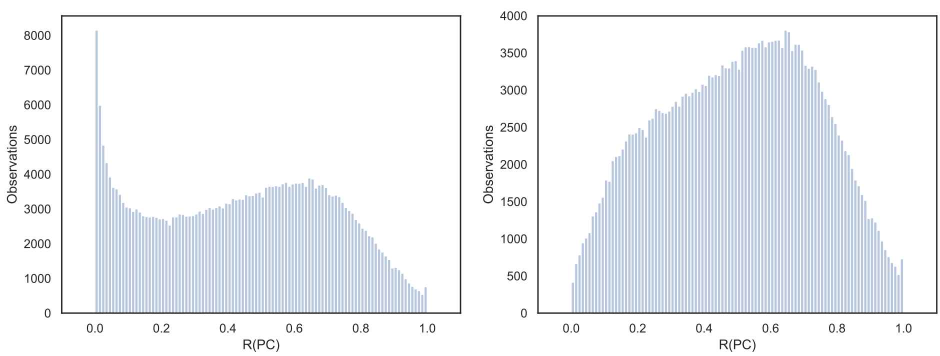
By specifying situational constraints in the tracking data when collecting observations, the number of similar situations collected can be increased without running through all the data. There are a number of different types of situations that could be analysed through the lens of the models presented in this thesis. With football being a two-way game, the objective for both teams is the same. Both teams seek to score goals while not conceding, meaning that both defensive and offensive situations can be analysed through the same lenses. Currently, the PC and PI models have primarily been used to assess performances during offensive situations, so continuing with this focus is deemed to be the most useful alternative. Passes are suitable as markers for when situations of interest occur as they are discrete and recognisable events. Since they move the ball from one place to another, passes are assumed to be catalysts for movement as players seek to re-position themselves for the next situation. The assumption, therefore, becomes that passing events are situations where players are forced to make positional decisions, and it is these types of decisions that are deemed to be most intriguing to analyse. Passes made in the attacking half of the pitch are therefore chosen as markers for the situations to be analysed in

[1]The GPU used in for this work: Nvidia GTX 1060 graphical processing unit

this thesis. The initial positions of players are at the exact time of the pass, with optimal positions computed for the following second.

4.2.2 Removing non-influential and invalid situations

Situations where the player is experiencing limited ability to influence, meaning that the player's decision has little impact on the PC score, are deemed to not be of interest. These situations are, therefore removed from the data set. The reason for this is that the player will likely focus on other aspects when not in a position to influence the PC score, e.g. moving to an area where the player can influence. This is often the case for defenders positioned far away from the lower and upper bound, x_l and x_h, for where the PC and PC*PI score is computed. This can provide inconsistent contributions to the performance evaluation of individual players later conducted in this thesis. A threshold is set for the variability of the $S(PC)$ obtained for the 200 sampled points of the reachable area. The standard deviation is used, and the threshold set to 1. This implies that the standard deviation of the $S(PC)$ values in the reachable area must be above the equivalent of capturing one full square meter of total control from the opponent within the lower and upper bound of the scoring area. This process removes approximately 15% of the observations. The distribution of $R(PC)$ scores obtained by players before and after removing these observations can be seen in Figure 4.3.



(a) Before **(b)** After

Figure 4.3: Illustration of the PC score distribution before and after removing observations deemed non-influential.

Observations of goalkeepers and the player receiving the pass are removed, as they are assumed to follow different criteria and rules for positioning. Goalkeepers' objective is to stay back to protect their own goal, and pass receivers follow the trajectory of the ball. Furthermore, players scoring well above or below the score bounds of 0 and 1 are removed, as they violate the score criteria. This can happen for several reasons; The main reason being errors in the tracking data, further elaborated in Chapter 5. Another reason is the sampling process of positions in the reachable area. As the full reachable area is not covered completely, the player might be able to find a position yielding a score slightly over or under the optimal and worst points of the reachable area. These observations are not filtered, but the scores are transformed to the closest scoring bound. A third reason is found when attacking players move behind the offside line, scoring higher than 1 on PI. A scoring threshold of 1.1 is used, removing the most severe cases. This is the case for approximately 1% of the observations.

Chapter 5

Data

The following chapter details the process of preparing the data for use in the models created in Chapters 6 and 8. An introduction to the data sources is given, followed by an explanation of the structuring, merging, and processing of the data.

5.1 Signality Tracking Data

The data used throughout the thesis was provided Signality. Signality provides a video-based tracking system that automatically tracks and tags all 22 players, ball, and referees on the field. The data used is from the entire 2019 season of the Allsvenskan, which corresponds to a total of 240 games. The data files provided by Signality are structured in a folder for every half of each match i.e., 45 minutes of game time. Each folder contains the following separate json-files; events, tracks, info, and stats. An explanation of the content of the files is given in Table 5.1.

File	Content
events.json	Information about events during the game. Typically runs, passes and interceptions.
tracks.json	Gives the positions of players, ball and refferees on the pitch 25 times a second.
info.json	Details on players on home- and away teams. Mapping of player name to tag ID in the tracking file.
stats.json	Aggregated stats for both players and teams throughout the game.

Table 5.1: File types in supplied data set and the respective content of each file.

By using a video-based tracking system, Signality is able to identify the locations of all players on the pitch, without using wearable sensors. The system tracks the different players by linking their jersey number to a tag id, which uniquely pairs the player to the tracked position.

5.2 Extracting Relevant Situations

As discussed in Section 4.2, a choice was made to focus on attacking passing situations. One way to identify attacking passing situations is to go through all matches in the data and manually label the relevant situations. But with a large number of matches to analyse this would be an extremely time-consuming process and therefore not feasible. Instead, a set of conditions are defined with the intention of using them to automatically identify relevant situations and establish ball possession for players and teams. These conditions can also be used to define the exact timing for when these situations occur, as well as the consistency in timing and state when extracting similar situations. The following game states and timing can be drawn from associated conditions:

i) **Identifying the player with possession of ball:** Ball speed under 10 m/s and at least one player is positioned within 3 meters of the ball. If these conditions are not met, no player has possession. If they are, the player closest to the ball is the player with possession.

ii) **Identifying the team with possession of ball:** If a player is deemed to be in possession of the ball, following the first condition, the team of the player is deemed to be in possession. If no player has possession, the team of the next player with possession of the ball has possession. One team is therefore always in possession of the ball.

iii) **Identifying passes and their exact time of occurrence:** Signality provides an event file that contains information on passes, their timing, and the players conducting them. However, there exist small inconsistencies in the timing of the passes from the event files to the tracking files. To cope with this, conditions are added to define the exact timing of passes. The player making the pass has possession of the ball within ± 2 seconds of the pass in the event file. The time when the player making the pass loses possession of the ball, again following the first condition, is then deemed to be the exact time of the pass.

A short video can be accessed from the following link[1] showing an excerpt of extracted successful passes on the opponent's half made by Hammarby IF against GIF Sundsvall, 17. August 2019. Positions of the players and ball are gathered from *tracks.json*. The player with possession of the ball is identified by being bold. The start and end coordinates of passes, registered from the *events.json*-file, are represented by arrows, with their appearance indicating the exact time set for the pass, found using conditions previously presented. Hammarby is represented in red, while GIF Sundsvall is purple.

[1] `https://drive.google.com/open?id=1JGloipK3E6SptjMrWrD1Fb0hEOn1roGS`

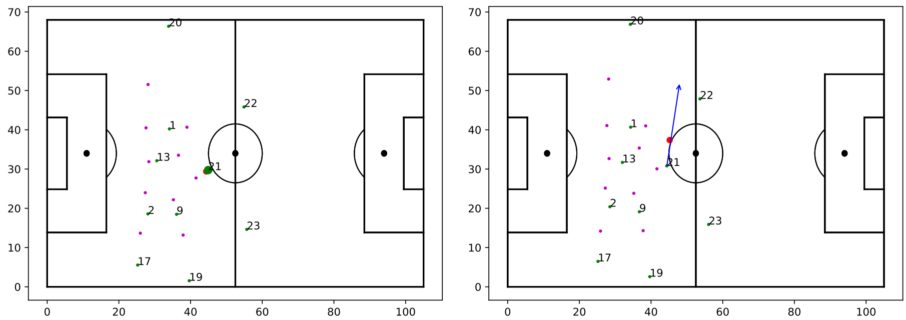


Figure 5.1: Two screenshots from the linked video showing a pass between player 21 and 22 during the match between Hammarby IF and GIF Sundsvall, 17. August 2019.

5.3 Handling Errors in the Data Set

The downside of using video-based tracking systems compared to GPS or radio-based systems is the lower precision and its higher unreliability leading to some missing positional data. The following sections present the issues identified in the work done in this thesis, and the actions taken to resolve the problems.

5.3.1 Missing positions

Throughout the data set, there are many cases where the players positions are not captured as the system temporarily loses track of the players. In these moments the x- and y-coordinates are set to -1. These occurrences normally last a short time before the system again is able to track the players' positions. As this is a regular happening, discarding all events where the position of a single player is missing for a short period of time would lead to the omission of large amounts of data. Therefore, the positions of players with missing data points are estimated using Bézier curves. The curve is created by using the players' positions before and after the missing data points. Missing data points can then be estimated for a desired time, t, by using Equation 3.11. Figure 5.2 shows an example of a situation where tracking data is missing for a part of the sequence. The Bézier curve allows for estimation of the player's position at any point in time throughout the interval. Another advantage is that the Bézier curve better perseveres the shape of the trajectory compared to a regular linear interpolation, which again means that vector components calculated for the players' speed are more accurate.

5.3.2 Players moving unreasonably fast

Another issue encountered with the data set was that the players occasionally seemed to change their position on the field too fast after not being tracked for a short period of time. Figure 5.3 shows an example where the position of the player drops out and is reestablished. Between the blue and red dot, there are 57 frames where the player's

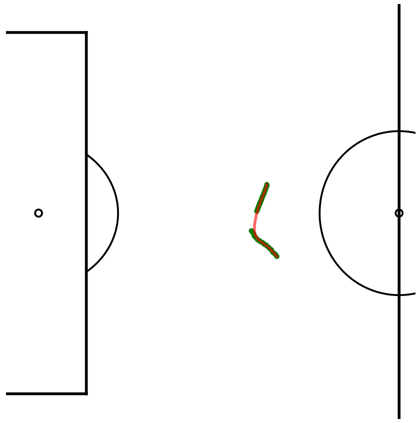


Figure 5.2: Example of a situation where a player's position is missing from the tracking data. The green dots are the positions of the player one second prior and post to the moment where the tracking data is lost and the red line is the Bézier curve representation of the player's trajectory.

position is not tracked, which equates to a total of 2.32 seconds of game time between the two dots. The distance between the two points is $\simeq 27.98$ meters, which implies that the player would have needed to have moved with an average speed of 12.06 m/s over the time period where the player was not tracked. Peralta Alguacil (2019) employed a maximum speed of 7.8 m/s when modelling player movement, and this maximum speed has also been used in this thesis. Findings from Gregory (2019) also confirm that players rarely reach speeds of 7.8 m/s in games, so a speed of 12.06 m/s appears to be unreasonable to achieve. This case is believed to be due to inconsistencies in the tracking system where the system briefly switches the id's of two or more players, before switching them back. This can be a result of the video-based tracking system not being able to accurately identify the jersey number of the players in situations where several players are close to each other. Cases like the one shown in Figure 5.3 are therefore discarded as they most likely include errors in the tracking data.

5.3.3 Discrepancies between file types

Another issue observed in the data set is that in some matches the players' position in the *events.json* file does not conform with the corresponding position of the players found in the *tracks.json* file. The issue has been partially dealt with by allowing for ± 2 seconds when searching for the pass in *events.json* file. However, increasing this time window could possibly allow for the system to tag a different involvement, not necessarily a pass, made by the same player which will lead to inconsistencies. The result of this issue is that a substantial amount of the matches have been omitted from the final data set, as the passes could not be identified in the *events.json* file. By comparing live video feeds from a few of these games, there appeared to be a shift in the timestamps of the tracking data compared to the event data, which lead to inconsistencies between the files.

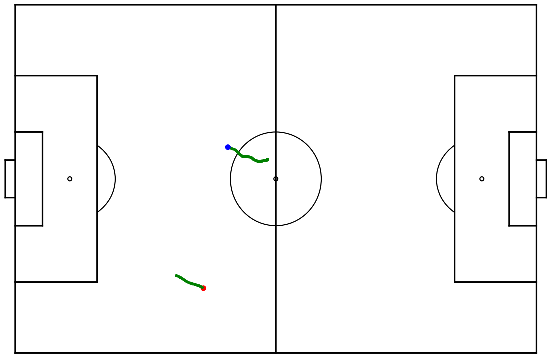


Figure 5.3: Illustration that shows the movement of IFK Göteborg player Tobias Sana through a 5 second interval in the game between IFK Göteborg - Östersund on 2 November 2019. The green line shows Sana's position, while the blue and red points show respectively the last position before-, and the first position after his position was not tracked. Sana is running from right to left through the entire sequence and a total of 2.32 seconds elapse between the red and blue points.

5.4 Final Data Set

The final data set used for analyses in this thesis consists of data from 144 games. From 34 208 passing situations, a total of 250 245 off-ball observation are collected, after removing invalid observations. 379 players from 16 teams playing in the 2019 season of the Allsvenskan is represented in the data set, with the average number of observation being 660 for each player and 15 640 for each team.

5.5 Player Roles

Football players are often described and grouped by the role they are assigned to within the team. The role of every player for each match in Allsvenskan was provided by Hammarby IF. A player often has different roles throughout the season. Therefore, the player's most-played role through the season is the role used in this thesis. This data is used to look for role-specific behaviour in positioning. In some of the models introduced in the following chapters, the role of the player is joined with associated observation, potentially providing descriptive information regarding the player's behaviour by serving as an indicator of areas that the player is attracted to in a given situation. The different roles assigned to the players in this thesis are detailed in Table 5.2.

5.6 Football Manager 2020 Player Ratings

Football Manager 2020 (FM20) is a simulation-based video game that allows the user to take the role of the manager of a football team. The game is famous for its vast scouting network of over 1300 scouts, and professional managers have admitted to using the game to aid them in their job (Smith, 2015). These scouts rate and grade football players by

Abbreviation	Role
GK	Goalkeeper
CB	Centre back
FB	Full back
MF	Midfielder
AM	Attacking midfielder
FW	Forward

Table 5.2: The different roles that are assigned to the players.

many technical, mental, and physical attributes. The ratings for the different attributes vary from 1 to 20, where a higher value implies that a player is better at the given attribute. The data regarding the ratings of the player attributes in FM20 is obtained from FmDataba and is consistent with version 20.4.0 of FM20 (FmDataba).

Chapter 6

Modelling of Situation Specific Positional Strategies

This chapter presents the model set-up for analysing situation-specific positional strategies of players, based on the metrics presented in Section 3.1. Two different classes of models, generalised additive models, and artificial neural networks are developed with the goal of predicting, analysing and evaluating players' positional decisions. The specific situations included in the analysis are passing situations in the attacking half, with the reasoning for choosing these situations following Section 4.2. In the first part of this chapter, the dependent and explanatory variables, and the methods for creating them, are explained. This is followed by a presentation of the different methods used to model player decision making. Then, the methods for validating the models and the results from this validation process is presented, followed by a presentation of the contributions of some different variables impacting the predicted outcome of the models. Finally, the results are discussed and concluded before moving on to the next chapter where some applications of these models are presented.

6.1 Dependent Variables

The three metrics PC, PI, and PC*PI, previously introduced in Section 3.1, will be used to analyse individual positional strategies of players. They form the basis for three dependent variables, $Y(PC)$, $Y(PI)$, and $Y(PC*PI)$, working as proxies for different strategies a player can make. The purpose of modeling these dependent variables is to see how a player performs compared to expected performance for the individual metrics.

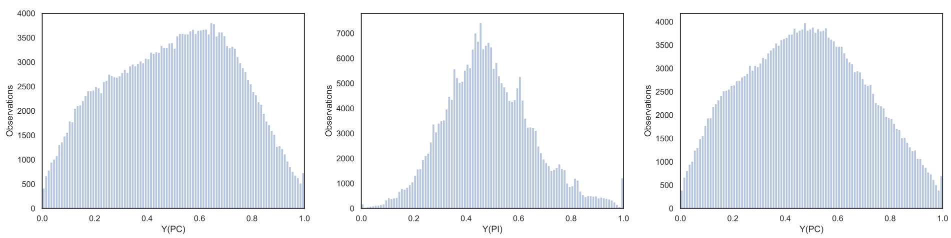
As detailed in Section 4.1.3, the normalised scoring criterion gives players a score in the range 0 to 1 for each metric during a situation. This is because the score $S(m, j)$ is computed relative to the maximum $S_{max}(m, j)$ and minimum $S_{min}(m, j)$ sampled values in the player's reachable area, functioning as upper and lower bounds for the possible performance. The score can, therefore, be seen as a score of how close, in the range 0% to 100%, a player was of obtaining the optimal value for a metric one second into the future of an initial position and game state. This approach makes the scoring uncritical of position on the pitch and the role type of the player, making it possible for all players, regardless of external factors, to achieve the whole range of values for all the metrics. Strategies can, therefore, be identified by looking at how the player is able to perform on the different metrics. Using metric m for situation j, the dependent variable $Y(m, j)$ follows Equation (6.1)

$$Y(m, j) = R(m, j) = \frac{S(m, j) - S_{min}(m, j)}{S_{max}(m, j) - S_{min}(m, j)} \tag{6.1}$$

The three dependent variables are presented in Table 6.1, with their distributions illustrated in Figure 6.1.

Dependent variable	Metric	Range
$Y(PC)$	PC	[0, 1]
$Y(PI)$	PI	[0, 1]
$Y(PC * PI)$	PC*PI	[0, 1]

Table 6.1: Dependent variables used to model situation specific position strategies.



(a) Y_{PC} **(b)** Y_{PI} **(c)** Y_{PC*PI}

Figure 6.1: Distribution of the three dependent variables for the GAMs.

6.2 Explanatory Variables

Explanatory variables are chosen based on whether or not they are believed to affect the dependent variables, with a goal of describing the game state as well as possible. Table 6.2 shows the initial list of explanatory variables included in the analyses. To make the variables more interpretable for the models, a transformation to vector components is introduced for several of the variables. This is further explained in Section 6.2.6. All explanatory variables describe the game state at the start of the 1 second time interval, except for the variables containing information on the optimal points, namely the locations of optimal points and their standard deviation. These are variables generated from a future state. A discussion around the inclusion of the different explanatory variables follows, explaining how they are believed to describe a specific situation.

6.2.1 Proxies of the implication of player's position on the pitch

The initial coordinates of the player are included in variables X_1 and X_2, while the initial velocity is represented by X_3 and X_4. The initial coordinates and velocities are thought to be fundamental properties for the situation a player faces. They describe the position the player has on the pitch, as well as the player's instantaneous momentum and direction of movement. As the game of football is constricted to a confined area, the coordinates are bounded by an upper and lower limit, $x \in [0, 105]$, $y \in [0, 68]$. The position gives an indication of the player's involvement and the boundaries of where the player can move to. Certain areas on the pitch can also affect the intensity and style of play. The velocity of the player can have an effect on the initial intensity level and the desire to move to a different position.

6.2.2 Proxies of player's position relative to teammates, opponents, and ball

The positions of teammates, opponents, and the ball are represented by variables $X_8 - X_{15}$. These features are represented as vector components from the player's initial position to the feature's position. This process is further explained in Section 6.2.6. X_8 and X_9 represents the vector components, in x- and y-direction, towards the position of the ball. With the ball being the centre of attention in the game of football, the ball position relative to the player is an important feature describing how close the player is to the center of focus. Further, $X_{10} - X_{13}$ represents the vector components towards the centroid of both the player's own team and the opposing team. These values are calculated using Equations 3.13 and 3.14, and are the means of the teammates' and opponents' positions in x- and y-direction. These features are included to describe how the player is positioned relative to the center of both teams, possibly affecting the role a player chooses to take in a situation. The vector component towards the nearest opponent is represented by X_{14} and X_{15} and is included as a proxy for the pressure the player is currently under from the opposition. This can indicate if a player is closely marked by an opponent, or not.

The velocities of team and opponent centroids are represented by $X_{22} - X_{25}$. This is calculated as the mean velocity of the respective players in x- and y-direction, and gives information on the direction and momentum of play. Additional variables describing the compactness of teammates and opponents are represented with $X_{26} - X_{29}$. These variables indicate how much space there is between the players on the pitch, and are calculated using Equations 3.15 and 3.16. As with variables X_{14} and X_{15} describing the closest opponent, the variables describing compactness can provide information on the pressure a player faces, as these variables represent how much space there is between players.

Variable	Description	Type
X_1	Position of player x-coordinate	Continuous
X_2	Position of player y-coordinate	Continuous
X_3	Velocity of player x-coordinate	Continuous
X_4	Velocity of player y-coordinate	Continuous
X_5	Standard deviation of PC	Continuous
X_6	Standard deviation of PI	Continuous
X_7	Standard deviation of PI*PC	Continuous
X_8	Position of ball x-coordinate	Continuous vector component
X_9	Position of ball y-coordinate	Continuous vector component
X_{10}	Team centroid x-coordinate	Continuous vector component
X_{11}	Team centroid y-coordinate	Continuous vector component
X_{12}	Opponent centroid x-coordinate	Continuous vector component
X_{13}	Opponent centroid y-coordinate	Continuous vector component
X_{14}	Closest opponent x-coordinate	Continuous vector component
X_{15}	Closest opponent y-coordinate	Continuous vector component
X_{16}	Optimal point control x-coordinate	Continuous vector component
X_{17}	Optimal point control y-coordinate	Continuous vector component
X_{18}	Optimal point impact x-coordinate	Continuous vector component
X_{19}	Optimal point impact y-coordinate	Continuous vector component
X_{20}	Optimal point PI*PC x-coordinate	Continuous vector component
X_{21}	Optimal point PI*PC y-coordinate	Continuous vector component
X_{22}	Team centroid speed x-direction	Continuous
X_{23}	Team centroid speed y-direction	Continuous
X_{24}	Opponent centroid speed x-direction	Continuous
X_{25}	Opponents centroid speed y-direction	Continuous
X_{26}	Team compactness x-direction	Continuous
X_{27}	Team compactness y-direction	Continuous
X_{28}	Opponent compactness x-direction	Continuous
X_{29}	Opponent compactness y-direction	Continuous
X_{30}	Game time	Continuous
X_{31}	Passing angle	Continuous
X_{32}	Direction of pass (left or right)	Binary

Table 6.2: List of explanatory variables used in the situation-specific positional strategies models.

6.2.3 Proxies of distance to optimal points

Vector components towards the optimal positions within the reachable area for PC, PI, and PC*PI are included in variables $X_{16} - X_{21}$. These optimal points are not visible for the player, with the individual player's intuition forming these points. However, they are included to give an indication of where and how far away the player should move to obtain a high score for the respective metric. A distance far away from the initial position demands more movement, hence higher intensity, making it a more demanding task to obtain an optimal score.

6.2.4 Proxies of influence on situation

The standard deviation of PC, PI, and PC*PI are included in variables $X_5 - X_7$. These variables show the variation in the metric scores, $S(m)$, possible to obtain within the reachable area of a player. A high standard deviation means that there is large variation in the possible scores, making movement highly influential on the score. A small value indicates that the player is not in a position with the ability to impact the metric score, and the influence on the game is lower. A high variation may give an increased desire to obtain the optimal score, as the game state is more dependent on the player, and the player's contribution may be important to other players.

6.2.5 Proxies describing the time and the pass

The last explanatory variables included are variables seeking to describe aspects of time and situation. Explanatory variable X_{33} describes the game time of the passing situation. Game time may affect positional decisions in several ways. Among those effects are physical and mental fatigue that increases over time and changes in offensive or defensive mentality as time is running out. To describe the direction of the pass, two variables are constructed. X_{34}, that describes the angle of the pass on a continuous 0 to 1 scale, from straight backward to directly forward, and X_{35} describing the direction of the pass as a binary left/right variable. Together, they capture the direction of the pass and could have been joined into one variable. However, as the two aspects are interesting on their own, they are separated into two variables.

6.2.6 Data representation

To make the data more interpretable for the models, several steps have been taken to modify the explanatory variables. These steps include vector transformation, scaling, and mirroring of the data.

Transforming to vector components

As involvement and contribution to the game state are heavily dependent on the player's initial position, variables $X_8 - X_{21}$, are transformed to the difference between the initial position to the coordinate of the variable, separated on x- and y-direction. This is to provide the game state from the perspective of the player. The transformation is illustrated in Figure 6.2. Testing shows that this approach gives more precise results than using untreated coordinates.

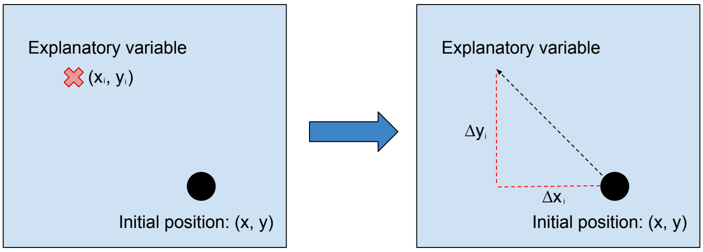


Figure 6.2: Transforming an explanatory variable to a vector component from the initial position of a player.

Feature scaling

All continuous variables are scaled to be in range zero to one. This implies finding the highest and lowest values for each feature in the data set and set these values to one and zero respectively. The rest of the variables are then set between 0 and 1, relative to the maximum- and minimum values of that feature. The variables are then scaled back to original values when interpreting the predictions.

Mirroring

The data is modified so that the attacking direction is always to the right. This means mirroring all coordinate dependent variables on situations where the team in possession of the ball is attacking to the left. Hence, all situations are modelled with attacking direction facing to the right as illustrated in Figure 6.3. The direction of play relative to the camera position is assumed to be uninteresting. Mirroring of the data is therefore done to be better able to compare and assess situations happening with opposite attacking directions relative to the camera.

6.3 Modeling Techniques

For this thesis, two different modeling techniques are used; generalised additive models (GAM) and feed forward neural network(FFNN). All models are created with an intention to predict the numerical value of the dependent variables from a set of explanatory variables. The two techniques have different properties and benefits, with this section seeking to address these.

6.3.1 GAM

Generalised additive models are chosen for their combination of the desirable qualities of prediction and explanation. When features are suspected to affect the dependent variable

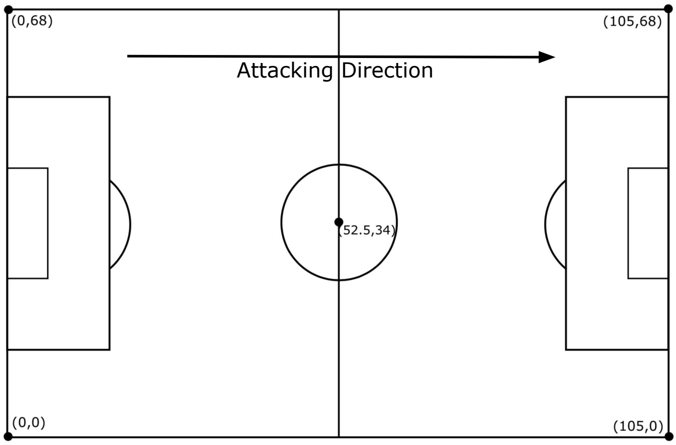


Figure 6.3: Illustration of attacking direction.

in non-linear ways, a GAM adds additional prediction power and flexibility by including smooth functions compared to a generalised linear model (GLMs). As many aspects of football are suspected to be non-linear (Cook et al., 2019), to capture these effects, GAM is chosen over GLM as the interpretive modeling technique. Furthermore, a GAM where all features are linear becomes a GLM, so using a GAM compared to a GLM is expected to be an improvement both in prediction power and interpretability. As with a GLM, the results produced by a GAM is interpretative beyond simply the model output for the dependent variable. For all features, whether they are linear, categorical, binary, or non-linear smooth functions, their contributions to the outcome of the dependent variable can be analysed. Each feature can, therefore, be analysed individually for their contributions to the dependent variable, leading to the GAM offering insight into both the result, through its predictive power, and the feature contributions through the interpretability of the model.

Explanatory variables used in GAM

An important part of creating a GAM is deciding how to represent the features involved in the GAM. Using non-linear smooth terms to model relationships that are highly linear, categorical, or binary is not ideal, and an understanding of the features involved in the model is therefore important. Table 6.2 shows that most of the initially proposed features are properties related to positions and movement on the pitch, properties that are suspected to have non-linear effects. Most of the features included in the GAM will, therefore, be represented as non-linear smooth functions to capture these non-linear effects. Furthermore, many explanatory variables, X_1 and X_2 as an example, describe the same properties but in different directions. As these pairs of variables are associated together, they are represented as tensor products of smooth-terms. A representation of the features initially included in the GAMs are presented in Table 6.3, where $f_i(X_j, X_k)$ describes a 2-D

Variable	Description	Type
$f_1(X_1, X_2)$	Position of player, x and y-coordinate	2-D smooth
$f_2(X_3, X_4)$	Velocity of player, x and y-direction	2-D smooth
$f_3(X_5)$	Standard deviation of PC	1-D smooth
$f_4(X_6)$	Standard deviation of PI	1-D smooth
$f_5(X_7)$	Standard deviation of PC*PI	1-D smooth
$f_6(X_8, X_9)$	Position of ball, x and y-coordinate	2-D smooth
$f_7(X_{10}, X_{11})$	Position of team centroid, x and y-coordinate	2-D smooth
$f_8(X_{12}, X_{13})$	Position of opponent's centroid, x and y-coordinate	2-D smooth
$f_9(X_{14}, X_{15})$	Position of closest opponent, x and y-coordinate	2-D smooth
$f_{10}(X_{16}, X_{17})$	Position of optimal PC point, x and y-coordinate	2-D smooth
$f_{11}(X_{18}, X_{19})$	Position of optimal PI point, x and y-coordinate	2-D smooth
$f_{12}(X_{20}, X_{21})$	Position of optimal PC*PI point, x and y-coordinate	2-D smooth
$f_{13}(X_{22}, X_{23})$	Velocity of team centroid, x and y-direction	2-D smooth
$f_{14}(X_{24}, X_{25})$	Velocity of opponent centroid, x and y-direction	2-D smooth
$f_{15}(X_{26}, X_{27})$	Team compactness, x and y-direction	2-D smooth
$f_{16}(X_{28}, X_{29})$	Opponent compactness, x and y-direction	2-D smooth
$f_{17}(X_{30})$	Game time	1-D smooth
$f_{18}(X_{31})$	Passing angle	1-D smooth
X_{32}	Direction of pass (left or right)	Binary

Table 6.3: Features used to develop positional strategies GAMs.

smooth function of variables X_j and X_k.

Model setup

Three GAMs are created for the three different dependent variables presented in Table 6.1, referred to as GAM_{PC},GAM_{PI} and GAM_{PC*PI}. Observable from Figure 6.1, the dependent variables are bounded between 0 and 1, a fitting distribution for the dependent variable is the beta-distribution referenced in Section 3.4 described by Equation (3.20). The link function is set to the logit link function as it transforms the output to the interval $[0, 1]$. Equation (6.2) describes the logit function.

$$g(x) = ln\left(\frac{x}{1-x}\right) \tag{6.2}$$

The GAMs built in this thesis are created in the statistical programming language R, using the $mgcv$ library. Features, smoothing parameters, λ, and the number of splines are chosen using the methods described in Section 3.4.2. The shape parameters, α, β, of the beta distribution for the dependent variables are automatically estimated by the $mgcv$ library.

6.3.2 FFNN

The feed-forward neural network (FFNN), can be considered more of a black-box modelling approach, with analysis of the importance of individual features being difficult.

However, its handling of non-linear relationships between variables might provide accurate predictions, making it a desirable modelling approach for comparing and ranking individual players. A single network is created for modeling all 3 aspects, essentially functioning as a multivariate regression. This model with three outputs is referred to as $FFNN$. It is also possible to create three separate neural networks using each metric as an output, but as this requires fitting an additional two models, and initial results suggested it performed worse than the single network, so this approach is deemed unnecessary.

The structure of the problem indicates that a FFNN should be sufficient to model the dependent variables. As the situations are modelled from single snapshots of the game state, and not a sequence of game states, no feedback connections, found in recurrent neural networks (Sutskever et al., 2014), should be necessary to account for sequence-specific effects.

The FFNN is implemented in Keras. Keras is an open-source neural network library written in Python, running on top of Tensorflow, a Google developed open-source library for machine learning. The network is built only using the Dense layer class, with weights initialised using default Keras parameters.

Grid search

The FFNN has many tunable hyperparameters, which can have a large impact on the accuracy of a trained model. Tuning these hyperparameters is often necessary to improve accuracy. A series of problem-specific alterations is therefore made to the network architecture to enhance the model performance, with a combination of theory and testing forming the basis for decisions.

In the grid search, a large space of hyperparameter values is systematically explored on a validation set, with the objective of finding the right combination of hyperparameter values. The grid search builds a model for each parameter combination wanted to explore, searching for the combination giving the most accurate model. The combination space can be large, making this a time-consuming process. Some hyperparameter values are therefore fixed and are individually tested with the obtained combination of hyperparameters from grid search. The data set is split into a separate training and test set, consisting of 75% and 25% of the data. The model is trained with 100 epochs, storing the weights of the network from the epoch providing the lowest MAE for the validation set.

The values of four hyperparameters are tested using grid search and can be seen in Table 6.4. The network structure column represents the number of nodes and layers included in the network.

Grid search findings

Results from the grid search showed that the network is sensitive to changes in hyperparameter values. Larger network structures generally demonstrated better accuracy than the smaller ones. However, differences are small between networks with 512 nodes and 256

Batch size	Loss function	Learning rate	Network structure
32	MSE	0.001	64;64
64	MAE	0.0001	128;128
128		0.00001	256;256
			512;512
			256;256;256
			512;512;512
			256;256;256;256
			512;512;512;512

Table 6.4: Grid search parameters for FFNN.

nodes. As training time increases with increased network complexity, 256 nodes is chosen over 512, using four hidden layers. MSE loss function and 64 batch size is chosen with marginally better accuracy than their alternatives. The learning rate of 0.0001 is the preferred one, giving better results than 0.001. 100 epochs is not enough when the learning rate is set to 0.00001. This is therefore rejected as it is deemed too slow.

Additional hyperparameters and testing

Some additional hyperparameters are tested and set individually. The Adam optimiser is utilised, a popular optimiser choice in the field of deep learning. Adam's default parameters are used, except for the learning rate, using the one found from the grid search procedure. A linear decay factor is added to the learning rate, set to the learning rate divided by the number of epochs, making the learning rate approach zero as the number of epochs increases. This can ensure a more accurate training process, making smaller weight changes as the network approaches the lower loss limit. A linear activation function is applied to the first hidden layer, making this layer able to interpret negative contributions from specific features. The Rectified Linear Unit (Relu) is selected for the rest of the layers due to its good properties when the gradient is backpropagated. The Relu activation function has a lower bound on 0, an advantageous property as the output variable has the same lower bound. Relu lacks the upper bound found in the output variables, but from training is able to fit to the upper bound. Both L1 and L2 regularisation are also tested. However, they do not increase the generality of the model and slows the training process. They are therefore dropped from this model.

The final model is summarised in Table 6.5, with Figure 6.4 showing a training process. Using early stopping, the weights are stored at the epoch with the lowest loss on the validation set, illustrated with the dotted line.

Handling random weight initialisation

The weight initialisation when training a neural network is non-deterministic, meaning that predictions from two networks trained on the same data, using identical hyperparameters and structures, can exhibit different behaviours. The reason for this is that the networks

Hyperparameter	Value
Number of epochs	100
Learning rate	0.0001
Learning rate decay	0.0000067
Batch size	64
Loss function	MSE
Hidden layers structure	256;256;256;256
Hidden layers activation functions	Linear(1) and Relu(2-4)
Output layer activation function	Relu
Additional hyperparameters	Keras default values

Table 6.5: Final model FFNN

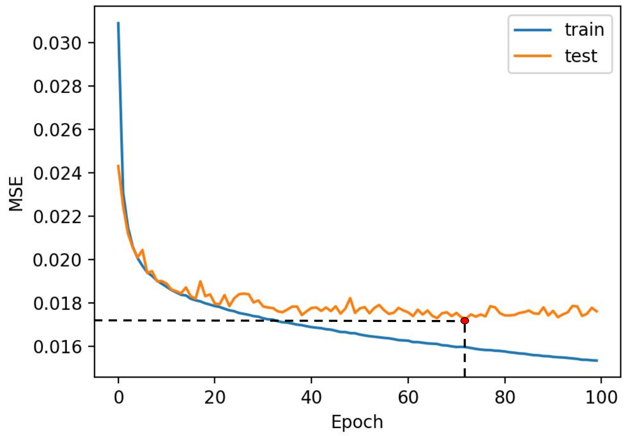


Figure 6.4: Training procedure with epoch 73 obtaining the lowest loss.

can be unfortunate with weight initialisation and get stuck in a local minimum during gradient descent. Instead of just using a single network, a total of 10 networks are trained on the same data, using the same hyperparameters. The median value predicted from the 10 models will form the prediction. The median is usually the preferred measure of central tendency when the distribution is not symmetrical.

6.4 Validation

After the models are developed according to the procedure described in this chapter they are validated on the accuracy of their predictions. With a continuous range for the dependent variable, a common validation method is to investigate the errors in model prediction on a test set. A test set of 25 % is selected randomly from the data and withheld from the fitting and training procedures. The different models' performance will be validated on this test set.

In this thesis three metrics derived from the prediction errors are used to validate the models. The four metrics are mean average error (MAE), root mean squared error (RMSE), mean bias error (MBE), and R squared (R^2), given by Equations (6.3-6.6) respectively.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6.3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6.4}$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{y}_i \tag{6.5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{6.6}$$

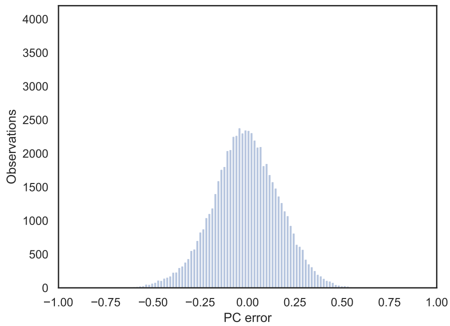
Here n is the number of predictions, $\hat{y}_i$ are model predictions and y_i are the actual values from the test set. Both the MAE and RMSE seek to evaluate how large on average the errors from the models are, without distinction between negative and positive errors. MAE and RMSE are therefore two metrics used to answer the same question but are calculated in different ways. As each individual error is squared in the RMSE, larger outliers of individual errors will have a stronger influence on the final value than they will for the MAE. If the absolute value of the model error is the same for all predictions in the test set the MAE and RMSE will be equal. Therefore, by calculating and comparing both the MAE and RMSE, an understanding of how the errors are distributed is included in the validation. The MBE is included to see if the errors are biased in one direction, or if the positive and negative errors cancel each other out on average. As positive errors cancel out negative ones, the MBE is a poor indicator of model precision and should not be compared with the results from the MAE and RMSE. If a model tends to error in its predictions in one direction, either predicting too high or too low results, the MBE will either have a positive or negative value. This information is not contained in the MAE or RMSE and is the reason for the inclusion of the MBE. The R^2 metric is included to see how much of the variance in the dependent variable that is explained by the explanatory variables.

6.5 Validation Results

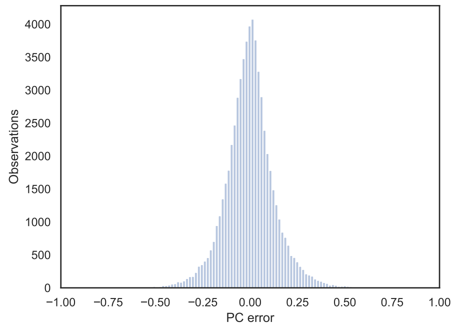
In this section, the results of the validation methods described in Section 6.4 are presented. For each dependent variable, the results from both the GAMs and FFNN are reported and compared. Plots of the different models error distributions are also included.

6.5.1 PC

The validation results for the dependent variable $Y(PC)$ from GAM_{PC} and the $FFNN$ is presented in Table 6.6. The table shows that for all validation metrics, the $FFNN$ performs



(a) GAM_{PC}



(b) FFNN

Figure 6.5: Error distributions for predictions of PC.

better than the GAM. The RMSE is larger than the MAE which indicates that all errors are not of the same magnitude, but the difference is not large enough to suspect that large outliers are mainly responsible for the errors. Figure 6.5 shows the error distributions for GAM_{PC} and $FFNN$, and the figure seems to support the assumption that the errors are normally distributed without many large outliers. Both the GAM_{PC} and $FFNN$ has a small negative bias, with the bias in the GAM_{PC} being almost six times that of the $FFNN$. Finally, $FFNN$ has a higher R^2 value than the GAM_{PC}.

Model	**MAE**	**RMSE**	**MBE**	$\mathbf{R^2}$
GAM_{PC}	0.1398	0.17778	-0.0087	0.4333
$FFNN$	0.0962	0.1312	-0.0016	0.6873

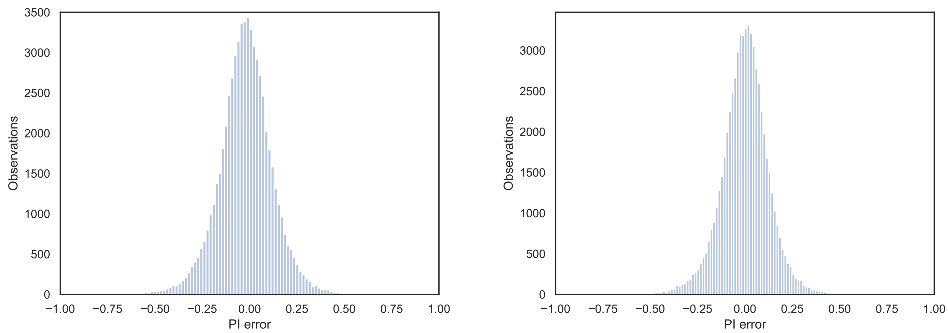
Table 6.6: Result of validation metrics for modelling $Y(PC)$.

6.5.2 PI

The validation results for the dependent variable $Y(PI)$ is presented in Table 6.7 and Figure 6.6 showing the error distributions. Again $FFNN$ performs better than theGAM_{PI} on all metrics, with lower MAE, RMSE and MBE, and a higher R^2. $FFNN$ performs quite similarly on $Y(PI)$ as $Y(PC)$ for the MAE and RMSE, but the MBE is better and the R^2 is worse. This means that the network makes errors of similar magnitudes for $Y(PC)$ and $Y(PI)$, but is able to explain more of the variance for $Y(PC)$ than $Y(PI)$. The GAM_{PI} is able to perform better according to the MAE and RMSE than GAM_{PC}, but as with the $FFNN$, the GAM_{PI} is able to explain less of the variance than GAM_{PC}. In contrast to the $FFNN$, the magnitude of the MBE for the GAM_{PI} is higher than for its PC counterpart, still having a negative bias.

Model	MAE	RMSE	MBE	$\mathbf{R^2}$
GAM_{PI}	0.1071	0.1411	-0.0201	0.3240
$FFNN$	0.0932	0.1230	-0.0002	0.4503

Table 6.7: Result of validation metrics for models using $Y(PI)$.



(a) GAM_{PI} **(b)** FFNN

Figure 6.6: Error distributions for predictions of PI

6.5.3 PC*PI

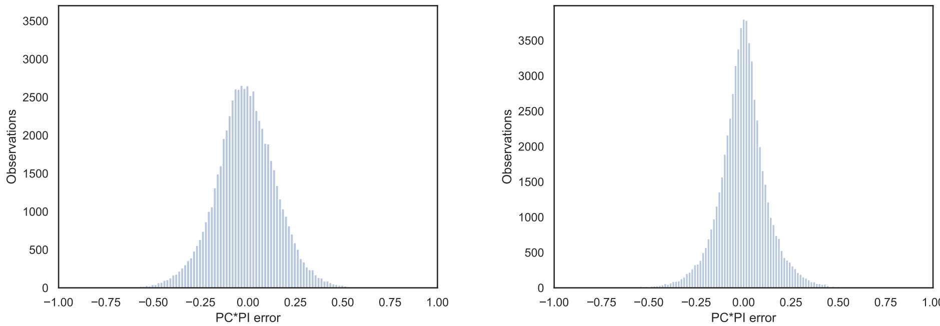
The validation results from the dependent variable $Y(PC * PI)$ is presented in Table 6.8, with error distributions shown in Figure 6.7. Again, both the GAM_{PC*PI} and $FFNN$ performs similarly on MAE and RMSE as their $Y(PC)$ and $Y(PI)$ counterparts. The bias is small and negative for both models, with the bias being five times the size for the GAM_{PC*PI} compared to the $FFNN$. Both models perform better according to the R^2 with values similar to their $Y(PC)$ counterparts.

Model	MAE	RMSE	MBE	$\mathbf{R^2}$
GAM_{PC*PI}	0.1251	0.1612	-0.0110	0.5095
$FFNN$	0.0937	0.1285	-0.0022	0.68215

Table 6.8: Result of validation metrics for models using $Y(PC * PI)$

6.6 Feature Contributions for GAM

In this section, some of the smooth functions estimated in the three GAMs are presented. As the shrinkage term for the smoothing penalties of some of the smooth functions is set by the fitting algorithm to be large, these smooth terms approach zero and their contributions to the final model are negligible. This section therefore only presents and discusses a subsection of interesting features with non-negligible contributions. Plots of all smooth terms can be found in Appendix B. Not all combinations of feature values are present in



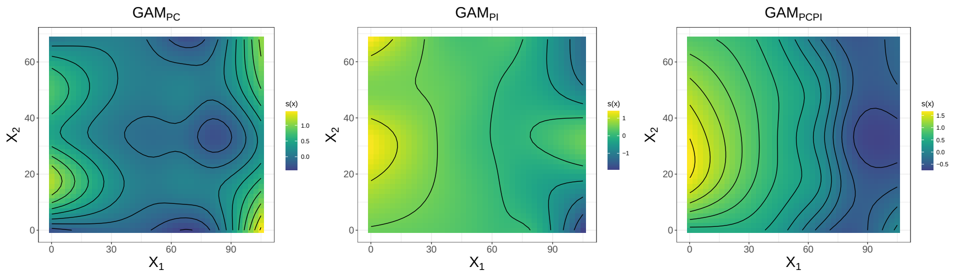
(a) GAM_{PC*PI} **(b)** FFNN

Figure 6.7: Error distributions for predictions of PC*PI

the set of observations used to train the models, and in the figures below showing two dimensional smooth surfaces, the areas without training observations are showed as gray. The first feature presented is the contributions towards the three dependent variables by the player's starting position, in x- and y-direction. The contributions of this feature, $f_1(X_1, X_2)$ is shown in Figure 6.8 as a two dimensional smooth surface. Generally, the smooth functions seem to suggest that initial positions further from the goal are favourable in terms of performing better for all three dependent variables. The GAMs offers no explanation why, but one possible explanation could be that situations, where players are closer to the opponents goal, are more complex, making positional decisions more difficult. As players move closer to the goal, less space is generally available, and more teammates and opponents will probably be in proximity, making positional decisions more complex, as accurately predicting all these players' future positions get more difficult.

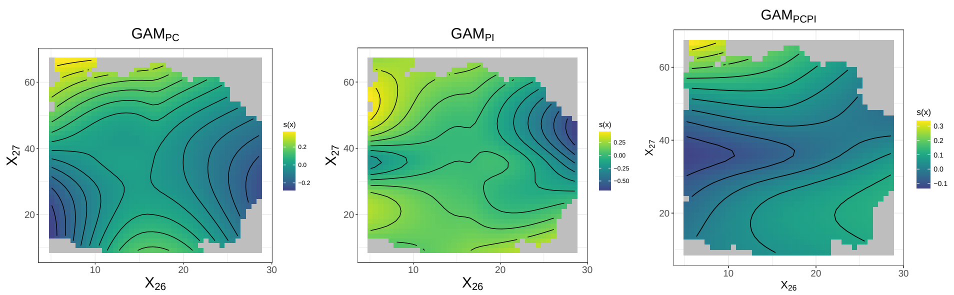
The smooth surfaces $f_{15}(X_{26}, X_{27})$ and $f_{16}(X_{28}, X_{29})$ shown in Figures 6.9 and 6.10 shows the contributions from team and opponent compactness respectively. These smooth surfaces indicate that low compactness, meaning closer distances between players, has a negative contribution towards the dependent variables, especially in the x direction for the opponent compactness. This somewhat supports the theory for why situations closer to the goal are harder to read as a result of the area around the player being more congested.

Figure 6.11 shows the contribution of the players initial velocity towards the dependent variable as a 2-D smooth surface $f_2(X_3, X_4)$. No definite conclusions can be drawn regarding velocity in the y-direction, but the figures seem to show that opposite relationships exist for PC and PI in x-direction. Having an initial velocity away from goal, represented by a negative X_3-value, seems to be preferable for attaining the highest amount of PC while velocity towards goal is preferable for PI. For PI this result is quite intuitive as the PI metric is generally increasing towards goal and initial movement in this direction should be an indicator that the players' general movement is also in this direction.



(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

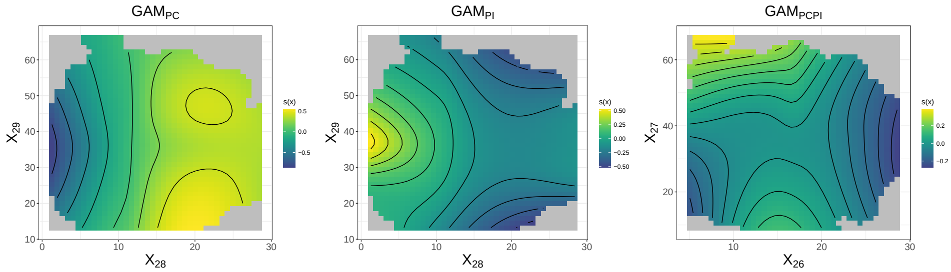
Figure 6.8: Heat map showing the contribution of smooths for the position of the player, $f_1(X_1, X_2)$



(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

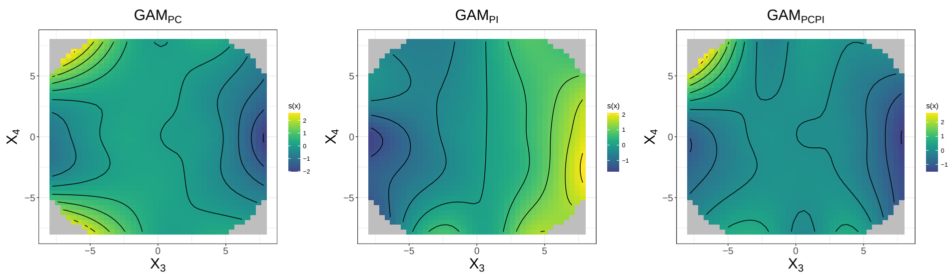
Figure 6.9: Heat map showing the contribution of team compactness, $f_{15}(X_{26}, X_{27})$

Another selected feature is the vector from the player to the ball position, $f_6(X_8, X_9)$. In football, the center of attention is often the ball itself, and the position a player has relative to the ball is therefore often related to the player's objective in that situation. When Hammarby IF uses PC, PI, and PP to analyse situations, players are assigned objectives based, among other things, on their position relative to the ball (Peralta Alguacil, 2019). Investigating through the use of the GAMs if actual player behavior is related to position relative to the ball is therefore interesting. Figure 6.12 shows the contribution from the relative ball position on the three dependent variables as two dimensional smooth surfaces. One general observation from these smooth surfaces is that when the distance to the ball is small, meaning that the ball is close to the player, the contribution towards all dependent variables seems to be small or negative. This result is and in line with previous results from this section. As opponents generally have some attraction towards the ball, the area around the ball will often be more crowded than areas far from the ball. This increases complexity in areas close to the ball and following the same reasoning as previously discussed, increases the difficulty of positional decision making.



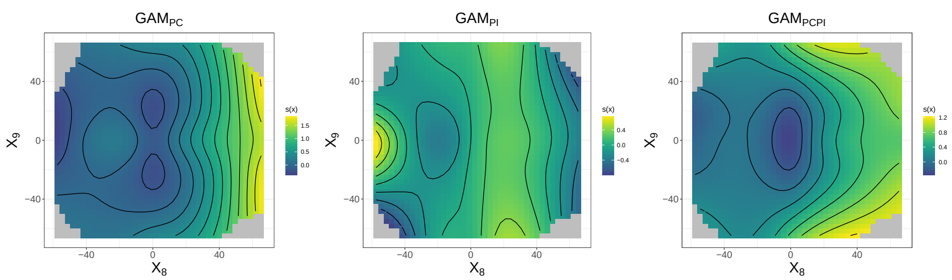
(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

Figure 6.10: Heat map showing the contribution of smooths for the opponent compactness, $f_{16}(X_{28}, X_{29})$



(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

Figure 6.11: Heat map showing the contribution of smooths for the velocity of the player, $f_2(X_3, X_4)$



(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

Figure 6.12: Heat map showing the contribution of smooths for position of ball, $f_6(X_8, X_9)$

6.7 Discussion

Before moving on with the applications of the models, a discussion of the models is presented.

The models introduced in this chapter are seemingly unique in the way the dependent variables are defined. Therefore, it is difficult to relate the results to excising work. Predicting decision making in football is a complex problem, showing signs of having stochastic elements. Therefore, none of the models in this thesis are expected to perfectly model player behaviour, which is also the case. However, they are able to capture a lot of trends and tendencies. By assuming that player performance is mean reverting, these models can provide insights into how players perform by using a number of observations to form the basis for evaluation.

Players do on many occasions show non-deterministic behaviour, a hypothesis that will be further investigated in Chapter 8. Without knowing the exact amount of randomness inherent in the problem, it is difficult to assess if the prediction models could be significantly improved, or if they are close to the limit of possible accuracy. However, inherent problems with the data quality suggest that the models could be improved. As mentioned in Section 5.3, the data is in some occasions unreliable, leading to the models being fitted and trained on some noise and missed features. Increased reliability of the data would also open up for generating additional features describing the game state, such as physical fatigue from distance traveled and intensity over a previous period. As such features are generated from accumulating movement, the problem with switching and missing player positions have to be dealt with before this is can be done reliably, a problem outside the scope of this thesis.

A feature believed to increase the accuracy of the models is the direction the player faces, now only included as a proxy from the direction of the velocity of the player. This feature could capture body direction relative to the movement, possibly providing useful information on the true focus of the player. Such a feature could be obtained using a radar-based tracking system, with players having two separate sensors positioned horizontally on their bodies. The National Football League (NFL), the top professional league in the sport of American football, uses a system where players wear one sensor on each shoulder. Burke (2019) includes the direction a player faces as a feature in the feed forward neural network developed to investigate passing in the NFL.

An increased number of observations increases model accuracy. This accuracy moves towards a limit, an effect discovered from training the models on data sets of different sizes. Increasing the number of observations to model the situation described in Section 4.2.1 is therefore believed to increase the accuracy slightly, but not by a lot. This is relevant when moving to the application of the models, as the dilemma of using many observations and computational expenses related to creating models appears. This dilemma is further addressed in Section 7.2.2.

Random effects could be introduced as parts of the models, increasing the ability to model individual players. This could be of use to assess the player with itself, i.e. a player's single-game performance in relation to the player's general performance. For the GAM, looking at the coefficients generated to account for random effects could be used to compare individual player performance. This is however not so easy for the FFNN, as there are no coefficients present to compare. For this thesis, the scope is to assess the general overall performance of players by looking at how players perform in relation to each other, and not necessarily to accurately predict individual player performance. Therefore, an alternative way for rating players is introduced, later explained in Section 7.2.1.

Following the validation results presented in Section 6.5, the FFNN is better at predicting how players make decisions for all three dependent variables. With a lot of variables seeking to describe the situation, the FFNN seems to be better at understanding the complex dynamics that influence players in the modelled situations. This is believed to come from the model's ability to interpret non-linear relationships between all variables, and not only from non-linear contributions of selected variable pairs, which is the case for the GAM. As the FFNN has better precision and less bias, the predictions from this model is deemed to be more reliable than from the GAM. This is also backed by the R^2, supporting the notion that the FFNN is able to capture more of the variance in all three dependent variables than the GAMs. This is an important finding for this thesis, as moving forward to Chapter 7, the models are used to assess the PC, PI, and PC*PI performances of both individual players and teams. The conclusion is that the FFNN is more suitable for further analysis than the GAM for all three dependent variables. Therefore, only the FFNN will be used in applications of the models presented in Chapter 7.

Chapter 7

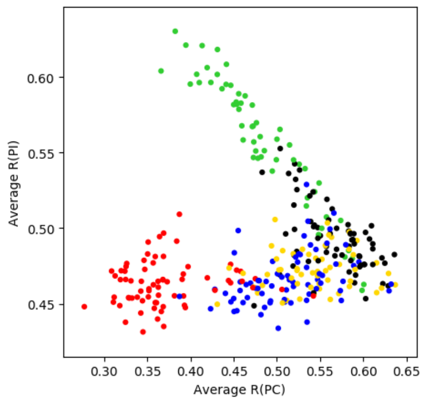
Evaluating Positional Strategies

The following chapter presents applications of the scores of the metrics from Chapter 4, and the models developed in Chapter 6, seeking to evaluate positional strategies. First, an investigation of player roles and positional strategies is presented, followed by a method for ranking individual players on how well they perform relative to the optimal position. Then comes an analysis of relations between team ratings and attacking success. The last part of this chapter is a case study of positional strategies for two individual players. From the validation results and discussion presented in Sections 6.5 and 6.7, the FFNN is used in all model applications presented in this chapter.

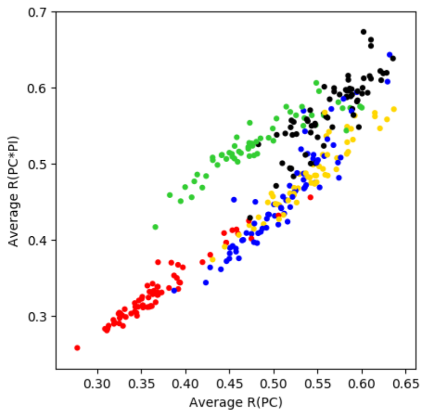
7.1 Positional Strategies Related to Player Role

The following section investigates how player roles relate to positional strategies on the pitch. As previously mentioned in Section 5.5, all players in the data are assigned a role according to their most played role throughout the season and Table 5.2 shows the six different roles the players are divided into. It is important to note that this approach does not capture the fact that some players can be assigned different roles during the season. Therefore, the assigned player roles are only indicative of the role a player has during the season.

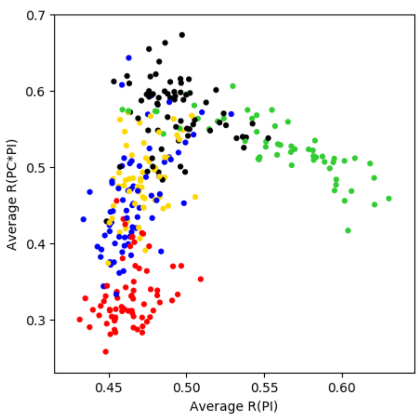
Figure 7.1 shows the average achieved value of the PC, PI and PC*PI score $R(m, j)$, as described in Section 4.1.3, for each player with over 100 observations during the 2019 Allsvenskan. Each dot represents a single player and the different colors of the dots represent the different roles that are assigned to the players. The figure allows for a comparison of how different player roles perform in relation to the three metrics. When interpreting the figure, it can be observed that FWs achieve higher average values for PI than the other players. Remembering from Section 3.1.3, that the PI model is a goal probability model, it is intuitive that FWs outperform the other players as their main responsibilities on the pitch usually is to create chances and goals.



(a) $R(PC)$ and $R(PI)$



(b) $R(PC)$ and $R(PC * PI)$



(c) $R(PI)$ and $R(PC * PI)$

Figure 7.1: Average PC, PI and PC*PI scores for individual players with colours representing the role of the player. Colours: ● - CB, ● - MF, ● - FB, ● - FW, ● - AM

Another notable finding is that CBs seem to underperform in all of the three metrics compared to the other roles. An explanation for this can be the fact that CB's main task on the field is not necessary to focus on maximising any of the three metrics. CB's assignment is to be the last line of defense between the opponent and their own goalkeeper, limiting their ability to partake as actively in attacking situations. In a sense, their main responsibility is not to optimise performance regarding the three metrics, but rather minimise these same metrics for the opposing players. This result is especially interesting when contrasting the CBs to the FBs. One could presume that FBs would perform similarly to CBs, but it appears that they in fact score closer to MFs in regard to the metrics. The modern FB is for many teams an important attacking resource, responsible for providing width and attacking support down the sidelines. In this context, it is not surprising that FBs seem to outperform CBs, as the FB role is considered to be more attacking focused than the CB role in the modern game of football.

When comparing the AMs to the MFs, it appears that AMs score higher in all three of the metrics. As AMs play in a more advanced position, it is reasonable that they perform higher on PI as it primarily measures attacking contribution. The fact that AMs also perform better in PC and PC*PI can be related to the fact that MFs also have defensive duties and are not able to focus as much as AMs on gaining control.

7.2 Ranking Individual Players

This section presents a method for using the prediction models created in Chapter 6 to rate and rank individual players on performance regarding PC, PI, and PC*PI. Firstly, an explanation of the two methods for ranking players is given, followed by a presentation of how the data is split to perform out of sample validation for all observations. Then, the results showing the top performers according to the ratings are presented, and lastly, the ratings produced in this section are compared to relevant off-ball ratings from the video game Football Manager 2020.

7.2.1 Metrics for ranking individual players

From the metrics presented in Section 3.1 it is possible to rank individual players on how they perform on the chosen criteria. Summing up or averaging the rating achieved $R_i(m, j)$, by player i in observation j using metric m, without using a prediction model is one possible way of rating a player's performance on off-ball metrics. This method is relatively simple as it only requires averaging the $R_i(m, j)$ on all observations j of the player i, resulting in what is referred to as the average actual rating, shortened to $AAR_i(m)$.

Comparing AAR performance across different situations could lead to an incomplete description of the true performance, as the difficulty and incentive of obtaining a high AAR are presumed to be dependent on the situation. By using prediction models, this problem can be solved by predicting performance for the situation using variables describing the situation, and then compare this to the true performances of the player. A player can, therefore, be evaluated on his performance in a specific situation compared to how the models expect other players to perform in the same situation. It can also be used on an entire season of situations, which is what has been done in this section. Then the aggregated difference in performance between the chosen player and the rest of the league, or other individual players, can be evaluated. With $R_i(m, j)$ as the rating player i achieves in situation j and $R_{predicted}(m, j)$ as the rating predicted by the model in the same situation, the average situational adjusted rating $ASAR_i(m)$ is given by Equation (7.1).

$$ASAR_i(m) = \frac{1}{N} \sum_{j=1}^{N} (R_i(m, j) - R_{predicted}(m, j)) \tag{7.1}$$

Both the top 10 performers in $AAR_i(m)$ and $ASAR_i(m)$ are presented in this section. Even though the $AAR_i(m)$ does not account for the characteristics of the situation, it is an interesting rating as it rewards players who are able to put themselves in favourable situations where high $R_i(m, j)$ values are easier to obtain. Some characteristics of these types of favourable situations are discussed in Section 6.6.

7.2.2 Training a prediction model for the situational adjusted ratings

When evaluating a player using a prediction model, the more observations included in the training process of the model will give a better evaluation of how the player compares to general expected performance. One approach would be to create separate models for all players, training and validating the model on all observations except for the ones associated with the player, and use this model to evaluate the player. However, with over 300 models to be created, one for each player, this is a computationally expensive procedure and therefore rejected. Instead, an alternative approach is used. All observations are shuffled and separated into five equally large groups. Then, a model is trained and validated using observations from four of the groups. This model is then used to predict the observations in the remaining group. A total of five models, one for each group, will therefore have to be created. Using this approach, a lot of observations can be used to train and validate the models, without having to train 300 models. A potential problem is that a player's observation is predicted using a model trained on the player's own observations, as the player's observations can be present in all five groups. The model could therefore potentially learn the player's tendencies and take this into account in the predictions. However, as each player only represents a small fraction of the total observations, the effect from this is deemed negligible. Further, an advantage of this is that all players are evaluated on all models. This way, potential biases of individual models will not influence the final ranking of players.

7.2.3 Ranking results

To ensure that only players that frequently play and contribute to their teams success are included, a threshold of 500 observations was set to qualify for the rankings. Using this threshold, 201 players qualified for the rankings and the top 10 players for both $AAR_i(m)$ and $ASAR_i(m)$ for $m = \{PC, PI, PC * PI\}$ is presented. Tables 7.1, 7.2 and 7.3 shows the top 10 performers in $AAR_i(m)$.

One observation made from the tables of AAR is the role of the top performers in each category. While AMs and FBs seem to obtain the highest actual PC values, PI is dominated by FWs with the combined PC*PI metric again dominated by AMs. These results are quite expected, and in line with the results from the analyses done in Section 7.1. Looking at the individuals represented in the Tables 7.1-7.3, 6 of the top 10 PC achievers are also top 10 in PC*PI, while zero of the top PI achievers are able to reach the top 10 in PC*PI. This could indicate that the players that are most focused on PI tend to neglect PC to a higher degree than the opposite. One explanation could be that forwards have more specialised roles in attacking situations than AMs. A FWs primary task is, as mentioned in Section 7.1, to create shots and goals. Therefore, employing a more single-minded approach of

#	Player	Team	Role	Obs	AAR(PC)
1	Romain Gall	MFF	MF	633	0.6321
2	Kevin Adrian Wright	ÖSK	FB	1475	0.6291
3	Giorgi Kharaishvili	IFK	AM	1204	0.6288
4	Ahmed Yasin	BKH	AM	1478	0.6248
5	Felix Beijmo	MFF	FB	769	0.6211
6	Søren Rieks	MFF	FB	2290	0.6153
7	Max Svensson	HIF	AM	1968	0.6109
8	Nasiru Mohammed	BKH	AM	534	0.6072
9	Johan Blomberg	GIF	AM	1018	0.6027
10	Oliver Berg	GIF	AM	1427	0.6014

Table 7.1: Top 10 performers $AAR(PC)$.

#	Player	Team	Role	Obs	AAR(PI)
1	Vidar Örn Kjartansson	HAM	FW	992	0.6305
2	Carlos Strandberg	ÖSK	FW	910	0.6209
3	Marc Mas Costa	GIF	FW	670	0.6183
4	Kolbeinn Sigþórsson	AIK	FW	585	0.6063
5	Guillermo Molins	MFF	FW	1110	0.6019
6	Linus Hallenius	GIF	FW	671	0.5965
7	Nsima Peter	FFF	FW	522	0.5964
8	Alhaji Gero	HIF	FW	860	0.5952
9	Aron Johansson	HAM	FW	505	0.5875
10	Mohamed Buya Turay	DIF	FW	1904	0.5835

Table 7.2: Top 10 performers for $AAR(PI)$.

#	Player	Team	Role	Obs	AAR(PC*PI)
1	Max Svensson	HIF	AM	1968	0.6627
2	Romain Gall	MFF	MF	633	0.6430
3	Giorgi Kharaishvili	IFK	AM	1204	0.6192
4	Ahmed Yasin	BKH	AM	1478	0.6185
5	Maic Sema	GIF	AM	1186	0.6155
6	Johan Blomberg	GIF	AM	1018	0.6120
7	Paulo De Oliveira	BKH	AM	1059	0.6100
8	Oliver Berg	GIF	AM	1427	0.5996
9	Muamer Tankovic	HAM	AM	2231	0.5977
10	Francisco Wánderson	HIF	AM	1042	0.5969

Table 7.3: Top 10 performers $AAR(PC * PI)$.

#	Player	Team	Role	Obs	AAR(PC)	Predicted	**ASAR(PC)**
1	Alexander Farnerud	HIF	MF	736	0.5452	0.5232	**0.0219**
2	Nasiru Mohammed	BKH	AM	534	0.6072	0.5868	**0.0203**
3	Max Svensson	HIF	AM	1968	0.6109	0.5925	**0.0184**
4	Felix Beijmo	MFF	FB	769	0.6211	0.6039	**0.0171**
5	Viktor Lundberg	BKH	AM	837	0.5957	0.5792	**0.0165**
6	Tarik Elyounoussi	AIK	MF	1231	0.5948	0.5783	**0.0165**
7	Daleho Irandust	BKH	AM	1608	0.5957	0.5807	**0.0151**
8	Søren Rieks	MFF	FB	2290	0.6153	0.6005	**0.0148**
9	Adi Nalic	AFC	FW	845	0.5873	0.5728	**0.0146**
10	Elias Andersson	IKS	MF	524	0.5765	0.5620	**0.0145**

Table 7.4: Top 10 PC performers $ASAR(PC)$.

#	Player	Team	Role	Obs	AAR(PI)	Predicted	**ASAR(PI)**
1	Kolbeinn Sigþórsson	AIK	FW	585	0.6063	0.5689	**0.0373**
2	Carlos Strandberg	ÖSK	FW	910	0.6209	0.5852	**0.0357**
3	Guillermo Molins	MFF	FW	1110	0.6019	0.5675	**0.0344**
4	Vidar Örn Kjartansson	HAM	FW	992	0.6305	0.5990	**0.0314**
5	Markus Rosenberg	MFF	FW	1707	0.5589	0.5281	**0.0307**
6	Marcus Antonsson	MFF	FW	1592	0.582	0.5519	**0.0302**
7	Marc Mas Costa	GIF	FW	670	0.6183	0.5884	**0.0299**
8	Per Frick	IFE	FW	760	0.5814	0.5555	**0.0260**
9	Nikola Djurdjic	HAM	AM	1812	0.5385	0.5136	**0.0249**
10	Alhaji Gero	HIF	FW	860	0.5952	0.5704	**0.0248**

Table 7.5: Top 10 performers $ASAR(PI)$.

#	Player	Team	Role	Obs	AAR(PC*PI)	Predicted	**ASAR(PC*PI)**
1	Nasiru Mohammed	BKH	AM	534	0.5931	0.5663	**0.0267**
2	Max Svensson	HIF	AM	1968	0.6627	0.6385	**0.0242**
3	Adi Nalic	AFC	FW	845	0.5731	0.549	**0.0241**
4	Deniz Hümmet	IFE	FW	745	0.5634	0.5425	**0.0209**
5	Paulo De Oliveira	BKH	AM	1059	0.6100	0.5893	**0.0208**
6	Romain Gall	MFF	MF	633	0.6430	0.6227	**0.0203**
7	Francisco Wánderson	HIF	AM	1042	0.5969	0.5772	**0.0197**
8	Tarik Elyounoussi	AIK	MF	1231	0.5937	0.5757	**0.0180**
9	Isak Magnusson	KFF	FB	625	0.5456	0.5286	**0.0170**
10	Alexander Kacaniklic	HAM	AM	1753	0.5845	0.5677	**0.0169**

Table 7.6: Top 10 PC*PI performers $ASAR(PC * PI)$.

obtaining a dangerous position appears to be prioritised for players in that role. The job of an AM is on the other hand oftentimes more diverse. Often relied on to be on both the creative and receiving end of opportunities, it is natural that these types of players are focused on both PC and PI.

The top 10 performers in situational adjusted rating, $ASAR$, for PC, PI and PC*PI are shown in Tables 7.4 - 7.6. Comparing the results from Table 7.1 and 7.4, four players make it into the top ten on both actual and adjusted ratings, with the top ten in adjusted rating consisting of a mix of four different positional groups compared to three for actual rating. Looking at the PI ratings from Tables 7.2 and 7.5 there are six names common to both top tens, but unlike the $AAR(PC * PI)$, the top ten for $ASAR(PC * PI)$ includes a non-forward in the attacking midfielder Nikola Djurdjic at number nine. As with $AAR(PC * PI)$, there are more similarities between the top ten performers in PC and PC*PI, four players making the top ten for both metrics, than PI and PC*PI, with no players making both top ten lists. Four players are also able to make the top ten in both $AAR(PC * PI)$ and $ASAR(PC * PI)$. The greater diversity of roles in the top ten lists for the $ASAR$ metric may be the result of this metric being able to adjust for how players in different roles often encounter different situations.

7.2.4 Comparing results to ratings in Football Manager 2020

In this subsection, a comparison between the results from the AAR and $ASAR$ ratings created in this section and a few selected attributes from FM20 have been made. As FM20 evaluates many different attributes for the players, it is necessary to only include attributes that are comparable to the ratings. As the AAR and $ASAR$ ratings created in this thesis seek to measure off-ball decision making, similar attributes from FM20 were identified. The following five attributes were chosen, accompanied by a brief explanation of how FM20 scores the attributes (guidetofm.com).

1. **Aggression:** How likely a player is to choose to get involved in a physical situation and how much he exerts physical force in such situations.
2. **Anticipation:** How well a player can predict the movements and other actions of his teammates and opposition players.
3. **Decisions:** How well a player can evaluate the options he is aware of and choose which action to perform, when to perform it and how to perform it.
4. **Off the Ball:** How well a player moves and positions himself, to either provide a passing option or create space for teammates to exploit, when he is off the ball and his team is in possession.
5. **Work Rate:** How much physical effort a player puts into his actions during a match.

When analysing the correlations between the FM20 attributes and the achieved player rating in Figure 7.2 the most intriguing result is the correlation coefficients for the *Off the Ball* attribute. The correlation between the FM20 off the ball attribute and the $ASAR$ statistics for PI and PC*PI is to be considered moderate and for PC it is considered weak

Figure 7.2: Heatmap that shows the Pearson correlation between five selected player attributes ratings in FM20 against the $ASAR$ and AAR values for the PC, PI and PC*PI ratings achieved by the 201 qualified players.

(Senthilnathan, 2019). This result shows that there is a noteworthy positive correlation between FM20's experts' qualitative assessments of the *Off the Ball* attribute of the players in Allsvenskan and the quantitative ratings presented in this thesis. This correlation appears to be stronger for AAR than $ASAR$, meaning that FM20's off the ball attribute has more similarities to the actual ratings the players achieve rather than their situation adjusted ones. This is an interesting observation, but the question remains if this is the result of FM2020 not properly adjusting for the situation when they evaluate players, or that being able to find favourable situations during attacking passing events are more important than performing above expectations during those situations. The remaining FM20 stats all correlate weakly with at least one of the ratings, but there appears to be no noteworthy correlation between the ratings created and the remaining FM20 attributes.

7.3 Relation Between Positional Metrics and Attacking Success

In this section, the relation between the metrics introduced and goal-scoring success is investigated. The teams of the 2019 season of Allsvenskan are divided into 4 categories based on goals scored throughout the season. The FFNN developed in Chapter 6 is used to analyse differences in how the different groups perform according to the three off-ball positional metrics. The FFNN is trained and tested on three groups to create a model that will be used to predict the observations for the remaining group. This is done to isolate individual effects present in the remaining group, preventing the model from accounting for these effects. Distinct performance of a group can be identified when actual performance does not match the model's predictions. The four groups and teams included in each are shown in Table 7.7. The $ASAR_i(m)$ metric proposed in Section 7.2 is used to identify distinct

Group number	Teams in group	Goals scored
Group 1	Hammarby IF	75
	Malmö FF	56
	FK Norrköping	54
	Djurgården	53
Group 2	AIK	47
	IFK Göteborg	46
	BK Häcken	44
	IF Elfsborg	44
Group 3	Örebro SK	40
	IK Sirius	34
	GIF Sundsval	31
	Helsingborgs IF	29
Group 4	Östersunds FK	27
	Falkenbergs FF	25
	AFC Eskilstuna	23
	Kalmar FF	22

Table 7.7: The four groups with their associated teams and goals scored.

performance for the groups $i = \{1, 2, 3, 4\}$, using metrics $m = \{PC, PI, PC * PI\}$.

During attacks, different positional groups often have different roles and are instructed by coaches to behave differently. Forwards may be instructed to try to take up positions closer to the goal where they hope to eventually receive the ball, while defenders may be instructed to control the opposing attackers as a preventive measure in case of a turnover. To further investigate the details of distinct performance, players are split into two groups based on their level of responsibility in the case of a turnover, forming the attacking group and the midfield/defender group. Following the assigned player roles described in Section 5.5, FWs and AMs are classified as attacking players, while MFs, FBs, and CBs are classified as midfielders/defenders. This leads to the creation of eight prediction models used to rate the four groups on performance for attacking and midfield/defending players.

To give additional insight into the four groups' performance, a map is presented of the attacking half of the pitch, with areas showing where the groups perform better or worse than what is expected.

7.3.1 Performance in the four groups by attackers

Attackers performance, PC

Figure 7.3 shows how the different groups perform with respect to the PC metric. Interestingly, it is Group 3 that seems to perform best according to the FFNN, followed by Group 2 and Group 4 with Group 1 performing the worst. As attackers on high scoring teams seem to neglect PC more than lesser teams, it may be the case that teams are better off instructing their attackers not to focus as much on PC. No positional characteristics for the different groups can with certainty be detected from the plots, as there seem to be no clear areas where specific groups perform notably different from their overall performance.

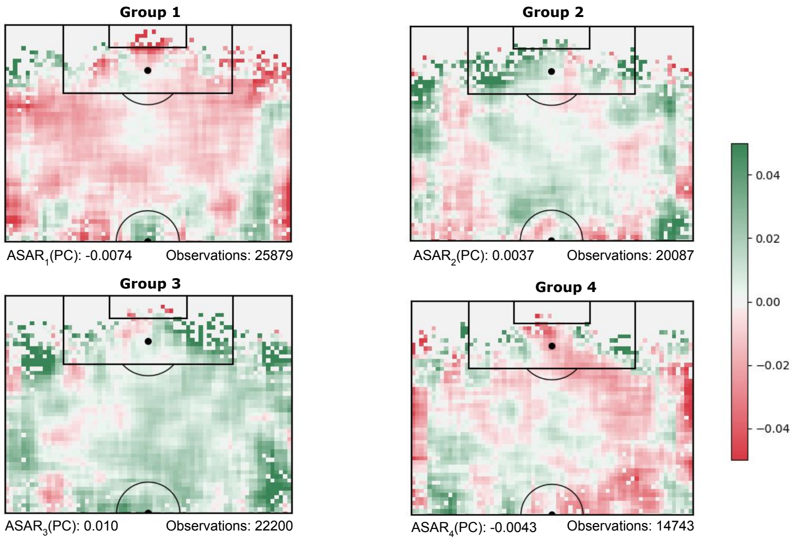


Figure 7.3: PC performance by attackers.

Attackers performance, PI

The PI performance of the different groups is shown in Figure 7.4. Using this metric, Group 1 performs best followed by Group 3, Group 4, and finally Group 2. With the exception of Group 2, the results seem to indicate some connection between PI performance and attacking success. Instructing attackers to focus mainly on getting into dangerous positions may, therefore, be more important for attacking success than teaching attackers how to control space.

Attackers performance, PC*PI

Figure 7.5 shows the performance of the four groups on the PC*PI metric. Similarly to the performance according to the PC metric, Group 3 performs best, now followed by Group 2, Group 1, and lastly Group 4. As with the results from other sections in this thesis, there seems to be more similarity between the results using the PC metric and PC*PI metric than PI and PC*PI.

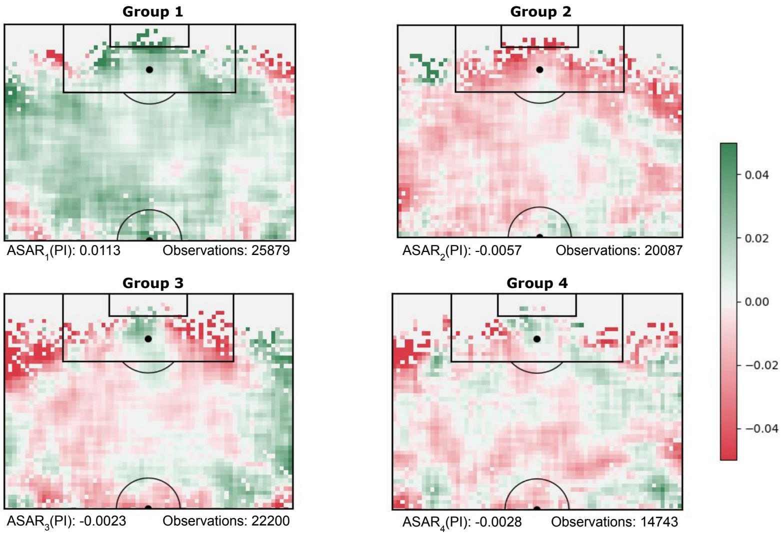


Figure 7.4: PI performance by attackers.

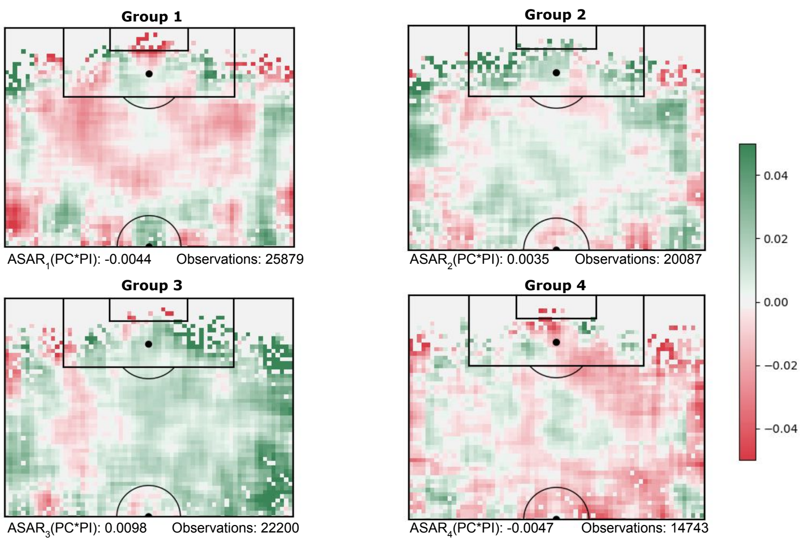


Figure 7.5: PC*PI performance by attackers.

7.3.2 Performance of the four groups by midfielders/defenders

Midfielders/defenders performance, PC

Figure 7.6 shows how midfielders in the two groups perform on the PC metric. Now the best performers are Group 1 followed by Group 2, Group 4, and Group 3. Interestingly, midfielders and defenders on good attacking teams seem to perform better at PC than their attackers relative to the other groups. This supports the hypothesis that good teams tend to differentiate more in positional strategies between different roles.

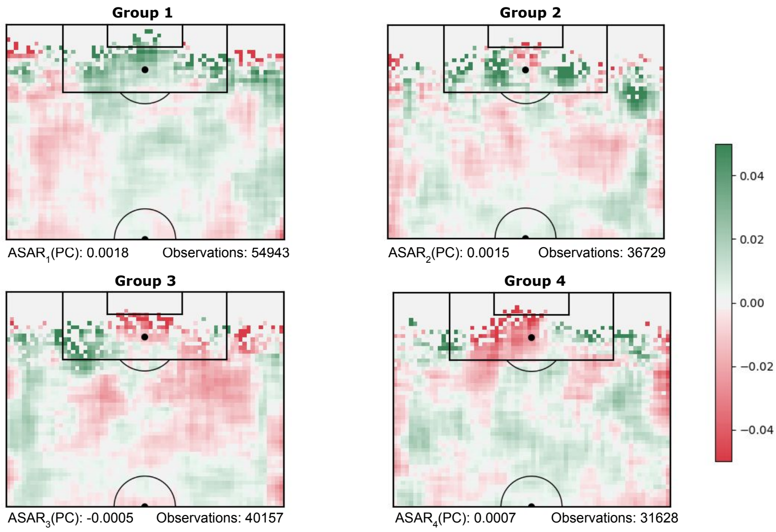


Figure 7.6: PC performance by midfielders/defenders.

Midfielders/defenders performance, PI

The performance of midfielders and defenders on PI is shown in Figure 7.7. Again Group 1 is the top performer, now followed by Group 4 with Groups 2 and 3 performing equally. In terms of ranking the different groups, the results for midfielders and attackers are quite similar to the results for attackers. Looking deeper into the results, one observation regarding Group 1 is quite interesting. Midfielders and defenders seem to perform well in most positions except positions close to and inside the penalty box. When getting close to goal, midfielders and defenders on the best attacking teams, seem to resist pushing further upwards, leaving that responsibility to their attackers which as shown in Figure 7.4 performs well in these areas.

Midfielders/defenders performance, PC*PI

Finally, Figure 7.8 shows how the four groups of midfielders and defenders perform on PC*PI. Using this measure Group 4 performs best with Group 1 in second, Group 3 third, and Group 2 in fourth. Even though Group 1 in total performs worse than Group 4, midfielders and defenders in Group 1 seem to perform very well in areas inside the penalty box for this metric. Other than that, no other significant positional trends are clearly visible from the figure, with no group seeming to perform much better than the others in clearly defined areas outside the penalty box.

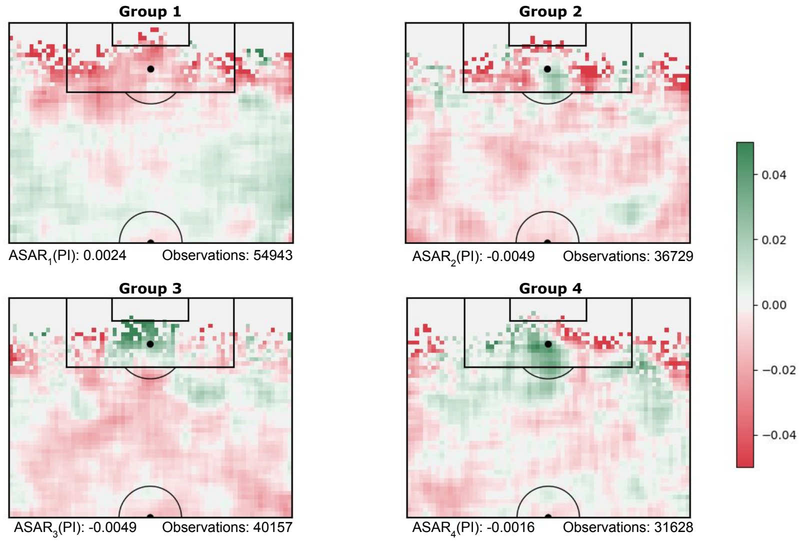


Figure 7.7: PI performance by midfielders/defenders.

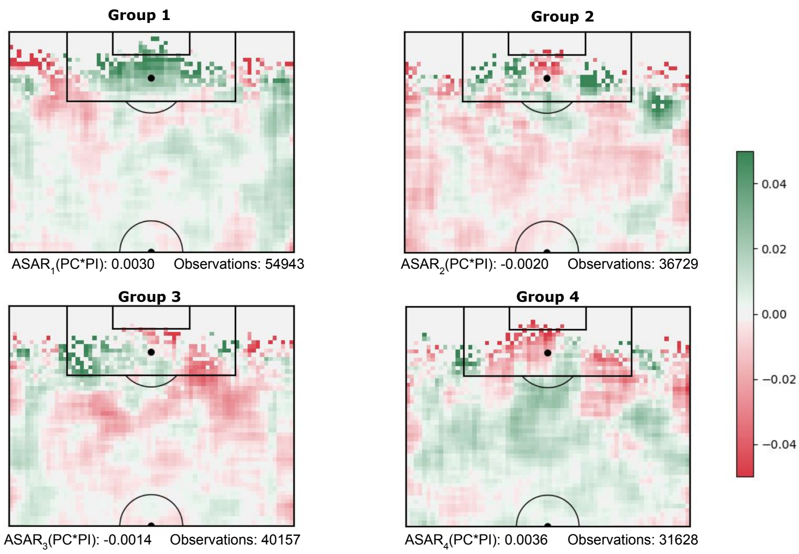


Figure 7.8: PC*PI performance by midfielders/defenders.

7.4 Analyses of Individual Player Performance

In this section, a detailed study of the performance of two players is conducted following the framework presented in Section 7.3. Individual player performance will be studied using a visualisation of location specific performance.

Muamer Tanković and Alexander Kacaniklic are chosen to be analysed in this section, both playing Attacking Midfielders for Hammarby IF during the 2019 season. They both have a high number of observations, a requirement for creating a map to illustrate pitch specific performance. To model their performance in relation to expected performance, all observations from Hammarby IF are withheld from the data set that the model is trained and validated on. This way, their performances are compared to the general performance of all teams in Allsvenskan.

7.4.1 Muamer Tanković

Muamer Tanković is an AM who played a total of 28 games and scored 14 goals, making him the second most scoring player in Allsvenskan during the 2019 season. 2231 observations are included to assess his off-ball performance. This is illustrated in Figure 7.9. Notable is his general performance above expectation on all metrics. Still, his performance is location specific, with some locations showing under-performance. This is especially the case for PI closer to the goal. This means that he is less active in pressing when he is closer to goal compared to others. As Tanković is one of the highest goal scoring players in Allsvenskan, this shows that dangerous players not necessarily have to press more than expected close to the goal.

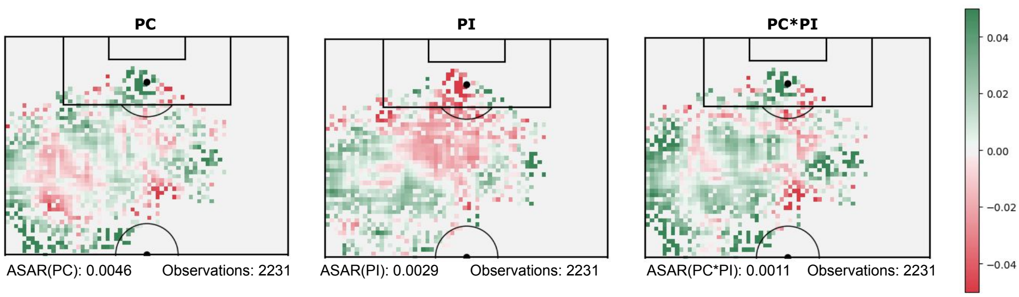


Figure 7.9: Performance of Muamer Tancović.

7.4.2 Alexander Kacaniklic

Alexander Kacaniklic is also an AM who played a total of 25 matches and scored a total of 10 goals during the 2019 season. 1753 observations are included to assess his off-ball performance. His performance is illustrated in Figure 7.10. Similar to his teammate, Tanković, he was able to achieve well over expected on all metrics, however, showing more variability over location specific performance. As can be interpreted from his PI

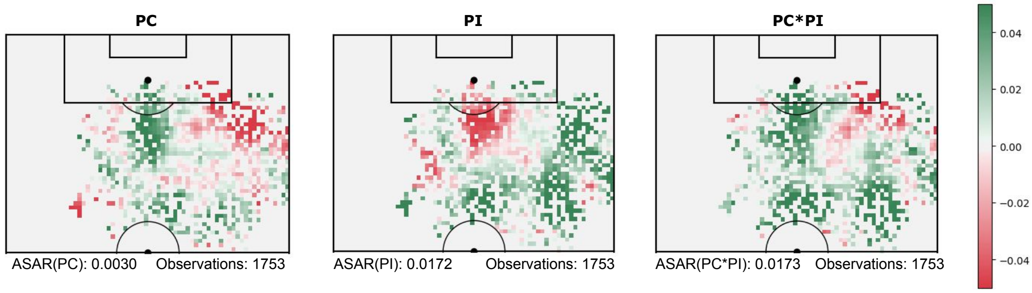


Figure 7.10: Performance of Alexander Kacaniklic.

performance, Kacaniklic shows resistance to move into the box when close to the goal. However, he is well above expected when it comes to capturing control in the same are. This indicates that he is good at finding space close to the goal. Further, he shows signs of being good at putting pressure in the early stages of attacks, being well over expected on PI further away from the goal.

Chapter 8

Predicting Player Movement

In this chapter, the focus is shifted towards predicting movement of players one second into the future after a situation of interest has occurred. The selected situations are the passes in the opponents' half, as described in Section 4.2.1. Similar to Chapter 6, two different classes of models are used, generalised additive model and feed forward neural network. The reasoning for choosing these specific types of models follows the discussions in Sections 6.3.1 and 6.3.2. The purpose of creating these models is to investigate if player movement can be accurately predicted over a short time interval and if there are differences in the predictability of individual players and roles. Another application is to investigate if the optimal points generated for PC, PI, and PC*PI can work as a good forecast for player movement, a hypothesis previously investigated by Peralta Alguacil et al. (2020). This chapter seeks to test this hypothesis on a larger scale, by using more data and comparing them to alternative modeling techniques and benchmarks.

8.1 Experimental Setup

The experimental setup in this section is similar to the one detailed in Chapter 6, using the same type of situations and filtering process as presented in Section 4.2. A brief summary; player movements are predicted one second into the future after a pass is made in the opponent's half, with only players on the attacking team included in the prediction. The models are in this chapter trained on 70% of the data and validated on the remaining 30%.

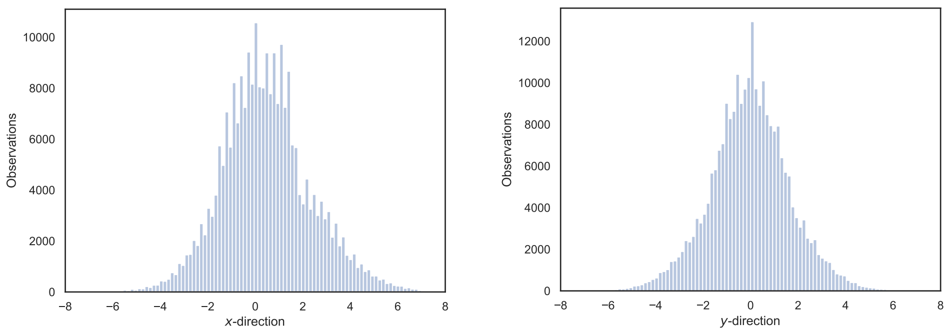
8.1.1 Dependent variables

The dependent variables are set to the x- and y-components of the vector between the player's initial position and their end position one second later. The range is set from -7.8 to 7.8 meters, as this is the maximum distance a player can move during the course of a second, following from Section 5.3.2.

The distribution of the dependent variables are illustrated in Figure 8.1.

Dependent variable	Range (meters)
Movement in x-direction	$[-7.8, 7.8]$
Movement in y-direction	$[-7.8, 7.8]$

Table 8.1: Dependent variables for movement prediction models.



(a) x-direction **(b)** y-direction

Figure 8.1: Distributions of the two dependent variables in Table 8.1.

8.1.2 Explanatory variables

Some changes are made to the explanatory variables compared to the problem in Chapter 6, as the models for this problem are created without any information about the future game state. As variables regarding PC, PI, and PC*PI are computed from a future game state, they are considered to contain information about the future. Therefore, the related variables are all dropped, namely the location of optimal points, and their standard deviation. In this section, the role of the player is added as the categorical variable $X_{33,r}$ with $r \in \{FW, AM, MF, CB, FB\}$. The full list of explanatory variables can be seen in Table 8.2.

8.1.3 GAM model set up

Similarly to Section 6.3.1, the GAM built in this section was constructed using the $mgcv$ library in R. The model is constructed as a multivariate GAM with the two dependent variables from Section 8.1.1. When modelling a multivariate GAM in $mgcv$, it is only possible to set the distribution of the conditional mean as a normal distribution. However, when observing the distribution of the dependent variables in Figure 8.1, using a normal distribution does not appear to be unreasonable. The explanatory variables used to predict player movement with the GAM are presented in Table 8.2. Unlike when modelling the features in Chapter 6, the features in this chapter are not modelled as tensor products, as the results were better when modelling an individual smooth term for each feature and dependent variable.

Variable	Description	Type
X_1	Initial position of player x-coordinate	Continuous
X_2	Initial position of player y-coordinate	Continuous
X_3	Initial velocity of player x-coordinate	Continuous
X_4	Initial velocity of player y-coordinate	Continuous
X_8	Initial position of ball x-coordinate	Continuous vector component
X_9	Initial position of ball y-coordinate	Continuous vector component
X_{10}	Team centroid x-coordinate	Continuous vector component
X_{11}	Team centroid y-coordinate	Continuous vector component
X_{12}	Opponent centroid x-coordinate	Continuous vector component
X_{13}	Opponent centroid y-coordinate	Continuous vector component
X_{14}	Closest opponent x-coordinate	Continuous vector component
X_{15}	Closest opponent y-coordinate	Continuous vector component
X_{22}	Team centroid speed x-direction	Continuous
X_{23}	Team centroid speed y-direction	Continuous
X_{24}	Opponent centroid speed x-direction	Continuous
X_{25}	Opponents centroid speed y-direction	Continuous
X_{26}	Team compactness x-direction	Continuous
X_{27}	Team compactness y-direction	Continuous
X_{28}	Opponent compactness x-direction	Continuous
X_{29}	Opponent compactness y-direction	Continuous
X_{30}	Game time	Continuous
X_{31}	Passing angle	Continuous
X_{32}	Direction of pass, left or right	Binary
$X_{33,r}$	Player Role	Categorical

Table 8.2: Explanatory variables for predicting player movement.

Hyperparameter	Value
Number of epochs	1200
Learning rate	0.00001
Learning rate decay	0.00001/1200
Batch size	64
Loss function	MSE
Hidden layers structure	256;256;256;256
Hidden layers activation functions	Linear(1) and Relu(2-4)
Output layer activation function	Relu
L1 Regulariser Penalty	0.0001
Additional hyperparameters	Keras default values

Table 8.3: Hyperparameters for final FFNN used to predict player movement.

Benchmark	
Initial position	IP
Initial direction	ID
Optimal control	OC
Optimal Impact	OI
Optimal combined control and impact	OCI

Table 8.4: Benchmarks for comparison with created models.

8.1.4 FFNN model set up

The FFNN developed to predict player movement is similar to the network developed in Chapter 6. However, some changes are made to better fit this problem, with testing forming the basis for these changes. An L1 regulariser is added, with the penalty set to 0.0001. This increases the model's ability to generalise, but also slows down the training process, making it necessary to raise the number of epochs to obtain the lower loss limit. The learning rate and its decay are also lowered, as for this problem the accuracy of the model is more sensitive to changes in the weights. The final model's hyperparameters are presented in Table 8.3. For this section, the technique of training ten separate models as presented in Section 6.3.2 is dropped, as tests only showed marginal improvements and the training time is ten times higher using this approach.

8.2 Benchmarks for Predictions

Player movement is a situation-dependent problem, where the situation and dynamics of play can have a large impact on the movement of players. Some situations will be easier to predict than others, i.e. if the player is standing still at the beginning of the situation. It is, therefore, necessary to assess the models by comparing them to some predetermined heuristic techniques, functioning as benchmarks. The first benchmark, IP, is simply using the player's initial position of the situation, a benchmark also used in Peralta Alguacil et al. (2020). This benchmark is included to see if the models are able to add prediction power above just the initial position. The second benchmark, ID, uses the initial position and predicts the movement by assuming the player follows initial direction and velocity over the next time period. Comparing the models to this benchmark is a way to investigate if the models are able to predict the changes in direction and velocity the players undertake during the time interval. The points of optimal PC, PI, and PC*PI are also included as benchmarks, following the concepts detailed in Section 4.1. These are included to assess if players make movement decisions on optimal position, and see how these models compare to other alternatives. The five benchmarks used are listed in Table 8.4. The accuracy of the models and benchmarks are measured in x- and y-direction, as well as the Euclidean distance between the true movement, and the movement predicted by the different models and benchmarks. These benchmark and models will be evaluated as a total average on all observations, an average for the individual players with their assigned role, and an assessment of how often the individual models perform best compared to the others.

8.3 Results

In this section, the results from the GAM and FFNN models created to predict movement are presented and compared with the benchmarks presented in Section 8.2. The models are validated using the same validation metrics as presented in Section 6.4, and will also be compared on how often they give the most accurate prediction.

8.3.1 Validation results

The results of the validation of x- and y-direction are presented in Tables 8.5 and 8.6, and the Euclidean distance presented in Table 8.7. All measures are given in meters. Both the GAM and FFNN outperform the alternative benchmarks on all validation metrics. This is expected when compared to IP and ID, as both the models are fitted on information present in IP and ID, and should be able to find a better fit with additional information as well as the ability to scale the contributions. None of the optimal point benchmarks show signs of being a good predictor of player movement, with all of them missing by an average of over 3 meters.

Again, the FFNN outperforms the GAM on all validation metrics, showing a better ability to predict player movement. As with the problem in Chapter 6, this is believed to come from the model's ability to interpret non-linear relationships between all variables, and not only from non-linear contributions of selected variable pairs, which is the case for the GAM.

An interesting observation is the differences in predictability in x- and y-direction. From the IP benchmark, the MAE is lower in y- than x-direction, meaning that players move a shorter distance on average in y-direction than x during the one second. However, both ID, GAM, and FFNN shows better ability to predict the movement in x-direction. A possible explanation is that players often tend to follow the course of the attack, moving up and down the pitch (x-direction) collectively depending on the momentum of the attack, while movement in y-direction is more stochastic.

Model	MAE	RMSE	MBE	R^2
IP	1.474	1.920	-0.490	0
ID	0.627	0.883	0.074	0.774
OC	2.304	2.607	0.450	-0.864
OI	2.398	2.652	2.344	0.213
OCI	2.190	2.501	1.210	-0.274
GAM	0.542	0.737	-0.001	0.842
$FFNN$	0.504	0.692	-0.014	0.861

Table 8.5: Validation x-direction.

Model	MAE	RMSE	MBE	R^2
IP	1.300	1.680	0.032	0
ID	0.789	1.128	0.001	0.549
OC	2.023	2.405	0.024	-1.049
OI	1.523	1.860	0.018	-0.225
OCI	1.949	2.329	0.033	-0.921
GAM	0.673	0.916	0.000	0.702
$FFNN$	0.627	0.860	-0.002	0.738

Table 8.6: Validation y-direction.

Model	MAE	RMSE	MBE[a]	R^2
IP	2.190	2.551	-2.190	0
ID	1.115	1.433	0.212	0.485
OC	3.396	3.547	1.090	-0.534
OI	3.078	3.239	1.369	0.012
OCI	3.272	3.417	1.277	-0.328
GAM	0.956	1.176	-0.242	0.558
$FFNN$	0.891	1.103	-0.225	0.588

Table 8.7: Validation Euclidean distance.

[a] Calculated as the bias in prediction of Euclidean distance traveled from the initial position.

8.3.2 Evaluation and discussion

As decision making in football is very situation-dependent, the models have to be evaluated beyond just their average validation score to assess their true performance. One model may perform very well in certain situations, but worse in others, therefore impacting the validation metrics negatively. By assessing how often a model generates the most accurate prediction, the model's ability to best describe movement in certain situations can be evaluated.

Table 8.8 shows the percentage of the situations where the model gave the most accurate prediction of player movement in x- and y-direction, as well as Euclidean distance. As movement in football is conducted by individual players, with different physical attributes, mindsets, and roles, it is reasonable to assume that player movement is stochastic. With enough data, it is therefore expected that all models will on some occasions deliver the most accurate prediction, a consequence of randomness in a stochastic process. This is observed for the prediction accuracy separated in x- and y-direction. The distribution of the most accurate model is, therefore, more even in x- and y-direction compared to the Euclidean distance which is a two-dimensional metric. For the accuracy on Euclidean distance, the results show similarities to the results from the validation in Section 8.3.1. The FFNN comes out on top, followed by the GAM, ID, and IP. The optimal points very rarely outperform the other models and heuristics, and in the few observations where they do, it may be the result of randomness or some very specific situations where they predict well.

	IP	**ID**	**OC**	**OI**	**OCI**	**GAM**	**FFNN**
x-direction	11.82%	24.99%	4.49%	4.27%	4.86%	22.66%	26.96%
y-direction	14.55%	22.32%	5.72%	10.17%	5.47%	19.11 %	23.11%
Euclidean	9.41%	26.41%	1.17%	2.26%	1.30%	24.37 %	35.13 %

Table 8.8: Percentage of test set where the model made the most accurate prediction for x-direction, y-direction and the Euclidean distance.

8.4 Individual Predictability and Role Differences

With FFNN and ID being the best model and heuristic technique respectively to predict player movement, these approaches will be used in this section to assess the differences in the predictability between individual players and roles. In Figure 8.2, the MAE's of individual players from the FFNN are plotted against the ID benchmark. From the figure, the differences in the predictability of the individual players can be spotted, with the MAE ranging from 0.568 meters up to 1.236 meters. The predictive power of the model is correlated with the ID benchmark, meaning that players that often change their movement from initial direction and velocity are harder to predict. As can be seen from the dotted identity line, as all players are located below this line, the FFNN outperforms the ID on all players on average. To assess the differences in predictability for different player roles, the MAE for each role is calculated, shown in Table 8.9. For the FFNN, CBs stands out as easier to predict than the rest, with an MAE of 0.826 meters. The rest of the roles shows more similar results, with AMs having the highest MAE of 0.925. These results are correlated with the results from the ID, having a correlation coefficient of 0.894. CBs tend to deviate less from their initial velocity and direction, making them easier to predict.

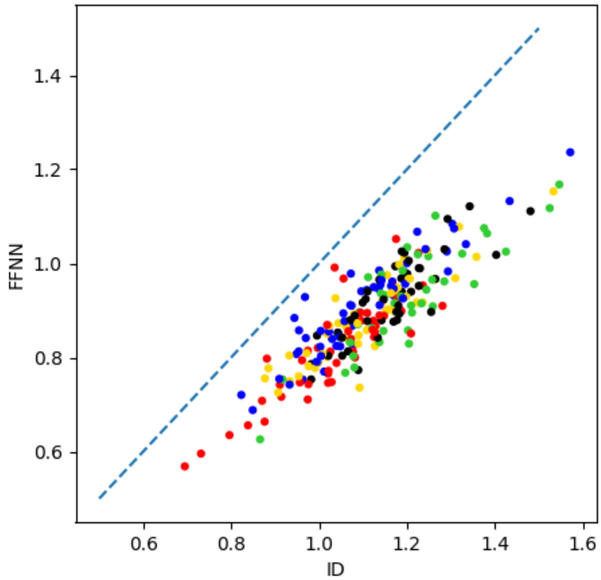


Figure 8.2: MAE of ID and FFNN for individual players. Colours represent the role of the players: ● - CB, ● - MF, ● - FB, ● - FW, ● - AM. The dotted line is an identity line.

	CB	FB	MF	AM	FW
ID	1.040	1.098	1.094	1.166	1.195
FFNN	0.826	0.881	0.902	0.925	0.920

Table 8.9: Average MAE for different roles.

Chapter 9

Answers to Research Questions

This chapter presents answers to the research questions posed in Section 1.2. The answers are based on the results and discussions provided in Chapters 6, 7 and 8.

RQ1: Can available state of the art off-ball metrics be used to evaluate off-ball movement and decision making in elite-level football?

A central question regarding the validity of the evaluation methods and applications presented in this thesis is whether the off-ball metrics used are adequate to describe the value of different off-ball movements. In football, there is only one commodity of value, goals, and actions only have value if they increase your team's probability of scoring a goal or decrease your opponent's probability. As the pitch control metric is developed with input and support from experts working in professional football clubs and the pitch impact metric specifically uses goals to infer value, both these metrics have some support for being used in this context. The more pressing question may ,therefore, be if there are other elements of off-ball positional value not captured by the metrics used in this thesis. As previously discussed, the metric calculating pass probabilities were too computationally expensive to use on a large scale for this thesis. It is, therefore, reasonable to suspect that including this metric could lead to a more comprehensive evaluation. There may also be other metrics describing other important parts of off-ball movement yet to be developed, and including these when evaluating players and teams could improve understanding of individual and role-specific movement behaviours.

Another characteristic of the framework presented in this thesis is that each evaluation only considers performance on one metric or an equal weighting of two. Assuming that achieving both control and impact are valuable; are they equally valuable, or is the ratio of value dependent on roles, position or game situation? Answering these questions would represent another step towards a comprehensive evaluation of off-ball movement. The results of ranking the players presented in Section 7.2, seem to show some support for the validity of the framework. The presence of several players generally considered to be top players in Allsvenskan in the top 10, and some correlation with subjective assessments made by the professional scouting network coming from FM20. The careful conclusion to this research question, therefore, becomes that while some questions and challenges still remain, the available metrics used in this thesis can be used to evaluate aspects of off-ball movement. This functions as an addition to existing player evaluation metrics, with new aspects of the game possible to be evaluated.

RQ2: Are there individual differences in off-ball decision making, and do decision making vary for different player roles?

Evaluation of individual players show that there are differences in how the individual players perform regarding the three off-ball metrics. Further, the results presented in Sections 7.1 and 7.2 suggests that players make different positional decisions depending on their role. Attacking players, especially forwards, seem to focus more on pitch impact than pitch control, meaning that they prefer to seek out positions closer to the goal where the impact of receiving the ball is higher. Another role which seems to differ in positional strategies is central defenders who seem to a lesser extent to make decisions based on impact and control than other players. Whether these differences are the result of differences in players instincts leading them to make different decisions and therefore makes them suited to different roles, role-specific instructions by coaches or a combination of these two factors is not yet known. Comparing the performance of individual players with the characteristic performance of different roles could be a useful tool for evaluating what roles fits the player. Further, players could be compared to asses on similarities and distinct performance, with regards to scouting and player acquisition.

RQ3: Are players on the best attacking teams in Allsvenskan making different off-ball positional decisions than other teams?

Section 7.3 seeks to answer this question by dividing the teams in Allsvenskan by the number of goals scored during the 2019 season into four groups and comparing their off-ball performance. The results did not prove that better performance on the three metrics, in general, was an indicator of attacking success, as clear trends between the average performance and success could not be identified. No single group appeared to consistently outperform the other groups, as three different groups found themselves in the top spot in one or more of the analyses. Three different groups also inhabited the bottom spot on one or more of the analyses. However, the most interesting finding from this section seems to be that the teams with the most goals scored divide the positional responsibility more than less successful teams. The attackers on the best teams performed the best on pitch impact while their midfielders and defenders performed the best on pitch control.

RQ4: How well can player movement be predicted over a short time interval, and what types of models are best suited to model this movement?

The results presented in Chapter 8 shows that feed-forward neural networks are better able to capture the dynamics of player movement than generalised additive models and the other benchmarks included in this thesis, during attacking passing situations. With players moving on average just over 2 meters from their initial positions, predicting the positions with an average error of approximately 0.9 meters using the feed-forward neural network means that more than half of the movement players undertake is predictable using the selected features. The ID benchmark, determining future positions by assuming the player continues with the initial velocity and direction, has an error of approximately 1.1 meters. This means that most of the information used in the FFNN to predict future movement is obtained from the player's initial movement. Assuming that player movement contains stochastic elements, errorless predictions are not possible. How much of the remaining error being a consequence of lacking information in the features used, weakness in the modelling approach, individual differences, or randomness, is not yet known.

Chapter 10

Conclusion and Recommendations for Further Research

This chapter concludes this master thesis with some closing remarks of the results from this thesis. The conclusion is followed by some recommendations for further research on off-ball evaluation.

10.1 Conclusion

Using tracking data, a framework for analysing off-ball performance has been presented. Metrics for defining off-ball success was introduced, with models created with the intention to predict and analyse the contribution of different features. Players were then evaluated on their obtained and situational adjusted score for off-ball performance. The results show a moderate positive correlation to existing ratings from experts, indicating that the ratings created capture some of the same characteristics as present off-ball ratings. The evaluation methods presented in this thesis are therefore recommended to be used not as a replacement, but as an addition to existing methods. As the framework only considers off-ball involvements, it needs to be combined with other metrics describing a players on-ball performance for a comprehensive evaluation of a player. The current framework also only considers attacking passing situations, and it is, therefore, better suited to evaluate the off-ball behaviour of more attacking-minded players than defenders. Evaluating players intended to have a defensive role may lead to the player being undervalued compared to players in other roles.

Throughout this thesis, the FFNN shows better abilities than the GAM for modelling off-ball decision making as it delivers more accurate predictions for both player performance and movement. Many features are present to model these aspects, and the FFNN shows signs of being better suited for understanding non-linear relationships between the features. However, the FFNN is less interpretable than the GAM, with no easy way of investigating contributions from individual features or feature pair. Further, results show that player movement is harder to predict for players deviating a lot from initial direction and velocity, with remaining effects yet to be discovered to better model player movement.

10.2 Recommendations for Further Research

In this section, some ideas and suggestion for further research on off-ball movement are presented. Some suggestions are intended as possible extensions to the work done in this thesis, while others can be considered as alternative approaches.

10.2.1 Expanding the framework

One possible approach to future research on off-ball movement is to expand on the analysis from this thesis with improved or different data and methods. The data used in this thesis, detailed in Chapter 5, lacks some details on match events and contains tracking errors. Improving the precision and adding more dimensions to the data used in the models may improve the framework through fewer errors and the addition of new features.

As previously mentioned, another possible improvement is the inclusion of pass probability as a metric inferring positional value. The code used to compute the PP model is currently not parallelised, making it computationally expensive to run it on large amounts of data using relatively modest systems. Other off-ball metrics could also be developed to describe other aspects of off-ball success. This would add further understanding towards off-ball movement, and could be an addition to the existing framework.

Another possible expansion is to consider different types of situations than what is considered in this thesis. Instead of analysing attacking situations, defensive situations could be used to evaluate defensive movement. Gaining pitch control means limiting your opponent's control, so this metric could be used in the same way as in this thesis. Impact is a different type of metric and an analysis of defensive performance could be to assess how defenders limit the impact their opponents achieve. If pass probability could be included, this metric could then be used to analyse how the defending team limits the total pass probability surface available to the attacking team.

Expanding to consider not just passing events, but other events such as tackles and dribbles is also a possibility, with the ultimate extension being a continuous evaluation of movement throughout entire matches. Evaluating movement on a broader range of situations could give a more complete picture of player movement, but increasing the number of situations comes with a computational cost.

Much of the work done in this thesis is based on the assumption that players are to some extent able to predict the movement of teammates and opponent over short time intervals. As measures like pitch control depend on other players' positions, as well as the player's own position, finding optimal positions requires some insight into the future movement of other players. Experimenting with the time interval of movement is, therefore, another possible extension. Increasing the time interval would expand the player's reachable area and therefore their available options, but requires the assumption of players predicting other players' positions to be extended further into the future. As an increased time interval means a larger reachable area, more than 200 points inside the reachable may need to be sampled to represent the player's opportunities adequately.

10.2.2 Alternative approach to off-ball metric evaluation

An alternative approach to evaluating off-ball movement is to directly link movement to the ultimate objective of football, scoring goals. This can be achieved with the use of a classification model, linking actions to goals in the near future. This has previously been done with on-ball events (Mackay, 2017; Decroos et al., 2019) and an extension to off-ball movement could be possible. The movement would, therefore, be evaluated on its contribution towards the probability of scoring or not conceding a goal in a given time frame. Metrics such as impact and control, could be used as features in the model, along with many of the same features used in this thesis. This approach requires a large amount of data containing information on player and ball movement, along with detailed information on on-ball events such as passes, shots and goals. The data used in this thesis does not contain such event information for this approach to be possible, but the right data set could be constructed by combining tracking and event data from different sources.

Bibliography

Barros, R.M.L., Misuta, M.S., Menezes, R.P., Figueroa, P.J., Moura, F.A., Cunha, S.A., Anido, R., Leite, N.J., 2007. Analysis of the distances covered by first division Brazilian soccer players obtained with an automatic tracking method. Journal of sports science medicine 6, 233–242.

Baseball-Reference, 2020. Baseball-reference.com WAR explained. URL: `https://www.baseball-reference.com/about/war_explained.shtml`. online; accessed 14-March-2020.

Basketball-Reference, 2020. NBA Win Shares. URL: `https://www.basketball-reference.com/about/ws.html`. online; accessed 14-March-2020.

Bransen, L., Robberechts, P., Van Haaren, J., Davis, J., 2019. Choke or shine? quantifying soccer players' abilities to perform under mental pressure, in: 13th MIT Sloan Sports Analytics Conference.

Burke, B., 2019. DeepQB: deep learning with player tracking to quantify quarterback decision-making & performance, in: Proceedings of the 2019 MIT Sloan Sports Analytics Conference.

Chawla, S., Estephan, J., Gudmundsson, J., Horton, M., 2017. Classification of passes in football matches using spatiotemporal data. ACM Transactions on Spatial Algorithms and Systems 3, 1–30.

Cook, T.K., Keil, E., Sætherø, I.D., 2019. Analysing Crosses from Open Play at Rosenborg Ballklubb. Project report in TIØ4550. Department of Industrial Economics and Technology Management, NTNU.

Decroos, T., Bransen, L., Van Haaren, J., Davis, J., 2019. Actions speak louder than goals: Valuing player actions in soccer, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1851–1861.

Di Salvo, V., Baron, R., Tschan, H., Calderon Montero, F.J., Bachl, N., Pigozzi, F., 2007. Performance characteristics according to playing position in elite soccer. Int J Sports Med 28, 222–7.

Di Salvo, V., Gregson, W., Atkinson, G., Tordoff, P., Drust, B., 2009. Analysis of high intensity activity in Premier League soccer. Int J Sports Med 30, 205–12.

Eager, E., Chahrouri, G., 2020. PFF WAR: Modeling player value in american football, in: Proceedings of the 2020 MIT Sloan Sports Analytics Conference.

Eilers, P.H., Marx, B.D., 1996. Flexible smoothing with B-splines and penalties. Statistical science , 89–102.

Fernandez, J., Bornn, L., 2018. Wide open spaces: A statistical technique for measuring space creation in professional soccer, in: 12th MIT Sloan Sports Analytics Conference.

Fernández, J., Bornn, L., Cervone, D., 2019. Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer, in: 13th MIT Sloan Sports Analytics Conference.

FIFA, 2007. 265 million playing football. URL: https://www.fifa.com/mm/document/fifafacts/bcoffsurv/emaga_9384_10704.pdf. online; accessed 7-December-2019.

FIFA, 2018. 2018 FIFA World Cup Russia, global broadcast and audience summary. URL: https://resources.fifa.com/image/upload/2018-fifa-world-cup-russia-global-broadcast-and-audience-executive-summary.pdf?cloudid=njqsntrvdvqv8ho1dag5. online; accessed 7-December-2019.

Floater, M.S., 2015. Bézier curves and surfaces .

FmDataba, . Player attributes in FM20. URL: https://fmdataba.com/. online; accessed 10-June-2020.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press. Adaptive computation and machine learning series, pp. 227–231.

Gregory, S., 2019. Ready player run: Off-ball run identification and classification. Barça Sports Analytics Summit .

Gudmundsson, J., Wolle, T., 2014. Football analysis using spatio-temporal tools. Computers, Environment and Urban Systems 47, 16–27.

guidetofm.com, . Player attributes. URL: https://www.guidetofm.com/players/attributes/3/. online; accessed 17-June-2020.

Haave, H.S., Høiland, H., 2017. Evaluating association football player performances using markov models. Master's thesis. Department of Industrial Economics and Technology Management, NTNU.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference and prediction. 2 ed., Springer.

Hastie, T.J., 2017. Generalized additive models, in: Statistical models in S. Routledge, pp. 249–307.

Hopfield, J.J., 1988. Artificial neural networks. IEEE Circuits and Devices Magazine 4, 3–10.

Håland, E.M., Wiig, A.S., 2018. Evaluating Passing Behaviour in Norwegian Association Football. Master's thesis. Department of Industrial Economics and Technology Management, NTNU.

Kharrat, T., McHale, I.G., Peña, J.L., 2020. Plus–minus player ratings for soccer. European Journal of Operational Research 283, 726 – 736.

Kim, S., 2004. Voronoi analysis of a soccer game. Nonlinear Analysis: Modelling and Control 9, 233–240.

Kingma, D.P., Ba, J.L., 2014. Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations .

Krishna, 2019. Rankings europe's top five leagues by goals per game ratio in 2018/19. URL: `https://www.90min.com/in/posts/6380147-rankings-europe-s-top-five-leagues-by-goals-per-game-ratio-in-2018-19?`. online; accessed 5-June-2020.

Lawrence, T., 2018. Introducing xGChain and xGBuildup. StatsBomb IQ Services. URL: `https://statsbomb.com/2018/08/introducing-xgchain-and-xgbuildup/`. online; accessed 11-March-2020.

Mackay, N., 2017. Predicting goal probabilities for possessions in football. Master of Science. Vrije Universiteit Amsterdam .

McHale, I.G., Relton, S.D., 2018. Identifying key players in soccer teams using network analysis and pass difficulty. European Journal of Operational Research 268, 339–347.

Memmert, D., Raabe, D., 2018. Data Analytics in Football: Positional Data Collection, Modelling and Analysis. Routledge.

Memmert, D., Raabe, D., Schwab, S., Rein, R., 2019. A tactical comparison of the 4-2-3-1 and 3-5-2 formation in soccer: A theory-oriented, experimental approach based on positional data in an 11 vs. 11 game set-up. PloS one 14.

Memmert, D., Rein, R., 2018. Match analysis, big data and tactics: current trends in elite soccer. German Journal of Sports Medicine/Deutsche Zeitschrift fur Sportmedizin 69.

Miller, A.C., Bornn, L., 2017. Possession sketches: Mapping NBA strategies, in: 11th MIT Sloan Sports Analytics Conference.

Mottley, C.M., 1954. The application of operations-research methods to athletic games. Journal of the Operations Research Society of America 2, 335–338.

Moura, F., Martins, L.B., Anido, R., Ruffino, P.C., Barros, R.L., Cunha, S., 2013. A spectral analysis of team dynamics and tactics in Brazilian football. Journal of Sports Sciences 31, 1568–1577.

Moura, F., Martins, L.B., Anido, R.O., De Barros, R.L., Cunha, S., 2012. Quantitative analysis of Brazilian football players' organisation on the pitch. Sports Biomechanics 11, 85–96.

Nørstebø, O., Bjertnes, V.R., Vabo, E., 2016. Valuing individual player involvements in Norwegian Association Football. Master's thesis. Department of Industrial Economics and Technology Management, NTNU.

Pappalardo, L., Cintia, P., Pedreschi, D., Giannotti, F., Barabasi, A.L., 2017. Human perception of performance. arXiv preprint arXiv:1712.02224 .

Patterson, H.D., Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58, 545–554.

Peralta Alguacil, F.J., 2019. Modelling the Collective Movement of Football Players. Master's thesis. Department of Information Technology, Uppsala University.

Peralta Alguacil, F.J., Fernandez, J., Piñones Arce, P., Sumpter, D., 2020. Seeing in to the future: using self-propelled particle models to aid player decision-making in soccer, in: 14th MIT Sloan Sports Analytics Conference.

Perl, J., Memmert, D., 2017. A pilot study on offensive success in soccer based on space and ball control–key performance indicators and key to understand game dynamics. International Journal of Computer Science in Sport 16, 65–75.

Power, P., Ruiz, H., Wei, X., Lucey, P., 2017. Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data, in: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1605–1613.

Rathke, A., 2017. An examination of expected goals and shot efficiency in soccer. Journal of Human Sport Exercise 12, S514 – S529.

Rein, R., Raabe, D., Memmert, D., 2017. "which pass is better?" novel approaches to assess passing effectiveness in elite soccer. Human movement science 55, 172–181.

Rein, R., Raabe, D., Perl, J., Memmert, D., 2016. Evaluation of changes in space control due to passing behavior in elite soccer using Voronoi-cells. Springer International Publishing. Advances in Intelligent Systems and Computing, pp. 179–183.

Rudd, S., 2011. A framework for tactical analysis and individual offensive production assessment in soccer using markov chains, in: New England Symposium on Statistics in Sports. http://nessis. org/nessis11/rudd. pdf.

Sæbø, O.D., Hvattum, L.M., 2019. Modelling the financial contribution of soccer players to their clubs. Journal of Sports Analytics 5, 23–34.

Schulze, E., Mendes, B., Maurício, N., Furtado, B., Cesário, N., Carriço, S., Meyer, T., 2018. Effects of positional variables on shooting outcome in elite football. Science and Medicine in Football 2, 93–100.

Senthilnathan, S., 2019. Usefulness of correlation analysis. SSRN Electronic Journal .

Sill, J., 2010. Improved NBA adjusted+/-using regularization and out-of-sample testing, in: Proceedings of the 2010 MIT Sloan Sports Analytics Conference.

Singh, K., 2019. Introducing expected threat (xT). URL: https://karun.in/blog/expected-threat.html. online; accessed 11-March-2020.

Smith, K., 2015. Why clubs are using football manager as a real-life scouting tool. URL: https://www.theguardian.com/technology/2014/aug/12/why-clubs-football-manager-scouting-tool. online; accessed 17-June-2019.

Spearman, W., Basye, A., Dick, G., Hotovy, R., Pop, P., 2017. Physics-based modeling of pass probabilities in soccer, in: Proceeding of the 11th MIT Sloan Sports Analytics Conference.

Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems, 27:3104–3112 .

Szczepański, Ł., McHale, I., 2016. Beyond completion rate: evaluating the passing ability of footballers. Journal of the Royal Statistical Society: Series A (Statistics in Society) 179, 513–533.

Taki, T., ichi Hasegawa, J., Fukumura, T., 1996. Development of motion analysis system for quantitative evaluation of teamwork in soccer games. Proceedings of 3rd IEEE International Conference on Image Processing 3, 815–818 vol.3.

Tovar, J., Clavijo, A., Cárdenas, J., 2017. A strategy to predict association football players' passing skills .

Twelve, . Twelve.football/analytics. URL: https://twelve.football/analytics. online; accessed 23-May-2020.

Ueda, F., Masaaki, H., Hiroyuki, H., 2014. The causal relationship between dominant region and offense-defense performance–focusing on the time of ball acquisition. Football Science 11, 1–17.

UEFA, 2018. UEFA stadium infrastructure regulations, edition 2018. URL: https://www.uefa.com/MultimediaFiles/Download/uefaorg/Stadium&Security/01/48/48/85/1484885_DOWNLOAD.pdf. online; accessed 12-June-2020.

Wood, S., 2006. Generalized Additive Models: An Introduction with R. 1 ed., Chapman and Hall/CRC.

Wood, S.N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 73, 3–36.

Worville, T., 2017. Expected assists in context. URL: `https://www.optasportspro.com/news-analysis/blog-expected-assists-in-context/`. online; accessed 14-March-2020.

Yam, D., 2019. A data driven goalkeeper evaluation framework, in: 13th MIT Sloan Sports Analytics Conference.

Appendix A

Abbreviations

Team	Abbreviation
AFC Eskilstuna	AFC
AIK	AIK
BK Häcken	BKH
Djurgården	DIF
Falkenbergs FF	FFF
GIF Sundsvall	GIF
Hammarby IF	HAM
Helsingborgs IF	HIF
IF Elfsborg	IFE
IFK Göteborg	IFK
IFK Norrköping	IFN
IK Sirius	IKS
Kalmar FF	KFF
Malmö FF	MFF
Örebro SK	ÖSK
Östersunds FK	ÖFK

Table A.1: Team names and abbreviations of teams in the Allsvenskan 2019 season.

Appendix B

Regression results for GAMs in Chapter 6

B.1 Significance of terms

Smooth terms	
Variable	Sign
$f_1(X_1, X_2)$	$***$
$f_2(X_3, X_4)$	$***$
$f_3(X_5)$	$***$
$f_4(X_6)$	$***$
$f_5(X_7)$	$***$
$f_6(X_8, X_9)$	$***$
$f_7(X_{10}, X_{11})$	$***$
$f_8(X_{12}, X_{13})$	$***$
$f_9(X_{14}, X_{15})$	$***$
$f_{10}(X_{16}, X_{17})$	$***$
$f_{11}(X_{18}, X_{19})$	$***$
$f_{12}(X_{20}, X_{21})$	$***$
$f_{13}(X_{22}, X_{23})$	$***$
$f_{14}(X_{24}, X_{25})$	$***$
$f_{15}(X_{26}, X_{27})$	$***$
$f_{16}(X_{28}, X_{29})$	$***$
$f_{17}(X_{33})$	$***$
$f_{18}(X_{34})$	$***$

Fixed effects	
Variable	Coefficient
X_{35}	-0.002 (0.004)
Intercept	0.040***(0.003)

Table B.1: Regression results from the GAM_{PC} model. ' ' $p < 1$; '***' $p < 0.001$.

Smooth terms		Fixed effects	
Variable	Sign	Variable	Coefficient
$f_1(X_1, X_2)$	***	X_{35}	-0.010**(0.004)
$f_2(X_3, X_4)$	***	Intercept	0.042**(0.003)
$f_3(X_5)$	***		
$f_4(X_6)$	***		
$f_5(X_7)$	***		
$f_6(X_8, X_9)$	***		
$f_7(X_{10}, X_{11})$	***		
$f_8(X_{12}, X_{13})$	***		
$f_9(X_{14}, X_{15})$	***		
$f_{10}(X_{16}, X_{17})$	***		
$f_{11}(X_{18}, X_{19})$	***		
$f_{12}(X_{20}, X_{21})$	***		
$f_{13}(X_{22}, X_{23})$	***		
$f_{14}(X_{24}, X_{25})$	***		
$f_{15}(X_{26}, X_{27})$	***		
$f_{16}(X_{28}, X_{29})$	***		
$f_{17}(X_{33})$	***		
$f_{18}(X_{34})$	***		

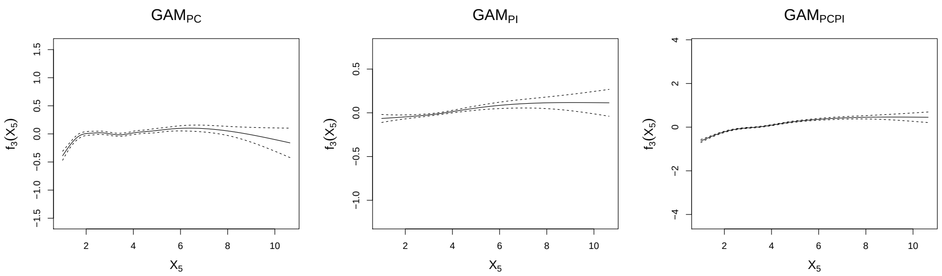
Table B.2: Regression results from the GAM_{PI} model. '**' $p < 0.01$; '***' $p < 0.001$.

Smooth terms	
Variable	Sign
$f_1(X_1, X_2)$	***
$f_2(X_3, X_4)$	***
$f_3(X_5)$	***
$f_4(X_6)$	***
$f_5(X_7)$	***
$f_6(X_8, X_9)$	***
$f_7(X_{10}, X_{11})$	***
$f_8(X_{12}, X_{13})$	***
$f_9(X_{14}, X_{15})$	***
$f_{10}(X_{16}, X_{17})$	***
$f_{11}(X_{18}, X_{19})$	***
$f_{12}(X_{20}, X_{21})$	***
$f_{13}(X_{22}, X_{23})$	***
$f_{14}(X_{24}, X_{25})$	***
$f_{15}(X_{26}, X_{27})$	***
$f_{16}(X_{28}, X_{29})$	***
$f_{17}(X_{33})$	***
$f_{18}(X_{34})$	***

Fixed effects	
Variable	Coefficient
X_{35}	0.011**(0.004)
Intercept	-0.063***(0.003)

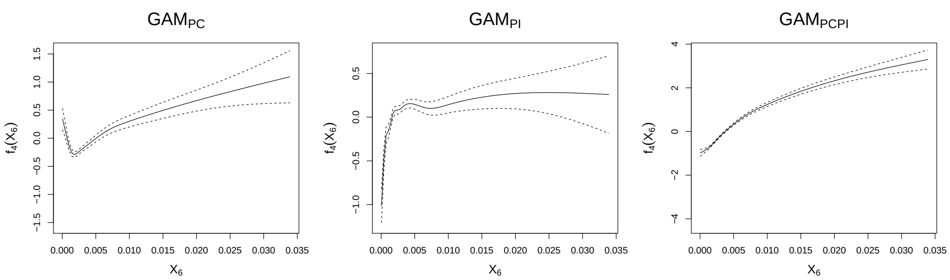
Table B.3: Regression results from the $\text{GAM}_{\text{PC*PI}}$ model. '**' $p < 0.01$; '***' $p < 0.001$.

B.2 1-D smooth function plots



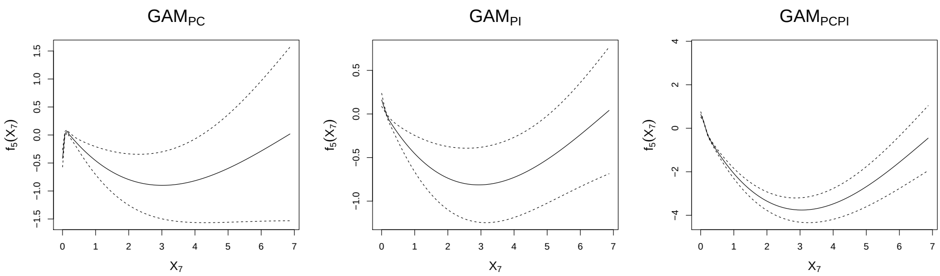
(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

Figure B.1: 1-D smooth functions for the standard deviation of PC, $f_3(X_5)$, for positional strategies GAMs.



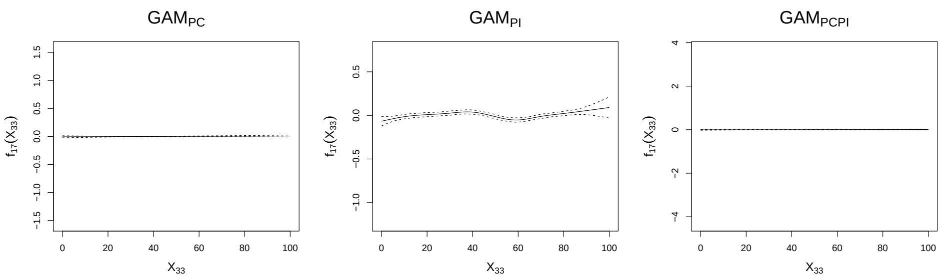
(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

Figure B.2: 1-D smooth functions for the standard deviation of PI, $f_4(X_6)$, for positional strategies GAMs.



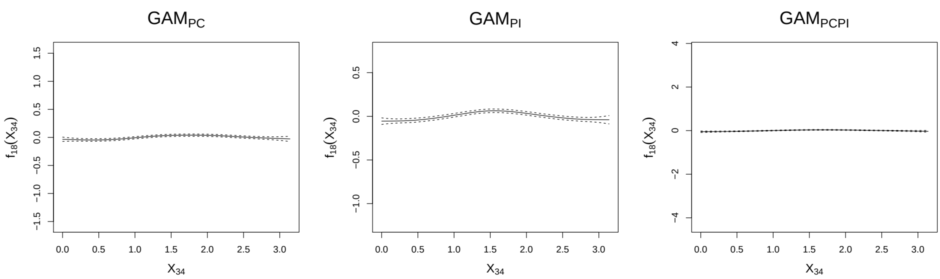
(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

Figure B.3: 1-D smooth functions for the standard deviation of PC*PI, $f_5(X_7)$, for positional strategies GAMs.



(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

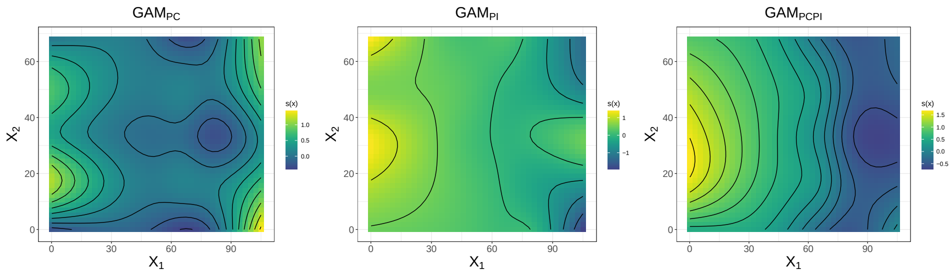
Figure B.4: 1-D smooth functions for the game time, $f_{17}(X_{33})$, for positional strategies GAMs.



(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

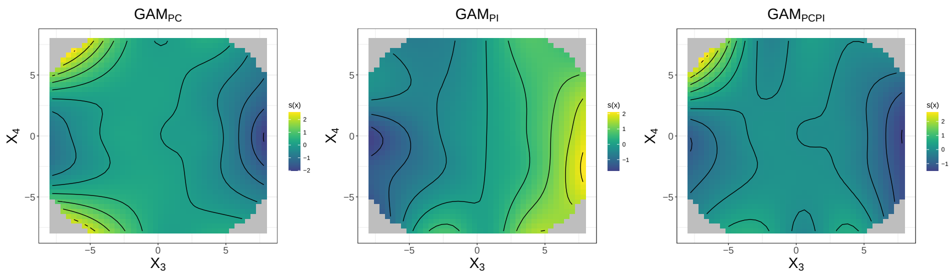
Figure B.5: 1-D smooth functions for the passing angle, $f_{18}(X_{34})$, for positional strategies GAMs.

B.3 2-D smooth functions plots



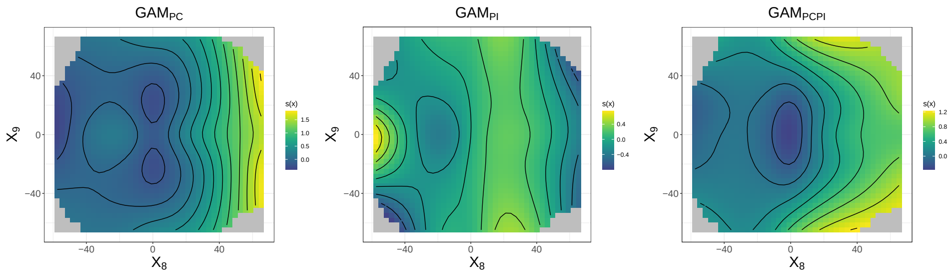
(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

Figure B.6: Heat map showing the contribution of smooths functions for the position of the player, $f_1(X_1, X_2)$, for positional strategies GAMs.



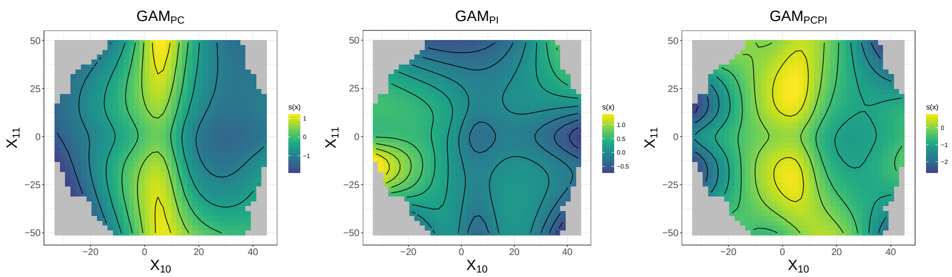
(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

Figure B.7: Heat map showing the contribution of smooths for the velocity of the player, $f_2(X_3, X_4)$, for positional strategies GAMs.



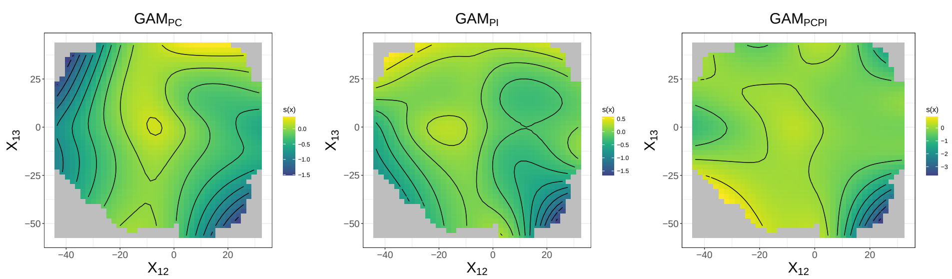
(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

Figure B.8: Heat map showing the contribution of smooths for the position of the ball, $f_6(X_8, X_9)$, for positional strategies GAMs.



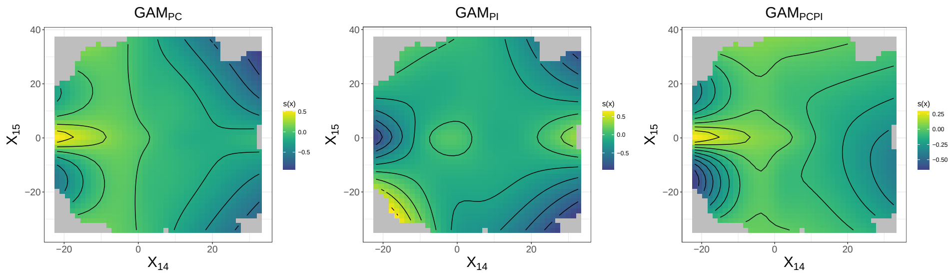
(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

Figure B.9: Heat map showing the contribution of smooths for the position of the team centroid, $f_7(X_{10}, X_{11})$, for positional strategies GAMs.



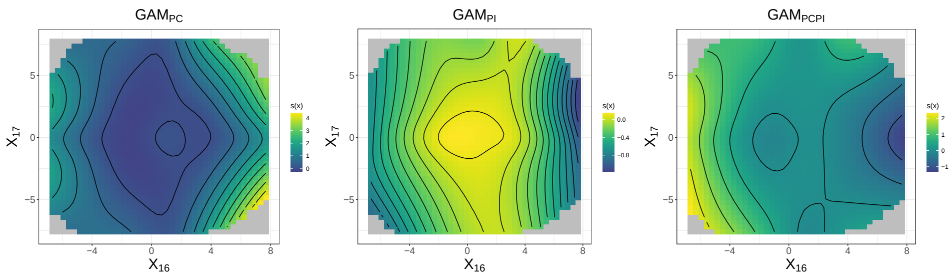
(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

Figure B.10: Heat map showing the contribution of smooths for the position of the opponent's centroid, $f_8(X_{12}, X_{13})$, for positional strategies GAMs.



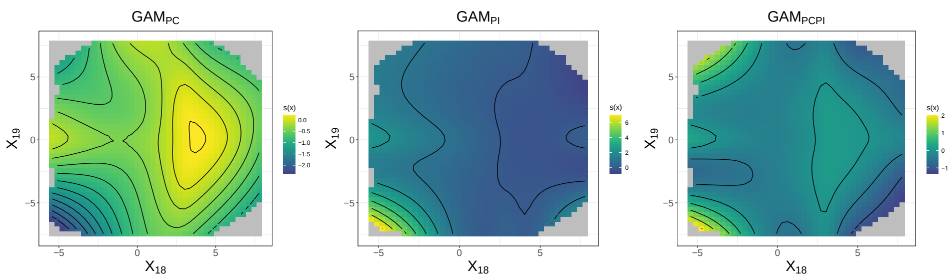
(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

Figure B.11: Heat map showing the contribution of smooths for the position of the closest opponent, $f_9(X_{14}, X_{15})$, for positional strategies GAMs.



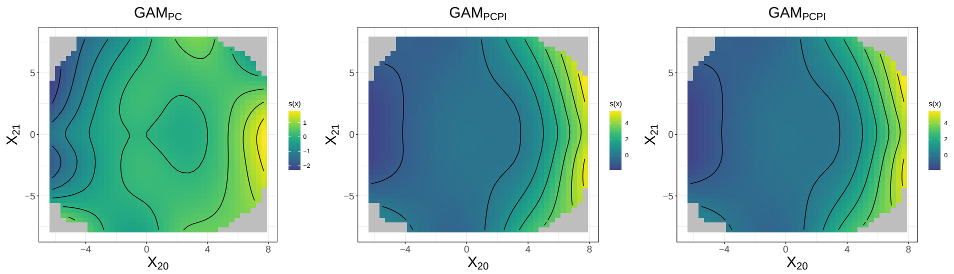
(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

Figure B.12: Heat map showing the contribution of smooths for the position of the optimal PC point, $f_{10}(X_{16}, X_{17})$, for positional strategies GAMs.



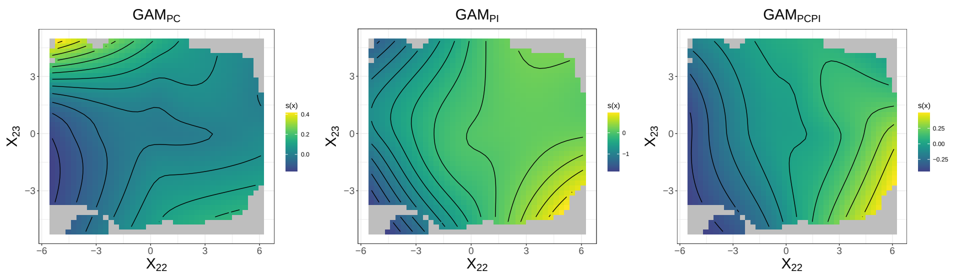
(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

Figure B.13: Heat map showing the contribution of smooths for the position of the optimal PI point, $f_{11}(X_{18}, X_{19})$, for positional strategies GAMs.



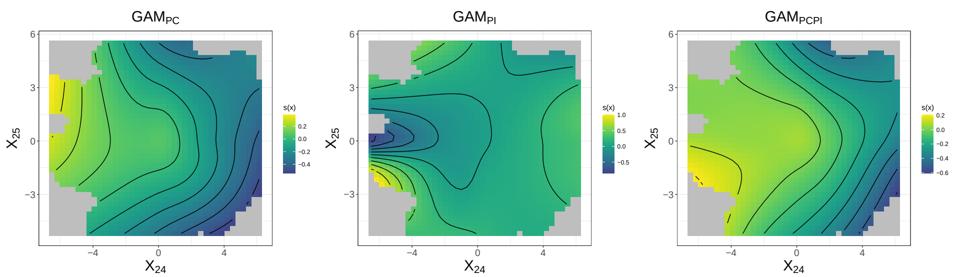
(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

Figure B.14: Heat map showing the contribution of smooths for the position of the optimal PC*PI point, $f_{12}(X_{20}, X_{21})$, for positional strategies GAMs.



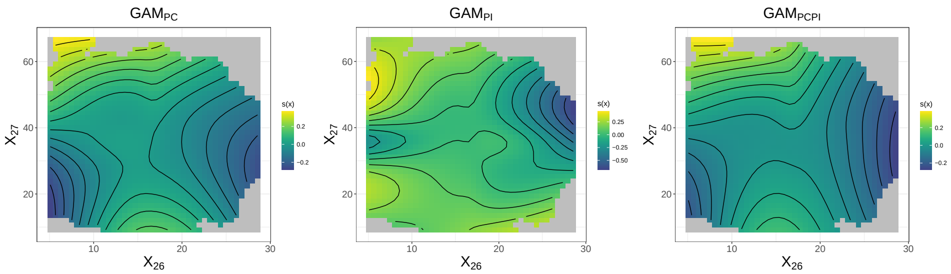
(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

Figure B.15: Heat map showing the contribution of smooths for the velocity of the team centroid, $f_{13}(X_{22}, X_{23})$, for positional strategies GAMs.



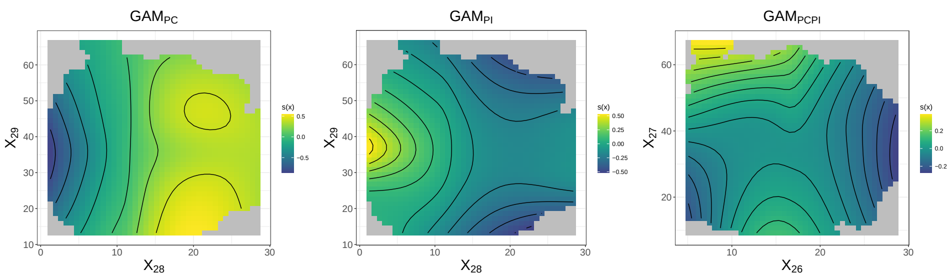
(a) GAM_{PC} **(b)** GAM_{PI} **(c)** GAM_{PC*PI}

Figure B.16: Heat map showing the contribution of smooths for the velocity of the opponents centroid, $f_{14}(X_{24}, X_{25})$, for positional strategies GAMs.



(a) GAM_{PC} (b) GAM_{PI} (c) GAM_{PC*PI}

Figure B.17: Heat map showing the contribution of smooths for team compactness, $f_{15}(X_{26}, X_{27})$, for positional strategies GAMs.



(a) GAM_{PC} (b) GAM_{PI} (c) GAM_{PC*PI}

Figure B.18: Heat map showing the contribution of smooths for opponents compactness, $f_{16}(X_{28}, X_{29})$, for positional strategies GAMs.

Appendix C

Regression results for GAM in Chapter 8

Smooth terms x		**Smooth terms y**	
Variable	Sign	Variable	Sign
$g_1(X_1)$	***	$h_1(X_1)$	
$g_2(X_2)$	***	$h_2(X_2)$	***
$g_3(X_3)$	***	$h_3(X_3)$	***
$g_4(X_4)$	***	$h_4(X_4)$	***
$g_5(X_8)$	***	$h_5(X_8)$	***
$g_6(X_9)$	***	$h_6(X_9)$	***
$g_7(X_{10})$	***	$h_7(X_{10})$	*
$g_8(X_{11})$	***	$h_8(X_{11})$	*
$g_9(X_{12})$	***	$h_9(X_{12})$	
$g_{10}(X_{13})$	***	$h_{10}(X_{13})$	.
$g_{11}(X_{14})$	***	$h_{11}(X_{14})$	*
$g_{12}(X_{15})$	***	$h_{12}(X_{15})$	***
$g_{13}(X_{22})$	***	$h_{13}(X_{22})$	
$g_{14}(X_{23})$	***	$h_{14}(X_{23})$	***
$g_{15}(X_{24})$	***	$h_{15}(X_{24})$	
$g_{16}(X_{25})$	***	$h_{16}(X_{25})$	***
$g_{17}(X_{26})$	***	$h_{17}(X_{26})$	***
$g_{18}(X_{27})$	***	$h_{18}(X_{27})$	***
$g_{19}(X_{28})$	***	$h_{19}(X_{28})$	
$g_{20}(X_{29})$	***	$h_{20}(X_{29})$	***
$g_{21}(X_{30})$	***	$h_{21}(X_{30})$	
$g_{22}(X_{31})$	***	$h_{22}(X_{31})$	

Fixed effects x		**Fixed effects y**	
Variable	Coefficient	Variable	Coefficient
Intercept	0.541***(0.005)	Intercept	0.040***(0.006)
X_{32}	0.002**(0.004)	X_{32}	0.019**(0.004)
$X_{33,CB}$	-0.233***(0.007)	$X_{33,CB}$	0.011 (0.009)
$X_{33,FB}$	-0.056***(0.006)	$X_{33,FB}$	0.005 (0.007)
$X_{33,MF}$	-0.137***(0.006)	$X_{33,MF}$	0.014 (0.007)
$X_{33,FW}$	0.139***(0.007)	$X_{33,FW}$	0.008 (0.008)

Table C.1: Significance of terms for the position prediction GAM developed in Chapter 8. Smooth terms and fixed effects for both the x- and y-direction for the multi variate model. For the categorical variable $X_{33,r}$ AM is chosen as reference. ’ ’ $p < 1$; ’·’ $p < 0.1$; ’*’ $p < 0.05$; ’**’ $p < 0.01$; ’***’ $p < 0.001$.

Tristan Kristoffer Cook, Eivind Keil, Inge

Modelling Off-Ball Decision Making

NTNU
Kunnskap for en bedre verden