University of Amsterdam

Masters Thesis

Simulating the pass after the pass: pass quantification through future scenario analysis

Author:
Ya'gel SCHOONDERBEEK

Supervisor:
M. Lees

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Computational Science

in the

Computational Science Lab
Informatics Institute

August 2021

Declaration of Authorship

I, Ya'gel SCHOONDERBEEK, declare that this thesis, entitled 'Simulating the pass after the pass: pass quantification through future scenario analysis' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at the University of Amsterdam.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date: 1 August 2020

"Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world."

Albert Einstein

UNIVERSITY OF AMSTERDAM

Abstract

Faculty of Science
Informatics Institute

Master of Science in Computational Science

Simulating the pass after the pass: pass quantification through future scenario analysis

by Ya'gel SCHOONDERBEEK

Inspired by the approach of chess-computers, this thesis investigates the possibilities of *future scenario analysis* in football, focused on movement and passing. In particular, the quantification of passes is studied. Pass Probability [1], Pitch Control [2] and Pitch Value [3] were reproduced and used to establish a pass option model. It predicted the passed-towards player for 22.4% of the passes; 42.6% if allowed two attempts. Both are 10% below field accomplishments [4]. The model, argued to generate optimal continuations, is still used. *Dead reckoning* is proposed as new baseline in movement prediction, which proves to be a robust predictor around 1 m. This model was extended with *ball orientation*, improving the predictions for the receiving player by 0.2 m. Both proposed models outperform field standards [5, 6].

Standard pass quantification models [7] were developed and extended: the number of out-passed opponents, computed via distance to the goal center, and the change in controlled pitch, based on Pitch Control and Pitch Value. Through the pass option and movement prediction models, future scenarios analysis was added to the pass quantification models. The open data set of Metrica Sports [8] was used to study correlations with performance indicators, after observational noise was reduced by a Kalman filter. Due to limited data, this was done per 5 minutes of match time. The number of out-passed opponents showed little correlation to shots, shots on target and ball-time in penalty area (respectively $r = 0.100, 0.046, 0.036$), the change in controlled pitch showed small correlations with the first two ($r = 0.342, 0.376$) and small negative correlation with the last metric ($r = 0.242$).

Incorporating future scenarios analysis for both pass quantification methods did not significantly improve the correlation with the descriptive metrics. The future number of outpassed opponents resulted in respectively $r = 0.016, 0.202, -0.085$; the future change in controlled pitch in $r = 0.378, 0.234, 0.048$.

Acknowledgements

First and foremost, I want to thank Mirjam Bruinsma, my supervisor from Ajax, for making this possible. Without her efforts and enthusiasm, I would never have had this opportunity. Her sharp mind and critical thinking were always of help and pushed the quality of my research to higher standards. The same goes for dr. Mike Lees, my supervisor from UvA, who was always able to pose the right questions and give the needed advise. I am grateful for my fellow students Bart van Laatum, Enrikos Iossifilidis, Kas Sanderink, Sam Verhezen and Roel van der Burght. I was always able to discuss my work with them, giving me new energy as well as a better understanding of my own work. Lastly, I want to thank my family and friends for all the support I have received.

Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

CSL	**C**omputational **S**ceince **L**ab
UvA	**U**niversiteit **v**an **A**msterdam
AI	**A**rtificial **I**intelligence
ABM	**A**gent **B**ased **M**odel
PPCF	**P**otential **P**itch **C**ontrol **F**ield
SPADL	**S**occer **P**layer **A**ction **D**escription **L**anguage
VAEP	**V**aluing **A**ctions by **E**stimating **P**robabilities
RMSE	**R**ooted **M**ean **S**quare **E**rror

Chapter 1

Introduction

Although it might be difficult to imagine now, someday the Football World Cup will be played once again in full glory; no COVID restrictions. Supporters in massive stadiums, cities decorated with all sorts of fan material; while countries from all over the world challenge each other in the world's most popular sport. Here, in Amsterdam, little orange plastic flags will hang from home to home across the little streets and windows will be showcasing posters, cuddly toys, drawings and shirts. No matter where you look, you will see orange and football. Somehow, during these magical weeks, everyone seems to love the game of football and joins together to support the national team.

In Netherlands, playing a World Cup final is a nightmare. We've lost it each of the three times: in 1974, 1978 and 2010. Johan Cruijff was not able to lead us to victory, Rob Rensenbrink remarkably hit the post in the very last minute (fig. 1.1) and Arjen Robben failed in two one-on-one chances against Casillas. The coming years, hope is in the new generation of F. De Jong, M. De Ligt and D. Van de Beek. Imagine that we would reach the final again during the next World Cup. The score remains 0-0 up until the 90th minute. The very last minutes of the extra time are being played. Frenkie de Jong recieves the ball at its own half. Beautifully, as well as typically, he dribbles deep into the opponents half, towards the targeted goal. You and your friends have gathered to watch this intense game. As you jump up from the sofa, you are shouting, screaming at the television: "Pass to the left! Pass to the left!!!" How could one know whether Frenkie de Jong should pass to the left? This thesis tries to explore exactly that question with a data-driven computational model. To ensure that De Jong will make the perfect pass and The Netherlands will be the next World Champion. Before diving into the details of the exact concept, let us briefly turn to the history of sport analytics and examples on the incredible influence of data.

FIGURE 1.1: The World Cup final in 1978: The Netherlands vs. Argentina. R. Rensenbrink hits the post and misses the opportunity to win the game in the very last minute of regular time. Unfortunately, the Netherlands lost in extra time.

1.1 A short history of sports analytics

The story of successful sports analytics starts in 2002: the Oakland Athletics achieve a winning streak of 20 games in American baseball. The spectacular achievement was last accomplished in the season of 1935 by the Chicago Cubs and is expected to be seen less than once per 25 years [9]. The well-known book and film *Moneyball* tells the story what enabled the Oakland Athletics to deliver this performance. The general manager Billy Beane used his so-called *sabermetrics* (statistics that measure in-game activity) to change the way of acquiring talent; with success. It is the first example of widely known sports analytics and embarks the start of data-driven sports management.

Ever since, the use of data has become rapidly popular in different fields. Several examples highlight how it became one of the primary factors for success in sport. Until 2009, the London-based rugby club Saracens had been one of the biggest under-achievers in the professional era, despite signing star international players. The appointment of Brendan Venter as director of rugby for the club was a turning point. Venter introduced an in-depth evidence-based approach with the coaching staff. They committed to using data to inform their decisions on team selection, game tactics, training priorities and player recruitment. Two years later, Saracens won the Premiership for the first time in their history and have developed to be one of the leading teams in European rugby.
When Bradley Wiggins won the Tour de France in 2012, he punched a button on his bike computer prior to punching the air with pride to celebrate. The power meter fed data through to an endurance training technology that gave insight into performance, form, and how to mould custom training programs. It enabled Wiggins to reach new personal performance heights and win one of the most prestigious tours.
Or take one of the greatest sports stories of all times: Leicester City winning the Premier

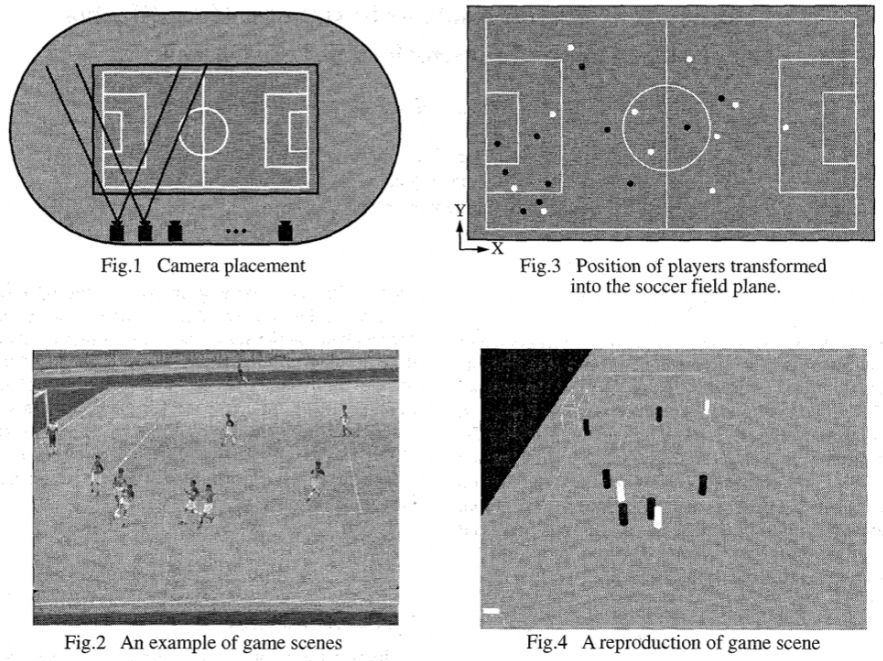


Fig.1 Camera placement

Fig.3 Position of players transformed into the soccer field plane.

Fig.2 An example of game scenes

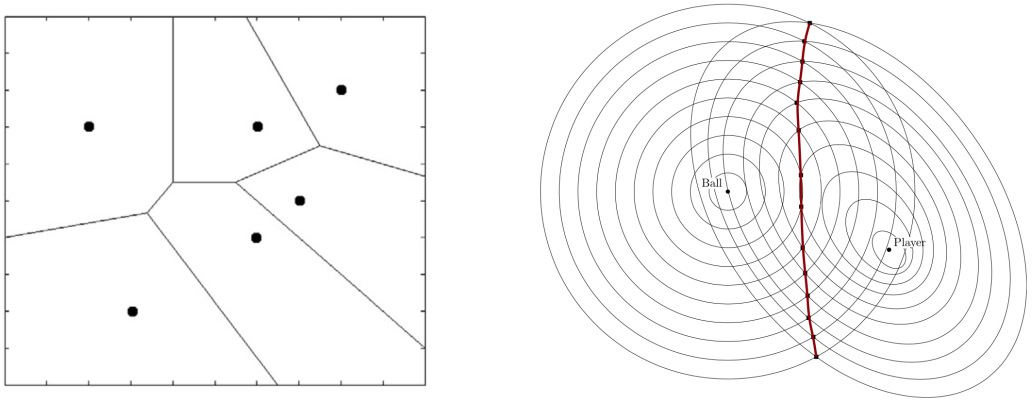
Fig.4 A reproduction of game scene

FIGURE 1.2: A system of tracking players by the use of cameras. Taken from one of the first published research on spatiotemporal tracking by Taki et al. [10].

League title of 2015/16. Throughout the history of the Premier League, every champion, until then, had finished in the top 3 in the season before winning the title. Leicester City was the exception, finishing the 2014/15 season in 14th place and 46 points behind winners Chelsea. Their incredible rise all came down to a simple, underlying reason – the club's use of data analytics and sports science was by far one of the most advanced in the Premier League. For years now, Leicester City has been using a number of different sophisticated data and analytical tools coupled with wearable technology to train right, play right, and strategise right. A culture of data-driven decision-making is deeply ingrained into the club's operational approach and proved to be the game-changer.

At this moment, football analytics is experiencing a shift from basic data analytics towards in-depth research on sport dynamics [11]. Pioneers in football analytics are pushing towards tactics based on computer models and elaborate data analysis. The reason for this is two-fold [12, 13]: i) the data available on football matches has experienced an explosive growth due to the development of automatic tracking systems; ii) the academic world of computational research on football is also rapidly growing and is starting to develop in a coordinated manner.

Key to this development is automated tracking data, which started around 1996. One of the first published research on spatiotemporal tracking data was by Taki et al. [10]. The researchers discussed a system of tracking the players in order to reconstruct the game (figure 1.2). Following this research, studies have focused on developing and automating



(A) A Voronoi diagram defines pitch control by the simple nearest-neighbor rule. The dots represent players and the lines mark territory borders. Taken from Fonseca et al. [14].

(B) An illustration of the model developed by Gudmundsson and Wolle [15]. Their discrete model calculates possible passing trajectories by drawing reachable eclipses for disretised time steps.

FIGURE 1.3: The first models that attempted to provide elaborate data analytics.

tracking systems in order to obtain high-resolution and high-frequency spatiotemporal trajectories of the players and ball.

After some time, the tracking data was precise enough to allow for more in-depth research and to dig for information past basic statistics such as the amount of covered space. In 2012, Fonseca et al. [14] proposed to model pitch control by Voronoi diagrams, which is considered the first model to map the influence of players on the field. It uses the simple nearest-neighbor rule: each player, represented by the coordinates of his/her location in the field, is associated to all parts of the field that are most near to that player. A simple visualisation of this is shown in figure 1.3a. They conducted a comparative analysis on the Voronoi diagrams of different plays, which showed different patterns of interaction between attackers and defenders. These results supported further investigation of the spatial dynamics of football. However, the model was clear to lack consideration of speed and direction of players. Spearman [2] proposed an elaborate model to overcome this disadvantage, to which we will turn later.

Two years later, Gudmundsson and Wolle [15] published their study, which discusses tools developed specifically for analysing the performance of players and teams. First, they conducted a passing analysis in the form of 2 tools. The first tool computes possible passing alternatives and the second tool computes frequent pass sequences. Possible passes are defined by trajectories for the ball for which one player of the same team reaches the ball first. This is calculated by eclipses of motion as shown in figure 1.3b. Next, they aimed to analyse the movement of players by the use of clustering: for a given trajectory, the goal is to find sub-trajectories. Lastly, they studied the correlation between clusters in order to identify team movements. For each of these tools, the

developed algorithm is outlined. However, as the researchers do not provide any validation of their models, this publication should be seen as an exploration of possibilities in modelling in football analytics rather than a reference for fundamental results.

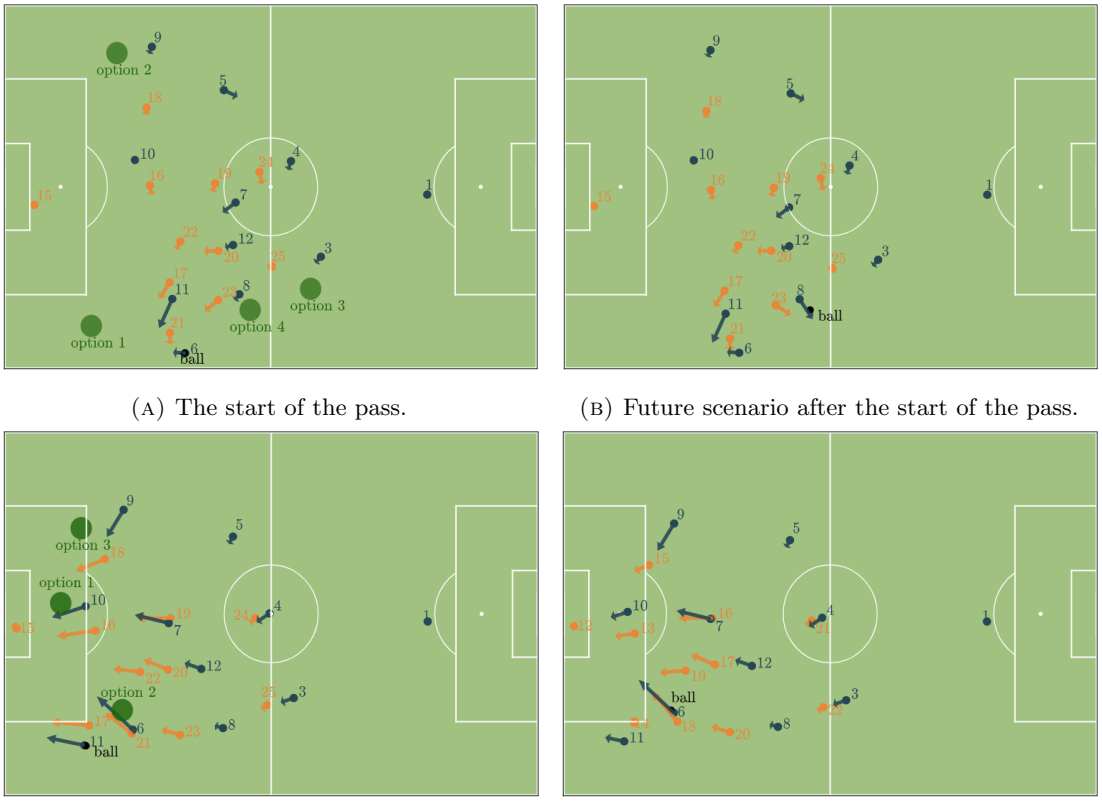
In more recent years, the MIT Sloan Sports Analytics[1] Conference has published a fair share of progressive studies within football analytics. For example, the article of Spearman et al. [1] on the probabilities of pass interceptions. The model forms a fundament to this thesis and will be revisited later on. Spearman [2] also proposed the aforementioned Pitch Control model in his study on a measure called *off-ball scoring opportunities* (OBSO). Similarly, the research done by Fernandez and Bornn [16] lays a foundation for recent and future studies. They also modelled the influence area of players based on their speed and acceleration in order to study space creation. Since the publication, the model has already been incorporated by many other researchers [17–19].
Other well-known examples are the study on rebounds by Maheswaran et al. [20]; the deep-learning study on expected possession value by Fernández et al. [21]; and the study on ghosting of players by Le et al. [22]. In line with other frontier models [23–25], current research often relies heavily on *Artificial Intelligence* (AI) models. Although AI-based models might lead to high predictive scores, an inherent problem arises for the scientific view. The accompanied black box phenomena obstructs a fundamental understanding of game dynamics [26]. It is exactly this trend that this study tries to counteract.

1.2 The approach of a chess-computer: pass quantification through future scenario analysis

In contrast to football analytics, algorithms and computers developed on chess analytics focus on complete game simulation. They attempt to evaluate all possible future scenarios by the means of complete quantification of the search space. Current algorithms are enhanced by simple tree pruning or AI to eliminate weak branches. Simulating chess has the advantage of a (relatively) comprehensible set of actions: there are 30 - 35 legal moves on average per turn [27]. Nevertheless, due to the exponential growth of the search space over the amount of turns, chess is unsolvable. John Tromp [28] estimated there to be roughly 10^{46} possible configurations on the board. This dazzling number is already far beyond the reach of any computer. Imagine how it translates to football; the number of possible positionings of 22 players and a ball is infinite.

Fundamentally, simulating all future scenarios of a football game is unrealistic. The computational power needed is beyond the physical limits of our world. This thesis sets to challenge the impossible by focussing on two types of actions: running and

[1] https://www.sloansportsconference.com



(A) The start of the pass.

(B) Future scenario after the start of the pass.

(C) The end of the pass

(D) Future scenario after the end of the pass.

FIGURE 1.4: The general concept of the envisaged model: the start of the pass (A) is quantified by its best future scenarios (B), which is compared to the end of the pass (C) and its best future scenarios (D). As such, it is expected to facilitate an in-depth analysis in passing.

passing. The resulting model, simulating the pass after the pass, forms a dissent from the AI-based research on pass quantification. It allows for an in-depth analysis and more fundamental understanding of the game. The aspiration to model future scenarios in football embarks the essence of this thesis:

How can the long-term impact of a single football pass be quantified? In particular, how do we evaluate all future scenarios after the pass and quantify their impact.

The first challenge in answering the above is establishing a reduction in the state space. The amount of considerable states must be downsized to both a tractable as well as representative part, i.e. states that represent prospective as well as valuable future scenarios. This thesis simplifies the search space by only considering two actions: running and passing. Further simplification can be achieved by approximating what movement of players is presumable and simulating only that. The arguments why the algorithm's approximation are valid will be emphasised in Methods (ch. 3) and Experimentation and Results (ch. 4). Once the possible future scenarios can be generated in manageable size, the value of these scenarios needs to be assessed.

A quantification system for future scenarios inherently allows for the envisioned in-depth analysis. Quantifying the value of situations and actions in football is a complicated problem. Where physicists are able to relate their mathematical models to measurements, there is no obvious, measurable truth to football. Incomplete passes are naturally invaluable, but received passes are difficult to relate to each other. Valuing passes is subject to tactical preference. One might envision that passing backwards is always bad; someone else might prefer passing on one specific side. The challenge is to capture the essence of football in a mathematical model, without a directly measurable metric available. Multiple studies have tried to overcome this problem by the use of multiple expert opinions or studying relations to goal-scoring statistics. Validation of the model and a proper comparison to other quantification methods is required. In summary, answering the following questions will result in an elaborate understanding of the research question:

How can we generate future scenarios in a representative, tractable way?
How is impact measured in order to quantify scenarios?
How can the quantification through simulation of future scenarios be evaluated?

The overarching approach of this study starts to take shape. In algorithm 1 and figure 1.4, the general concept is portrayed. Both the start and end of a pass are evaluated and analysed. For each, the idea is to generate possible progressions and simulated them. The value of those future scenarios provides extra insight in the value of the analysed scenario. As shown in figure 1.4, realistic simulation of the game is sought after in this concept.

Algorithm 1: General Concept
Result: Pass quantification
initialise *begin & end pass*;
compute *possible future scenarios for start & end pass*;
quantify *future scenarios*;
return *Pass value*

1.3 Outline

The aim of this thesis is to investigate the possibilities of simulating game progression in football. What can be expected in this exploratory study is briefly set out here. As illustrated, the computational size of the problem is massive. This thesis will be based on predictions in movement and passing, accompanied with certain accuracies. The achieved precision can be investigated via the data. On the basis of these results, the

significance of the resulting model can be discussed. The upcoming chapter Literature Review (ch. 2) is dedicated to such models. The existing research provides a context for the results that can be expected. In addition, different models will be compared in order to substantiates which models were incorporated in this study.

In the subsequent chapter Methods (ch. 3), the implemented work is elaborately set out, drawing upon the conclusions of the Literature Review (ch. 2). The biggest contribution of this study is in the field of movement prediction. The dead reckoning model that will be proposed shows exciting results. The model on passing options also resulted in accuracies comparable to the field standard. Lastly, future scenarios are generated and assessed by the use of standard models in football analytics.

The impact of these choices can be viewed in Experimentation and Results (ch. 4). In particular, the effect of the established future scenario analysis can be viewed. Together with Discussion (ch. 5), both the opportunities and shortcomings of the developed model are addressed and related to similar research. Finally, in the Conclusions and Future Work (ch. 6), the entire study is summarised. We look ahead on the possible improvements in the future. It is hoped that, some day, it will be possible to produce realistic game simulation.

Chapter 2

Literature Review

As has just been passed, data-based sports analytics is a young scientific field, which limits the comprehensiveness of this literature review. In this study, it was needed to reproduce existing work on Pass Probabilities, Pitch Control and Pitch Value in order to conduct a more in-depth analysis on passes. Below, models on the subjects are discussed and compared. The choices made are supported by contrasting them with the scientific background. Although the number of alternatives is limited, an argument for the use of this research can be based on the published results.

In addition, all known research on pass quantification [3, 7, 17, 25, 29, 30] is examined in detail. In contrast to the concept behind this paper, the majority of these studies do not focus directly on exploring possible game-developments. There is a research gap on the fact on the possibilities to analyse teams by complete game simulation, which this thesis aims to narrow. It is identified that the publications struggle to validate their models scientifically.

The concept of Agent Based Modelling (ABM) is considered to be the best approach on research through game simulation. This is a modelling technique where players (agents) would determine the game dynamics by means of individually simulated actions. The focus in this thesis is narrowed down to player movement and passing choices. Research on decision-making on these two actions are discussed below. In order to avoid false expectations: the model presented here is not an ABM, as it is not computationally possible yet. This thesis merely lays groundwork for future development of ABM's in football analytics. Current ABM-based work in both football- and other sport analytics is shortly discussed lastly.

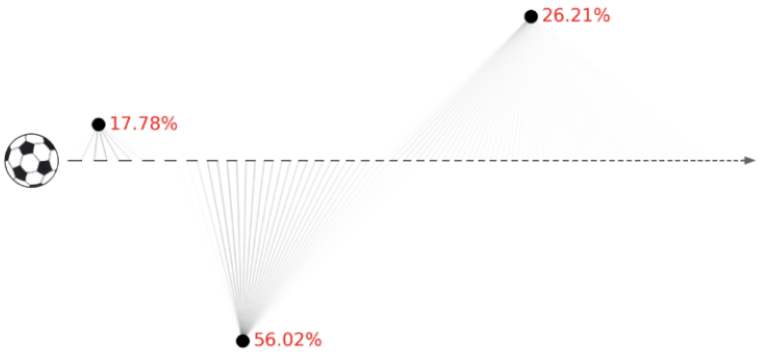


FIGURE 2.1: Pass Probability model by Spearman et al. [1]. Individual interception probabilities are found by approximating arrival times of ball and players.

2.1 Pass probabilities

In the philosophy of understanding passing dynamics, the paper of Spearman et al. [1] starts from a physics-based view. Rather than using predictive variables of passing situations, it aims to directly model the trajectory of a pass and interception probabilities along it. The probabilities of interception are approximated via stochastic distributions on the arrival times of players and the ball, schematically depicted in figure 2.1. Using a data set of 38 games played by Crystal Palace, the model was able to predict the outcome of a pass, i.e. completion or interception, with 81.9% precision and the specific receiving player with 67.9% (given the pass destination).

Two applications of the Pass Probability model are discussed. First, it is shown that a pass value can be based on this model. The resulting mean value per game correlates for 63% the number of shots of a team and for 83% with the number of passes in the attacking third. The same correlation coefficients for reception efficiency are respectively 64% and 70%, which highlights the usability of the model. Second, passing decisions were studied. The researchers investigated the relation between passing choices and the probability of the passes. Unfortunately, due to computational limitations, they were not able to complete a model that generates hypothetical passes and predict pass options.

Alternatively to the research of Spearman, pass probabilities have been analysed by the means of generalised additive mixed models [31, 32]. These studies focus on fitting smooth functions on variable related to passing situations, such as events before the pass, the pass being an header or the position of the pass. In figure 2.2, the heat maps on pass probabilities and the model error from McHale and Relton [31] are included. It compares the predicted probability of a pass being successful with the actual outcome. Despite its apparent predictive value, this method is unattractive for this study. The focus is not on

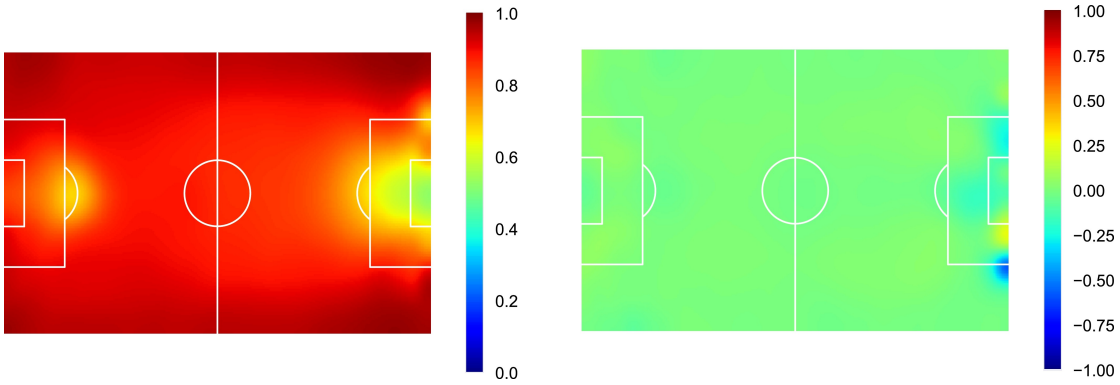


FIGURE 2.2: Heatmap on pass success probability (left) and model error (right) by the intended pass destination. Taken from the study of McHale and Relton [31]. Instead of the physical simulation of the pass, probabilities are estimated by situation related variables. The right figure indicates that model has very low error.

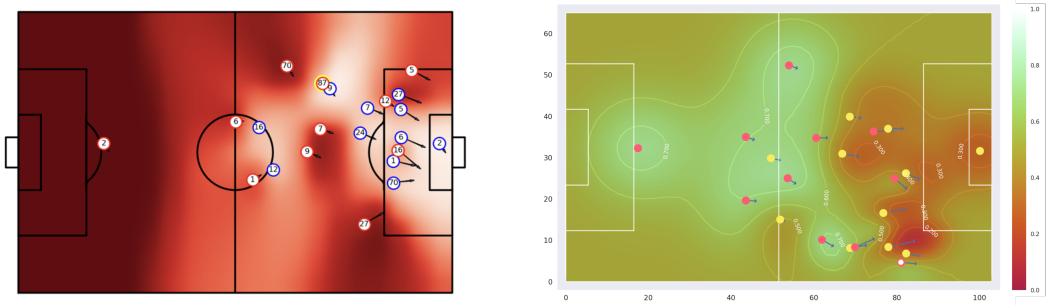
maximising the predictive value, but on a more fundamental understanding of passing through realistic simulation. Thus, the model of Spearman et al. [1] was incorporated. The technical implementation is discussed in section 3.2.1.

2.2 Pitch control

In addition to creating pass options, a team works to improve its influence or control on the pitch. During an attacking play, defenders position themselves with the aim to optimise their control on the defending zone. This can result in a positional advantage later on in the game. Vise versa, attackers reposition themselves when the team is defending. Players constantly try to control parts of the pitch that might be useful in the future. Modelling this control on the pitch provides a fundamental basis to understanding situations, especially when aiming to quantify future scenarios.

As aforementioned, Fonseca et al. [14] proposed the first model on pitch control based on a simple nearest-neighbor rule, shown in figure 1.3a. Although it captured the physical phenomenon of control in a straightforward manner, the model has two serious drawbacks. First, it does not account for the physical reality of running (accelerating and moving towards the ball). One player might be slightly closer, but running at full speed in the opposite direction. For such cases, this player should not be appointed as the player with the most influence. Second, the model considers control to be binary; either one team has control of a region or the other. In reality, both teams have comparable control over some regions.

The two drawbacks have led to the demand for a probabilistic model, based on more realistic arrival times of players. Two similar models were proposed by Spearman [2]



(A) Pitch Control as proposed by Spearman [2]. It is based on a logistic distribution taken over the arrival times of both players and the ball.

(B) Pitch Control as proposed by Fernandez and Bornn [16], based solely on the arrival time of players.

FIGURE 2.3: Two comparable probabilistic models on control over the pitch. The left figure models control over remote regions, e.g. the left upper corner. The other model considers only quickly reachable regions.

and Fernandez and Bornn [16], both visible in figure 2.3. The former aims to compute the trajectory of an air pass in order to estimate the arrival time of the ball to a pass destination on the pitch. This is combined with the expected arrival time of the player and the control of a player. The model is fitted to data of 58 matches. The proposed measure correlates 26% with future scoring, i.e. goals scored in a subsequent match. In contrast, the amount of shots correlates only 17% and the amount of goals only 12% with the amount of goals in the subsequent match; illustrating that the proposed measure is an improved indicator of scoring. A promising result, but it does not appear to lead to any strong conclusions on the model.

Fernandez and Bornn [16] do not simulate the actual pass and compute control probabilities without the arrival time of the ball. They consider the probability density function of a bivariate Gaussian distribution for the influence of players. The logistic function is used to map the subtraction of the accumulated team influence within the $[0, 1]$ range. Based on their Pitch Control model, they achieve a quantification of pitch value. The underlying assumption is that defenders try to cover important parts of the pitch. Therefore, the most valuable part of the pitch should be covered most in similar situations. The researchers analyse the covered space by training an AI model on an extensive data set of 20 matches of first and second Spanish division. They apply the Pitch Control and Value model to provide game analysis on space generation of players. Although there is no comprehensive validating work on which the two models can be compared, the model of Spearman was incorporated by this thesis. The underlying thought behind the model imposes a stronger approach to reproduce reality.

2.3 Pass quantification

In football analytics, there is currently no widely accepted method for pass quantification, whilst it would be valuable. As Spearman [2] argues: "*current match analytics shown to the public mostly discuss superficial statistics, e.g. shots and passing accuracy. This does not capture the intrinsic v[illegible] of specific passes or shots.*" Some research has been done on the matter; several pass quantification models are based on scoring probabilities [3, 25, 29], whilst others try to produce an analysis on the space-creation in a game [7, 17, 33]. Here, these methods are treated with the aim to identify successful aspects as well as shortcomings, setting a reference frame for the proposed model of this thesis.

2.3.1 Goal-scoring probabilities based

Decroos et al. [25] introduce an interesting framework to analyse actions of players, which is also applicable to other actions than passing. Most notably, a data storage format is proposed: *Soccer Player Action Description Language (SPADL)*, which has been incorporated by many clubs. The model analyses actions based on approximated probabilities of producing and conceding goals, computed for both the starting and resulting situations of an action. This measure is called *Valuing Actions by Estimating Probabilities (VAEP)*. The probabilities are calculated by two classification models trained on 11,565 European top leagues games. In addition to the SPADL features, the models consider distances covered, position to the goal and the score. With this model, Decroos et al. are able to deliver a top 10 player list, which summed market value is higher than tables based on goals or assists. They also show examples of evaluating and comparing players and their playing styles. The publication lacks an elaborate validation, substantiating that the model matches reality. The researches point out that the model also lacks analysis of any off-the-ball actions.

Likewise, Power et al. [29] quantify passes by studying the probability of scoring in 10 seconds after the pass. The study looks at two components of a pass: the risk and the reward. The study poses that these two can be quantified by the probability of the success and the probability of scoring, approximated by the use of logistic regressors. The researchers trained linear models on 352,466 examples. The best predictor resulted in a log-loss of 0.21 in predicting whether a pass is completed and a log-loss of 0.14 in predicting whether a shot occurred within 10 seconds after the pass. Where the article poses interesting ideas for the applications of the model, it too lacks validation or verification of the quantification system. Where the logistic regressor model is validated, the results of the final model are not compared to specific game dynamics.

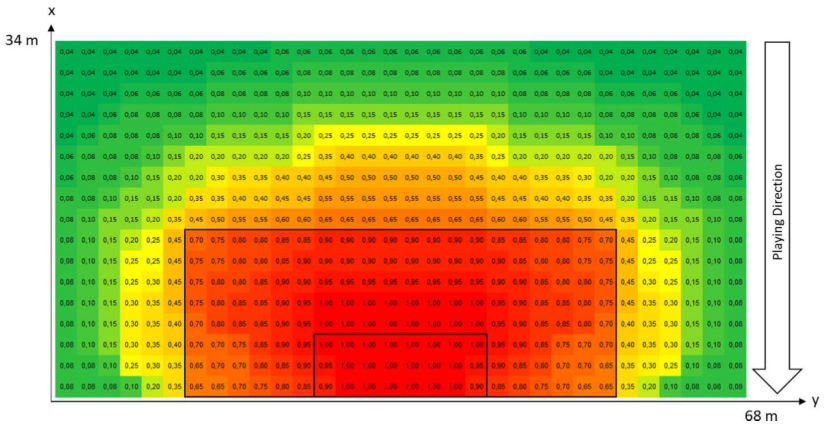


FIGURE 2.4: Pitch Value as implemented by Link et al. [3], which was used for the Pitch Value model in this thesis (ch. 3.2.3).

The study of Link et al. [3] also quantified passes based on goal-scoring probabilities, using the term *dangerousity*. The dangerousity is based on the spatial constellation of the players and ball, modelled by four components: i) Zone, the danger of goal-scoring, quantised by a theoretical grid (fig. 2.4); ii) Control, the extent of ball control, quantised by the speed of the ball & player; iii) Pressure, the possibility of defensive prevention, quantised by the defender's distance to the player in possession and defender's angle to the goal; and iv) Density, probability of success, quantised by distances of defenders in line of trajectory. The proposed dangerousity measure shows 82% correlation with the probability of winning a match in an analysis on 64 games in the Bundesliga. This is significantly higher than the other performance indicators examined: shots on goal (58%), passing accuracy (56%) and ball possession (71%). Win probabilities were derived from betting odds, which is questionable as scientific approach.

2.3.2 Space-creation based

Contrary to the above, Goes et al. [ ] try to quantify the defensive disruption due to a pass. They identify that the studies mentioned above lack insight in the match dynamics through their focus on goal scoring probabilities. The proposed measures *D-Def* and *I-Mov* are computed solely from the positional data of players. Using principal component analysis, they show that individual movement is highly correlated with descriptiveness of the defensive organisation. It is argued that different passes can be identified based on their D-Def score, as it yields major differences between top-, average-, and low performance passes. By the use of a multiple linear regression model, Goes et al. explore the possibilities of predicting the I-Mov and D-Def scores based on the pass characteristics. Unfortunately, the published results do not indicate a particular correctness of

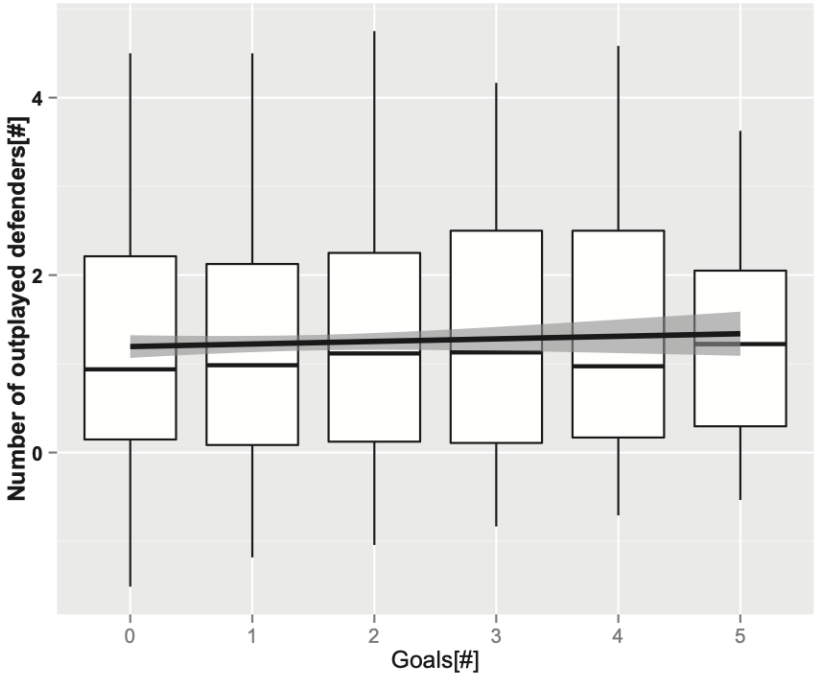


FIGURE 2.5: The validation by Rein et al. [7] on their pass quantification; the number of outplayed defenders against the number of scored goals during a game. The correlation of the average pass evaluation and the number of goals scored was found to be $\chi^2(1) = 4.67$, $p < 0.05$.

the model. Nevertheless, the concept shows potential and supports the aim for a deeper understanding of game dynamics.

Similarly, the research conducted by Rein et al. [7] aims to quantify space creation. The paper sets out to calculate the number of defenders between the ball carrier and the goal as well as the control of space by the use of Voronoi-diagrams [14] for 103 German First division games. These are used to quantify a pass. The researchers carry out a chi-squared test to compare the passing statistics to the number of goals scored and to the probability of winning a game. The test indicated small correlation with the number of goals for the average change in space control ($\chi^2(1) = 5.74$, $p < 0.05$) and for the average number of outplayed opponents ($\chi^2(1) = 4.67$, $p < 0.05$). The results for the latter are shown in figure 2.5. The correlation with game success was similar, found to be respectively $\chi^2(1) = 4.7$, $p < 0.05$ and $\chi^2(1) = 5.4$, $p < 0.05$. Surprisingly, no correlation was found with the number of shots. They are able to identify types of passes that score well/bad on either of the two measures. There is room for improvement, e.g. on the Pitch Control model, which can lead to a more precise pass quantification model and stronger correlation with the performance indicators. As the approach fits this thesis, this model will be build upon for assessing passes. It is hoped that by extending the pass quantification model, a stronger correlation with the performance metrics is found.

In the extension of the research done by Rein et al. [7], the company Impect[1] has introduced the measure Packing, which has grown popular in football analytics. As explained by Scott [33], it poses that the only measure of importance is the amount of outplayed opponents. The method has been picked up by many clubs, which gives strength to the article of Rein et al. Unfortunately, Impect has not published any (scientific) results to share its ideas.

2.3.3 Other techniques

Lastly, it the research of Chawla et al. [30] should be mentioned. They used expert analysis in order to cluster passes into categories *Good*, *OK*, or *Bad*. The study uses machine learning techniques to determine the classification based on features that were subtracted in their research. The subjectivity of the experts can be considered as a drawback of this method. However, they are able to achieve 90% accuracy in predicting the category chosen by experts for a pass. Moreover, the agreement between their model and a specific observer is close to the agreement between different observers, which suggests that significant improvements are improbable. Although this is study is promising in the use of post-match analysis for coaches and trainers, it does not provide insight in the underlying dynamics of the game.

In conclusion, the majority of current pass quantification models lack insight in passing dynamics. Although AI-based algorithms show promising results as predictors, they are limited in providing a better understanding of reality. In contrast, a study on pass quantification through game simulation might provide it. As such, it would be possible to recognise passes that do not seem interesting themselves, but result in valuable possibilities in the future. As seen above, most studies struggle with validating their pass quantification model. Seemingly, studying the correlation of the average pass value of a team with a descriptive statistic, e.g. number of passes in final third or goals scored, is the most straightforward option.

2.4 Action decision in football

In order to simulate future scenarios, the individual behaviour of players must be identified; a common practice in ABM. The model must be able to simulate the action decisions of the players in a game in a realistic manner. This thesis focuses on two

[1] https://www.impect.com/en/

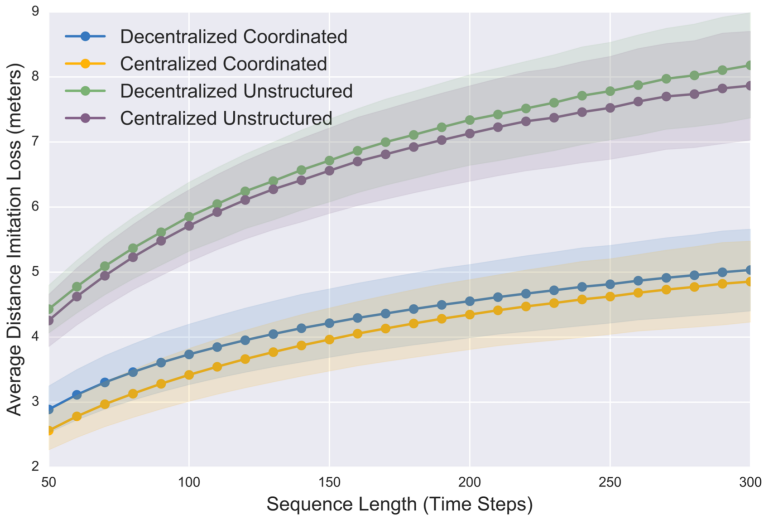


FIGURE 2.6: Results published by Le et al. [5] on movement prediction of players, for time steps of $0.1s$. The best model is able to predict the locations with an average accuracy ranging from 2.5 - $5.0m$ for sequences of 0.5 - $3.0s$.

types of actions: player movement and passing decisions, of which an overview is presented here. Future research might include other actions, such as dribbling, shooting and tackling.

2.4.1 Player movement

The prediction of the spatial behaviour of players is fundamental to simulating future scenarios. Current research on the movement prediction of players primarily use AI-models. For example, the research by Le et al. [5, 22] based on Deep Imitation Learning. Via a neural network model, Le is able to estimate the movement of a league average team, but also compare the defensive behaviour of a league top-team (Manchester City F.C.) to a mid-league team (Swansea City A.F.C). The results presented show an increase in the loss of distance to the real movement over time. For a sequence of 1.5 seconds, shown in figure 2.6, the model shows an average of 4 meters difference from the actual final position. In line with this research, Felsen et al. [34] were able to predict future positioning of basketball players with 5.74 $ft.$ $(1,75\ m)$ accuracy by the use of a neural network model.

Recently, Alguacil et al. [6] published their study on player movements with the shared aim to eventually model players as self-propelled particles, i.e. agent based modelling. Although the achieved predictive value proved to be invaluable, the idea behind to model of great interest for this study. After developing a Pass Probability-, Pitch Control- and Pitch Impact model, the researchers posed the hypothesis that offensive players will move such that these three aspects would be optimised. Unfortunately, the current position

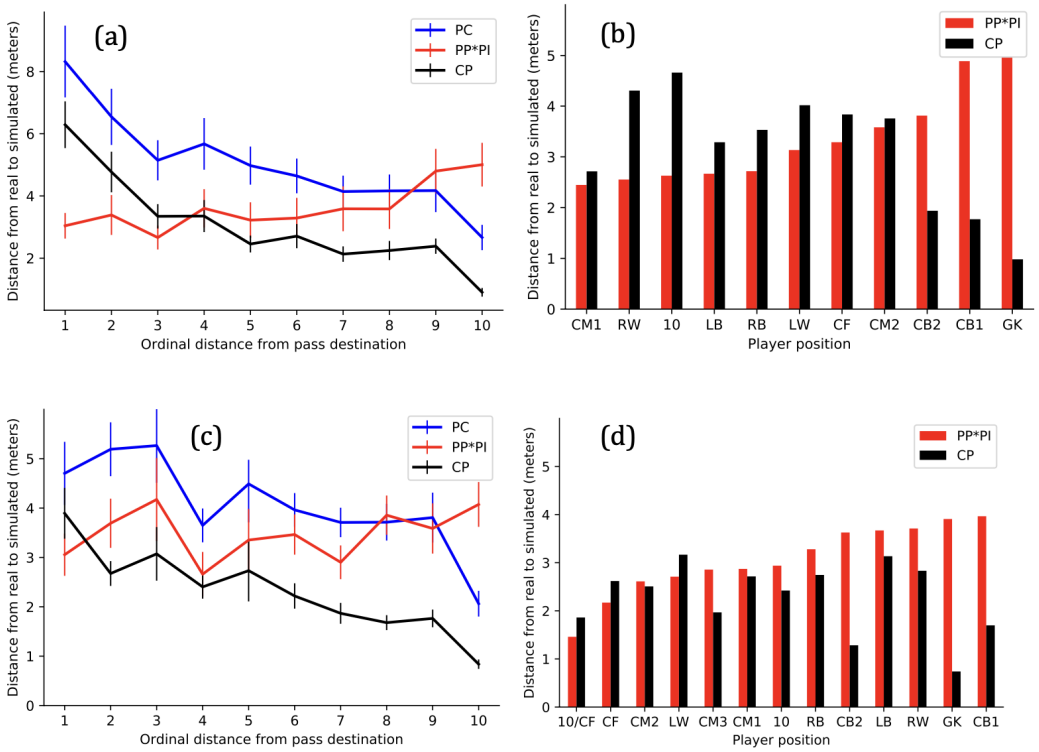


FIGURE 2.7: Best results of movement prediction as modelled by Alguacil et al. [6] for two different matches, one shown in the upper plots and the other in the lower plots. The distance to the observed final position is ploted. In (a) & (c), the i^{th} closest player to the ball is given on the x-axis with error bars indicating standard errors. In (b) & (d) show the results organised by player position.

proved to be a better predictor than any combination of these three aspects for the position of a player after a pass, shown in figure 2.7. This calls for further investigation of the influencive factors on the movement of players, done in this thesis.

2.4.2 Pass decisions

In order to identify possible passing options and the resulting future scenarios, it is also necessary to model passing decisions. Current research focuses on predicting which player is passed towards. Little is published on generating hypothetical options. Steiner et al. [35] were able to predict this with 23% accuracy. When allowed two guesses, this increased to 41%. In their paper, binary logistic regressions were implemented to test the relations between the predictor variables (openness of passing lane, position relative to the ball carrier, spatial proximity, and defensive coverage) and the dependent variable (passing decision). Similarly, Glöckner et al. [36] studied passing & shooting decision in handball. Their neural network model achieves an accuracy of the same order.

The research of Fournier-Viger et al. [4] analyses passing decision based on spatial characteristics. Their model is based on choosing the closest player to the ball after applying

penalties based on the distance to the defenders. They obtain a prediction accuracy of 33.8%, and 51.81% if two guesses are allowed. This simple estimation proves to be effective and is interesting to this thesis. It argues that the core dynamics at play are distance to the ball and coverage.

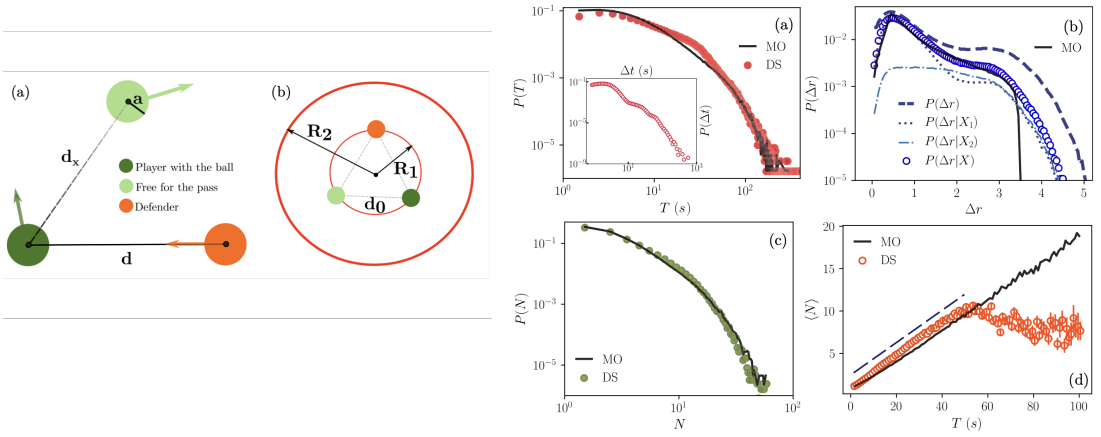
The model Fournier-Viger et al. [4] could be improved by the use of Pitch Control and Pitch Value, as can by supported by the study of Carlos Núñez and Dagnino [37]. They model decision making by applying a weighted evaluative function based on Pitch Control, Expected Possession Value and Expected Goals. In contrast to the other studies, this model was developed in the simulated environment provided by Google [38] for reinforcement learning. No particular comparison can be made with other studies on passing options, but their resulting competitive score is promising. Maintaining the focus on distance by Fournier-Viger et al. [4], passing options could be modelled by applying a distance gradient on Pitch Control and Pitch Value.

2.5 Agent based modelling and sport analytics

In order to put the previously discussed research into the perspective of developing an ABM, ABM-based research in sports analytics was shortly assessed. As mentioned, the current computational power is too little to develop an ABM on football. However, as turns out, particular questions can be studied by narrowing the scope of the model. Alguacil et al. [6] took the first step in the direction of this modelling technique, but an elaborate model is yet to be developed. A single publication can be found on football and one on basketball.

Chacoma et al. [39] developed a three-player ABM on football, i.e. two attackers and one defender (fig. 2.8a). The attackers' main goal was to pass the ball to each other. Movements were generated by random distributions and accepted if the position was not too close to the defender. Despite its simplicity, the model was able to capture, to some extent, statistical behaviour of possession times, pass lengths, and number of passes performed (fig. 2.8b). Although the resulting dynamical behaviour was realistic, the concept behind the model leaves a lot of room for improvement. Once it is computationally possible to extend the model to full game simulation, movement decisions can be investigated more fundamentally.

The study of Oldham and Crooks [40] investigates the psychological hot-hand effect in basketball. The model is developed in 3D NetLogo, a common language within the ABM field. The resulting scoring distributions showed resilience to observation, but not significant enough. Similar to the three-player ABM, the results are rather promising



(A) 3 player football ABM (B) Dynamical behaviour of model (MO) and data (DS)

FIGURE 2.8: Figures taken from the study of Chacoma et al. [39]. By simple simulation of three players (A), the researchers were able to achieve parallels with observed game dynamics (B).

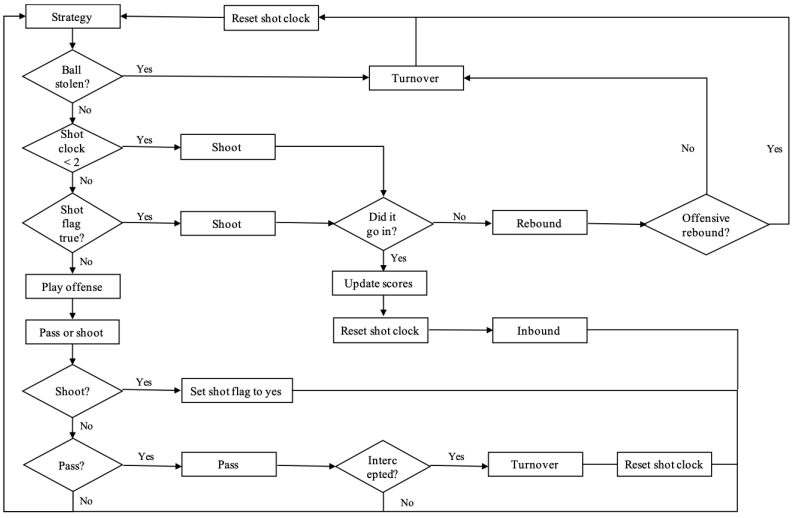


Figure 2: The play cycle of the model.

FIGURE 2.9: Decision scheme taken from the ABM model on basketball of Oldham and Crooks [40], which might serve as inspiration for future research on ABM in football.

than groundbreaking. However, the decision making for the basketball players might serve as inspiration, as it is the first study to explore an elaborate set of possibilities. It is split into a linear scheme, depicted in figure 2.9.

Chapter 3

Methods

Based on the arguments presented in the preceding chapter, existing models [1–3, 6] were reproduced. The models serve as base for the research done in this thesis. Building on this base, models were developed on movement prediction, passing options and pass quantification. Prior to the technical implementation of the existing and developed models, the data framework is discussed in order to ensure the reliability of the results.

Three existing models were reproduced. The work on pass probabilities and pitch control is based on the research of Spearman [1, 2]. Inspired by Alguacil et al. [6] and Shaw [41], approximations on ball- and player trajectory were applied in order to accelerate the algorithms. The pitch value model was based on the research of Link et al. [3] and statistical research on goal-scoring probabilities [24, 42].

After establishing the above, predictive models on player movement and pass options were required to generate future scenarios. In comparison to the research of Alguacil et al. [6], the positions of players were predicted after a pass. The simple concept dead reckoning [43–45] is proposed as new baseline. A ball-oriented variant was developed to improve the predictive value. For pass options, a combination of pitch control, pitch value and a distance gradient to the ball was developed to identify pass options. The model follows the thought of Fournier-Viger et al. [4], albeit it is fundamentally different.

Lastly, quantification models on scenario evaluation were developed, inherently allowing for pass quantification through future scenarios. The the change in control over import regions of the pitch [7] and the popular technique of outplayed opponents [7, 33] were studied. Both provide insight in the value of the play transition. By combining the models on movement prediction, passing options and situation evaluation, a pass quantification model through future scenarios was achieved.

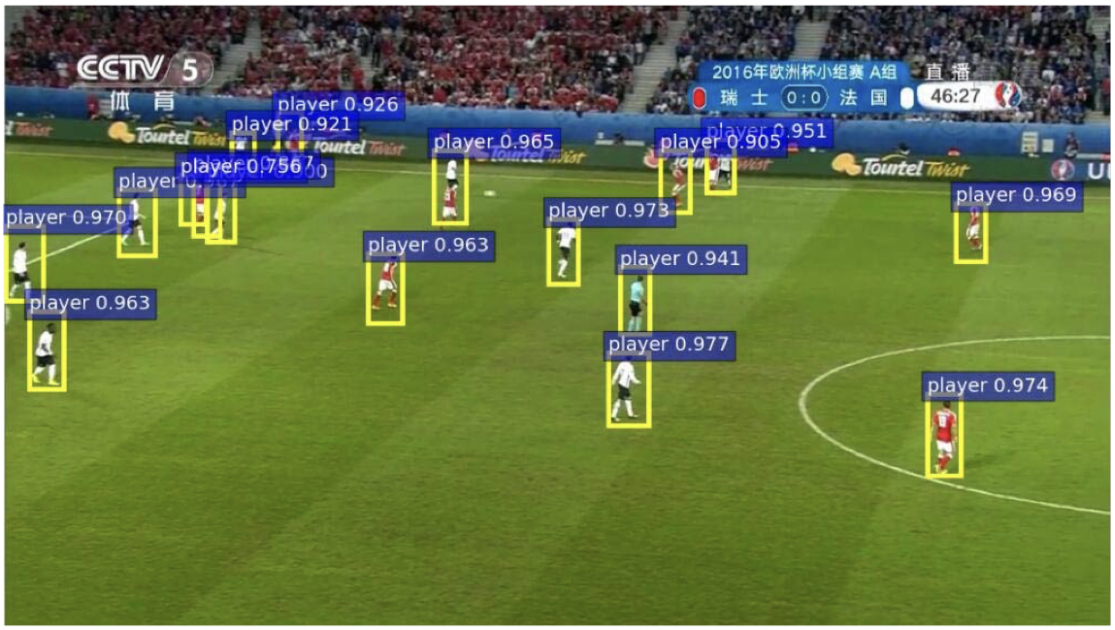


FIGURE 3.1: Image of the object detection model as published by Zhou et al. [50]. Observing the detection of some players closely demonstrates how errors arise in the tracking data. For example, the players at the top middle of the picture or the players at the corner of the penalty area seem ill-recognised.

3.1 Data framework

3.1.1 Software and open source codes

For this study the following (scientific) software programs were used: Python [46], NumPy [47], Numba [48] and Apache Spark [49]. Python and NumPy are two of the most popular programming languages for science and data analytics. However, as interpreted languages, the two are considered too slow for high-performance computing. Acceleration of the software might prove fundamental to achieving practical running time. Numba is an open source *Just-in-time (JIT) compiler* that translates a subset of Python and NumPy code into fast machine code. Apache Spark is an open source computing framework that unifies streaming, batch and interactive big data workloads; it realises computation in parallel. An extensive exploration of the data was realised by combining Numba and Apache Spark.

Additionally, models were partially based on open-source codes published by SciSports Labs[1], CleKraus[2] and Friends of Tracking Data[3].

[1] https://github.com/soccer-analytics-research [2] https://github.com/CleKraus/soccer_analytics
[3] https://github.com/Friends-of-Tracking-Data-FoTD

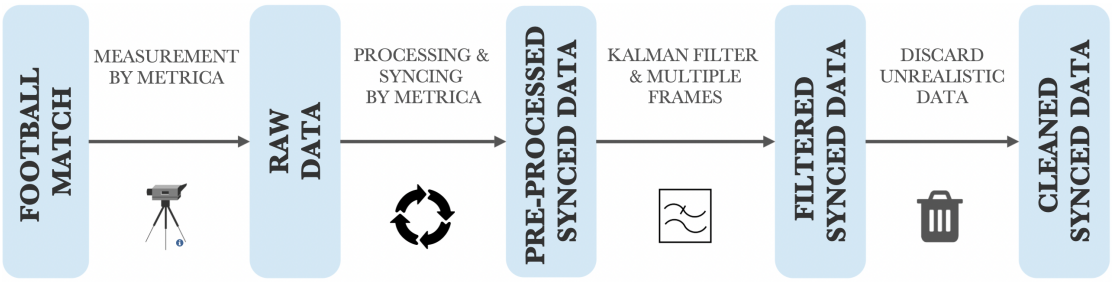


FIGURE 3.2: Schematic overview of the data processing. The tracking- and event data is generated by Metrica Sports Inc. [8], who also pre-process and synchronise the two. This study applies a Kalman filter to the tracking data in order to smooth out errorr and calculates speed based on multiple frames. Remaining data containing unrealistic speeds or accelerations are discarded.

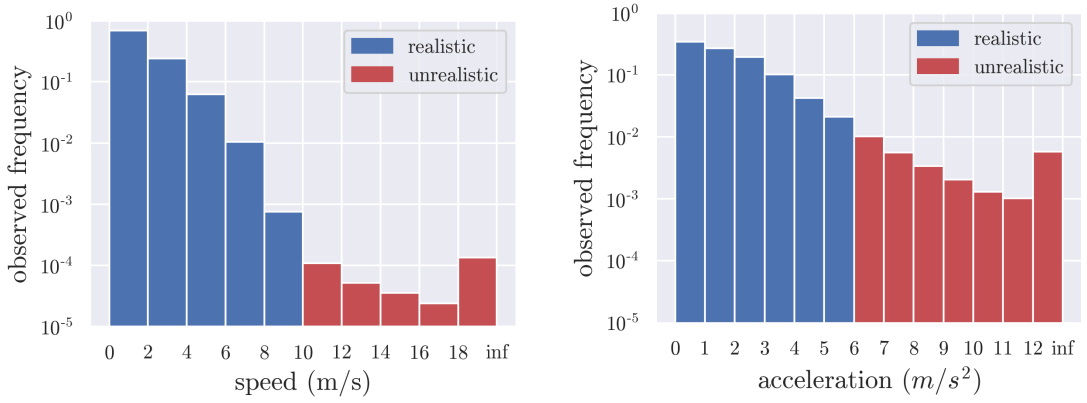
3.1.2 Synchronised event and tracking data

In order to analyse the positional settings during passes, it is necessary to have tracking data of passing events. This thesis aims to compute detailed models and requires the data of a substantial amount of passes; it requires synchronised event and tracking data. Generating this on large scale is currently one of the biggest challenges within football analytics. The reason behind this is the fact that tracking data is generated by image detection models and event data is hand-coded. The spatial-temporal features of an event as recognised by the eye often differ from the computer detected features. Disparities in time and position vary over different events, complicating the process of synchronisation.

Fortunately, Metrica Sports published an open data set [8] of synchronised event and tracking data, used in this study. The data set is anonymised and therefore it is not known which teams or players are playing. The tracking data consists of the two-dimensional positions of the players and the ball, generated at a sample-frequency of 25 Hz ($\Delta t = 0.04\ s$). The event data describes the type of the observed events for the corresponding time-frames, the player(s) involved and the location.

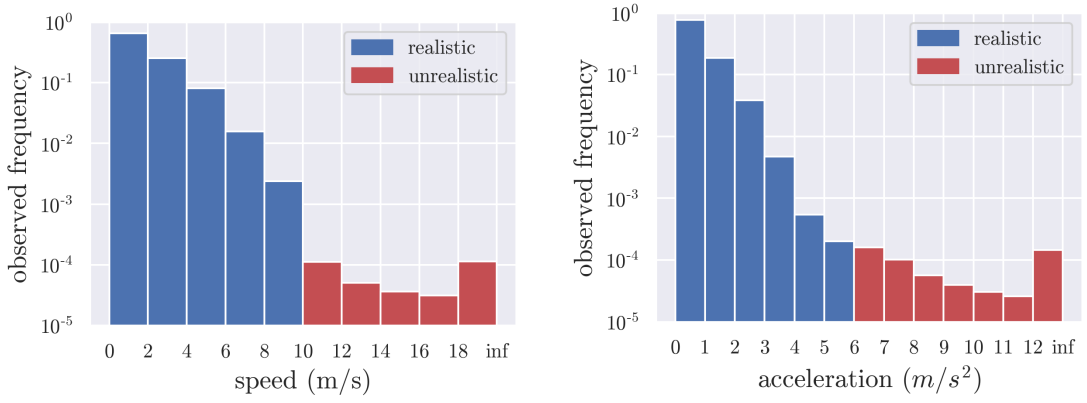
3.1.3 Observational noise in tracking data

Intrusive observational noise and errors are inherent to tracking data. Across all tracking techniques, e.g. GPS, object detection or WiFi, a substantial part of the research focuses on reducing the errors [51–53]. The tracking data treated here is generated by object detection. The errors found in this technique mainly arise from ambiguous images. When players are sprinting or stand too close to each other, the exact location of their bodies becomes ambiguous on a simple two-dimensional video image. The example from



(A) Speeds (B) Accelerations

FIGURE 3.3: The observed errors of *the raw tracking data* from the Metrica Sports open data set [8]. The data, sampled at 25 Hz, describes two games. The red bars indicate unrealistic observations in speed (A) or acceleration (B); this covers respectively 0.036% and 2.919% of the data. The latter in particular urges for thorough data processing.

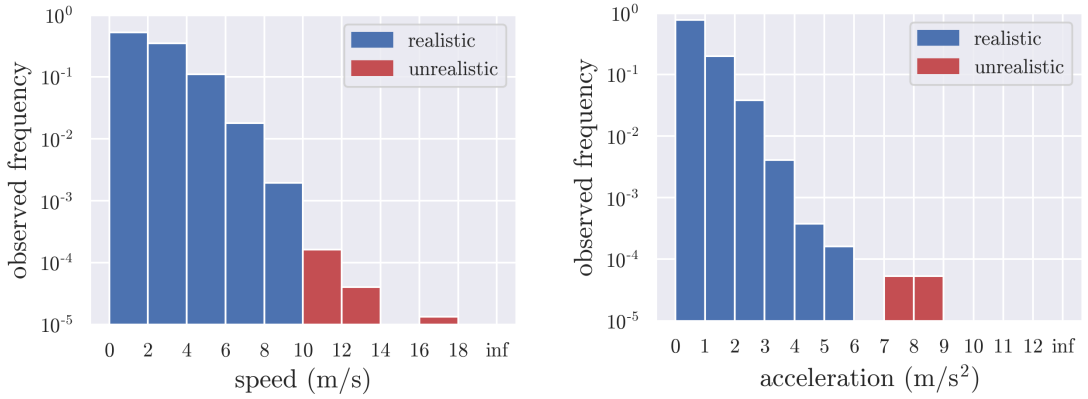


(A) Speeds (B) Accelerations

FIGURE 3.4: The observed errors of *the Kalman filtered tracking data* from the Metrica Sports open data set [8]. The data, sampled at 25 Hz, describes two games. The red bars indicate unrealistic observations in speed (A) or acceleration (B); this covers respectively 0.035% and 0.056% of the data. Compared to figure 3.3, this implicates a significant increase in the reliability of the data.

the publication of Zhou et al. [50] (fig. 3.1) illustrates this. By elaborate training of the algorithm, human correction and the use of multiple cameras, errors can be reduced.

Unsurprisingly, observational noise can be recognised in the open data set of Metrica Sports [8], as depicted in figure 3.3. The maximum speed seen across all professional football is approximately 10 m/s [54, 55]. Similar, football players have a maximum acceleration, situated somewhere between 5–6 m/s^2 [56]. The observed speeds and accelerations above these two threshold are signs of incorrect measurements. This covers respectively 0.036% and 2.919% of the data. The latter in particular urges for thorough data processing.



(A) Speeds (B) Accelerations

FIGURE 3.5: The observed errors after speed calculation over multiple time-frames in the passing events. The red bars indicate unrealistic observations in speed (A) or acceleration (B); this covers respectively 0.024% and 0.011% of all data. 9 of the 1682 passes are unusable.

3.1.4 Data processing

Figure 3.2 shows an overview of the data processing done for this thesis, including the work of Metrica Sports. From two football matches, Metrica Sports has generated event and tracking data, which were pre-processed and synchronised as well. In this study, the data was improved by Kalman filtering and speed calculation over multiple frames. The remaining passes with any unrealistic observation were ignored. The concept and technical implementation of the Kalman filter are treated in Appendix A.

3.1.4.1 Smoothing by Kalman filter

In figure 3.4, the results of the Kalman filter on the tracking data are visible. The frequency of unrealistic accelerations in the data has decreased from 2.919% to 0.056% (fig. 3.4b). One must note the sharp increase in low accelerations (1 - 2 m/s^2) from 34.2% to 76.9%. This possibly indicates over-smoothing by the Kalman filter; a negative impact on the data. The effect on the observed speeds (fig. 3.4a) negates this suspicion: the frequency of unrealistic speeds has merely decreased from 0.036% to 0.035%.

Surprisingly, the frequency of low speeds (1-2 m/s) has decreased from 68.7% to 64.7%, while the frequencies of higher speeds (2 - 10 m/s) have increased from 31.2% to 35.3%. This is possibly an effect of the glitches in the object detection along the running trajectory, where constant running at 6 m/s might be registered as 2 m/s and 12 m/s. Overall, the reliability of the data set appears to be significantly increased. The effect of the Kalman filter is further discussed in chapter 5.

3.1.4.2 Speed calculation over multiple time-frames

The remaining (small) share of unrealistic data were hoped to be overcome by speed calculation over multiple time-frames. This is a sensitive operation, as game dynamics might vanish due to the use of large time windows. The amount of frames was set 5. The resulting speeds and accelerations for all passes are shown in figure 3.5. Unfortunately, there is still unrealistic data present, made up by 9 passes per the total of 1682. These passes were discarded, due to the small share they occupy.

3.2 Reproduced work

3.2.1 Pass Probability

The first model developed for this thesis was a model that generates possible ground passes and corresponding success probabilities. The algorithm was heavily based on the work of Spearman et al. [1] and Alguacil et al. [6]. The published thesis of Alguacil [57] served as support. Inspired by the adaptation of Shaw [41], arrival times of players were approximated on the basis of reaction time, current velocity, maximum speed and maximum acceleration.

3.2.1.1 Trajectory simulation

In order to simulate the trajectory of a pass, an equation of motion is required. Based on a small personal study, Alguacil et al. [6] found that the trajectory of the ball in a ground pass is mainly determined by the aerodynamic drag for the first two thirds of the time. In the final third, the friction with the grass is the dominating force. This results in the following equation of motion:

$$\ddot{\vec{r}} \approx \begin{cases} -\frac{1}{2m}\rho C_D A \dot{r}\dot{\vec{r}} & \text{for } t \leq \frac{2T_{max}}{3} \\ -\mu g \hat{r} & \text{for } \frac{2T_{max}}{3} < t \leq T_{max} \end{cases} \tag{3.1}$$

where r is the two dimensional position vector of the ball, $m = 0.42\ kg$ is the mass of the ball, $\rho = 1.225\ kg/m^3$ is the density of the air, $C_D = 0.25$ is the drag coefficient, $A = 0.038\ m^2$ is the cross-section area of the ball, $\mu = 0.55$ is the coefficient of static friction (based on the average of FIFA recommendations for artificial grass) and, lastly, $g = 9.81\ m/s^2$ is the gravitational constant. Note that $\hat{r}$ is the normalised positional vector, i.e. the direction of the ball velocity.

$\boldsymbol{v_0}$ (m/s)	1	2	3	4	5	6	7	8	9	10
$\boldsymbol{T_{max}}$ (s)	0.528	1.065	1.575	2.043	2.472	2.856	3.201	3.510	3.783	4.029
$\boldsymbol{v_0}$ (m/s)	11	12	13	14	15	16	17	18	19	20
$\boldsymbol{T_{max}}$ (s)	4.250	4.447	4.627	4.787	4.933	5.068	5.189	5.230	5.402	5.497

TABLE 3.1: Resulting end times T_{max} for each starting speed v_0 of pass trajectories.

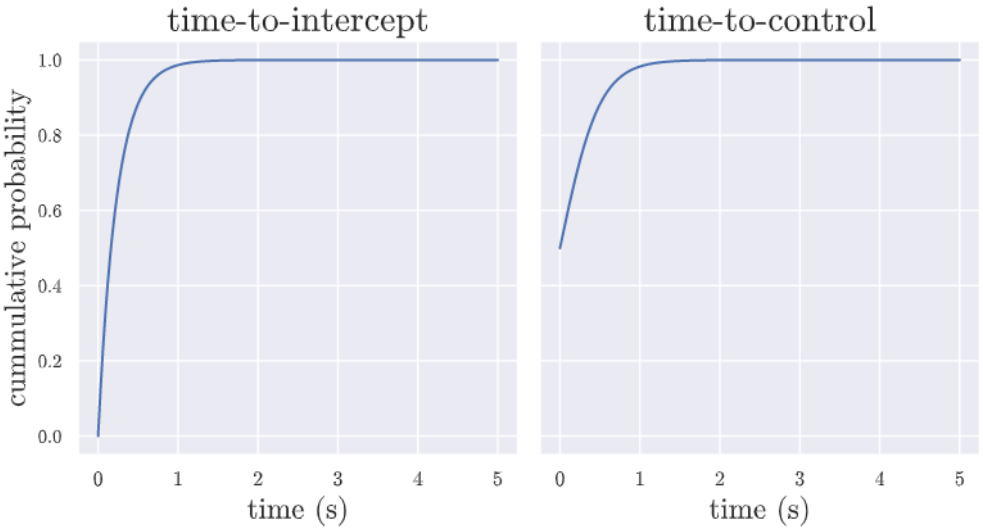


FIGURE 3.6: Distributions considered by Spearman et al. [1] to model interception (left) and control probabilities (right), respectively the logistic distribution and exponential distribution. The corresponding parameters $\sigma = 0.45$ and $\lambda = 4.3$ are taken from the parameter tuning done by Spearman et al.

Calculation of the trajectory is done by numerical integrating these two differential equations. However, this requires an end time T_{max} of the pass, dependent on the initial velocity. These values were calculated by simulating the trajectory for increasing T_{max} until the resulting velocity was approximately zero ($\leq 0.01\ m/s$). This was done for each of the velocities $v_0 = 1, 2, ..., 20\ m/s$. The results are shown in table 3.1

3.2.1.2 Interception time and -probabilities

By being able to generate all potential pass-trajectories, it is now possible to investigate the interception probabilities of all players during the trajectory. The model of Spearman et al. [1] is based on the time for players to reach the ball and the time for player to control the ball, respectively called time-to-intercept and time-to-control. Spearman et al. compute arrival times by solving a constrained minimisation problem on the equation of motion for players, a computational costly method. Therefore, it was chosen to follow the approach of Shaw [41] and Alguacil et al. [6] in considering a more simple approximation.

The simple approximation on the time-to-intercept is based on the reaction time of a player, i.e. the time between the player noticing the pass and the player actually acting upon it. The approximation is as follows: *first*, the player dispositions during the reaction time (eq. 3.2); *second*, the required acceleration time to achieve maximum speed towards the ball is computed (eqs. 3.4 & 3.3); *third*, the displacement during the acceleration time is estimated based on average speed (eqs. 3.5 & 3.6); and *fourth*, the remaining time required to move towards the ball is calculated, in case the ball was not reached yet (eq. 3.7). Mathematically, this is described as

$$r_{react} = r_{pl} + v_{pl} \cdot t_{react} \tag{3.2}$$

$$v_{optimal} = \frac{(r_{ball} - r_{react})}{\|r_{ball} - r_{react}\|} \cdot V_{max} \tag{3.3}$$

$$t_{acc} = \|v_{optimal} - v_{pl}\| / a_{max} \tag{3.4}$$

$$v_{acc} = \frac{v_{optimal} + v_{pl}}{2} \tag{3.5}$$

$$r_{acc} = \begin{cases} r_{react} + v_{acc} \cdot t_{acc} & if \ \frac{\|r_{ball} - r_{react}\|}{v_{acc}} > t_{acc} \\ r_{ball} & else \end{cases} \tag{3.6}$$

$$t_{int} = \begin{cases} t_{react} + t_{acc} + \frac{\|r_{ball} - r_{acc}\|}{v_{max}} & if \ \frac{\|r_{ball} - r_{react}\|}{v_{acc}} > t_{acc} \\ t_{react} + \frac{\|r_{ball} - r_{react}\|}{v_{acc}} & else \end{cases} \tag{3.7}$$

where r_{react}, t_{react} are the reaction dispositioning and -time; r_{pl}, v_{pl} are the position and velocity of the player, $v_{optimal}$ is the optimal velocity towards the ball; r_{ball} is the pass destination; V_{max} is the maximum speed of the player; a_{max} is the maximum acceleration; t_{acc}, v_{acc}, r_{acc} are the acceleration time, -average velocity and -displacement; and t_{int} is the time-to-intercept. The values used were $t_{react} = 0.26\ s$ [58], $v_{max} = 10\ m/s$ [54, 55] and $a_{max} = 6\ m/s^2$ [56].

Given the time-to-intercept, the probability that the player will intercept the ball at time T considered to be described by the logistic distribution [1]. Given variance σ, the probability density is

$$P_{int} = \frac{1}{1 + e^{-\frac{T - t_{int}}{\sqrt{3}\sigma/\pi}}} \tag{3.8}$$

In addition to the interception probability, Spearman et al. proposed to include a control probability in order to model the physical reality to be able to control the ball. This was taken into account via an exponential distribution, which depends on the control parameter λ. The probability of controlling the ball within t seconds is given by

$$P_{control} = 1 - e^{-\lambda t}. \tag{3.9}$$

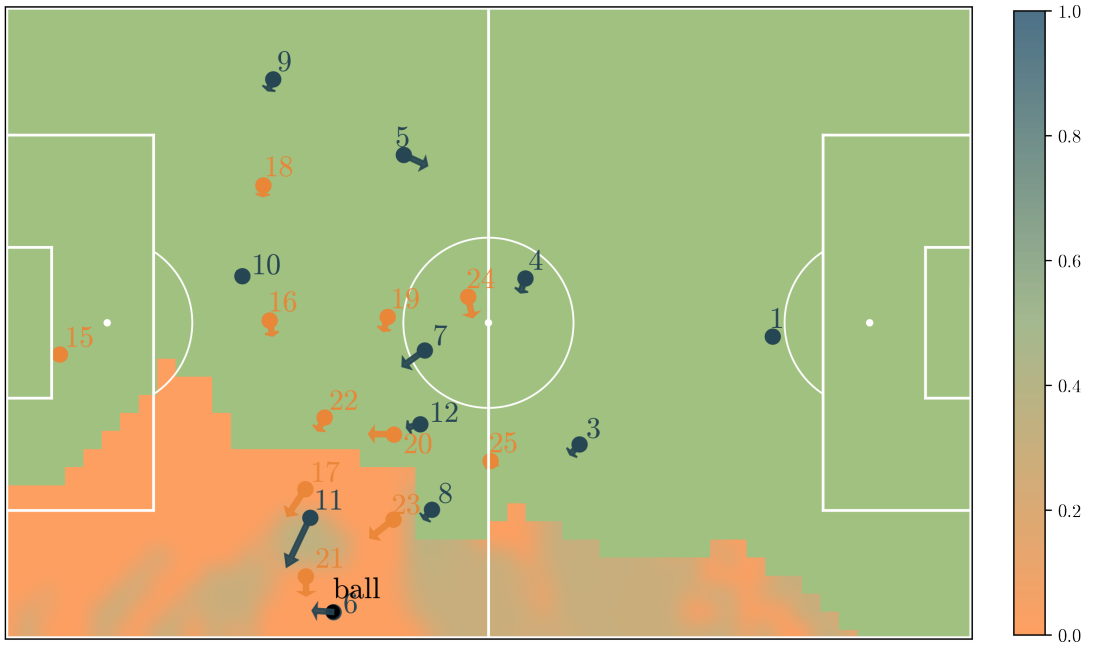


FIGURE 3.7: Resulting heat map from the developed Pass Probability model for a situation of the Metrica open data set [8]. As can be seen, only a part of the pitch can be reached by ground passes. Due to coverage of the orange team, only passes along the side of the pitch are expected to be successful.

The equations 3.8 and 3.9 each contain a free parameter. The values found in the parameter tuning of Spearman et al. were adopted: $\sigma = 0.45$ and $\lambda = 4.3$. The resulting distributions are plotted in figure 3.6. Using these values, it was possible to numerically approximate the integral

$$P_j(t) = \int_0^t \left(1 - \sum_k P_k(t')\right) P_{int}(t')\lambda \, dt' \tag{3.10}$$

describing the probability of player j receiving the pass at time t. The numerical solution was computed by discretising the time steps $dt \approx \Delta t = 0.04\ s$, corresponding to the sample frequency of the data. By summing over all players of a team, the total receiving probability can be obtained. An example of this is given in figure 3.7.

3.2.2 Pitch Control

In addition to modelling ground passes and their probabilities of success, it was necessary to gain insight on the area of influence of players in the field. The Pitch Control model as implemented by Spearman [2] was reproduced. As mentioned in the proceedings of Alguacil et al. [6], Pitch Control can be used to approximate success probabilities of air passes. Furthermore, a combination of Pass Probability-, Pitch Control- and Pitch Value models allow for an simple evaluation of situations.

In order to calculate this influence, Spearman proposed the following measure called *potential pitch control field (PPCF)*. It assumes that, while in proximity to the ball, a player's ability to make a controlled touch on the ball can be treated as a Poisson point process. It allows for a translation of the principle "who arrives first has a higher potential control". The original paper models the time of flight of the ball by considering air drag and the angle of the trajectory. In contrast, this research considers an average ball speed of 20 m/s to calculate the time of flight; accelerating the model.

In line with the Pass Probability model described above in section 3.2.1, the PPCF for each player j is computed via an integral. The differential term is given by

$$\frac{dPPCF_j}{dT}(t, \vec{r}, T|s, \lambda_j) = \left(1 - \sum_k PPCF_k(t, \vec{r}, T|s, \lambda_j)\right) f_j(t, \vec{r}, T|s)\lambda_j \tag{3.11}$$

where $f_j(t, \vec{r}, T|s)$ is the probability that player j on time t reaches location $\vec{r}$ within time T. The logistic distribution is taken to model this [2], which is given by

$$f_j(t, \overrightarrow{r}, T|s) = \left[1 + e^{-\pi \frac{T-\tau(t,\vec{r})}{\sqrt{3}s}}\right]^{-1} \tag{3.12}$$

where τ describes the approximated arrival time of the player (given by eqs. 3.2 - 3.7) and s describes the temporal uncertainty on player-ball intercept time. Spearman notes that "attacking players want to make a precise controlled touch that results in a shot or continued possession while a defender will often be satisfied with heading the ball away or kicking it out of play." To account for this, the control rate λ_j was considered to be different when defending or attacking, i.e.

$$\lambda_j = \begin{cases} \lambda & \text{if attacking} \\ \kappa\lambda & \text{if defending} \end{cases} \tag{3.13}$$

The parameters, found by Spearman in the parameter tuning of $s = 0.54$, $\lambda = 3.99$ and $\kappa = 1.72$, were adopted in this study.

Combining the above, results in an integral from the arrival time of the ball T_0 to infinity

$$\int_{T_0}^{\infty} \left(1 - \sum_k PPCF_k(t, \vec{r}, T|s, \lambda_j)\right) \left[1 + e^{-\pi \frac{T-\tau(t,\vec{r})}{\sqrt{3}s}}\right]^{-1} \lambda_j dT \tag{3.14}$$

By simultaneous numerical integration, this equation can be evaluated and the control of each player is obtained. Via summation it gives the control of the entire team. Note that the left term in eq. 3.14 ensures normalisation. Assuming small enough discretised time-steps, $PPCF_k(t, \vec{r}, T|s, \lambda_j) << 1$. As such, the integral will not become larger

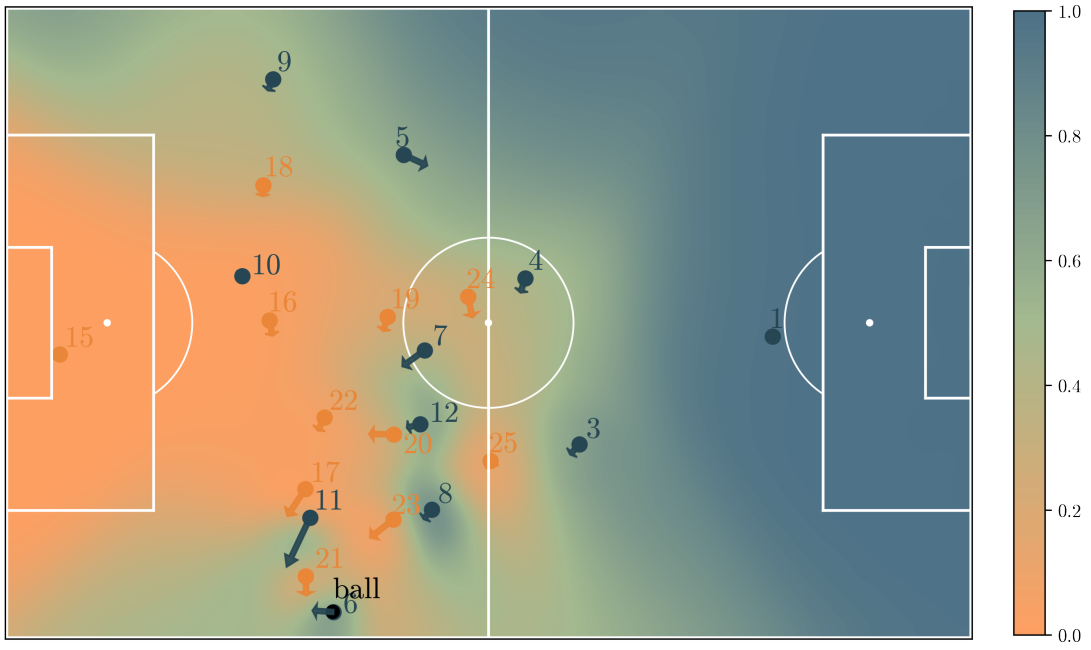


FIGURE 3.8: Heat map depicting the developed Pitch Control model for the same situation as shown in figure 3.7. In contrast to the Pass Probability model, Pitch Control produces values over the whole pitch. Again, the blue team has more control along the side of the pitch. Note that player 10 exercises no control as he is offside.

than 1. Moreover, the left term is a non-increasing function that approaches zero for $T \to \infty$, implying that the integral converges over time. The numerical computation can be stopped below a convergence tolerance, e.g. 0.01.

3.2.3 Pitch Value

In football analytics, Pitch Value models are linked to goal-scoring probabilities. The concept behind this originates in the fact that a team must score goals in order to win. The same view is incorporated in this thesis. However, rather than using a dynamical model (often AI based [6, 21, 59]) it was chosen to consider Pitch Value to be static.

Both simple data analytics as well as comprehensive studies [24, 42, 60] indicate that shots outside the penalty box rarely result in goals compared to shots taken inside the penalty box. Because of this, the Pitch Value is modelled in a static manner and focused on the 16-meter area. The idea is to use the Pass Probability- and Pitch Control models to analyse the dynamics of situations in combination with the Pitch Value model. The model implemented was based on the research by Link et al. [3], which shared a similar vision on the importance on the penalty box and its value compared to the rest of the field, see figure 2.4.

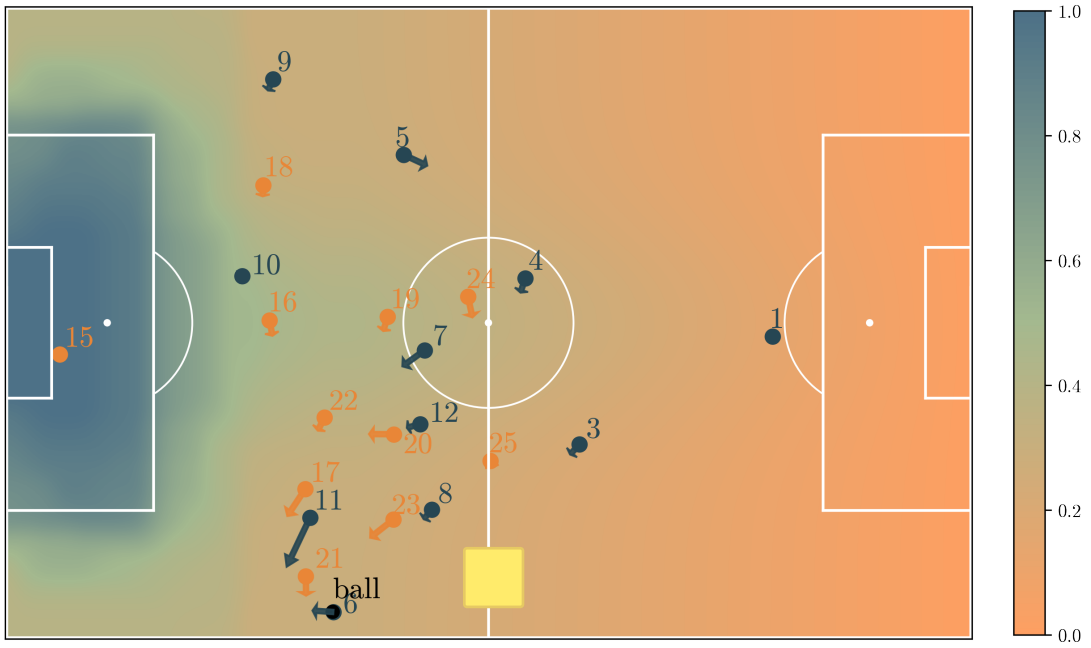


FIGURE 3.9: The developed Pitch Value model based on the model of Link et al. [3], shown in fig. 2.4. The final third is exactly the same. Additionally, a linear increase in the x-direction i added for the rest of the pitch and a Gaussian gradient i the y-direction to highlight the centre.

As Link et al. did not impose a quantification of the rest of the pitch, this was taken to be a simple model. Two concepts were imposed: i. possession increases in value towards the final third; ii. possession in the center is more valuable due to the fact that the number of possibilities is larger. This will stimulate advancement up the pitch and control in the center of the pitch. The former was modelled by a linear gradient in x-direction, the latter was modelled with a Gaussian gradient in the y-direction. This resulted in the model shown in figure 3.9.

3.3 Developed work: future scenario pass quantification

Having established a Pass Probability, Pitch Control and Pitch Value model, it is possible to conduct an in-depth analysis on passes. Three models, each discussed henceforth, were developed in order to realise a future scenario pass quantification model. The general concept of the model already passed by in pseudo-code (alg. 1). The value of a pass is quantified its possible future scenarios at the start and at the end. The quantification model works in four steps:

1. the future scenarios are determined by applying a pass option model (sec. 3.3.1);

2. for each passing option (future scenario), a movement prediction model is used to generate the positions of the players (sec. 3.3.2);

3. the value of each of the future scenarios is quantified by the situation evaluation models, i.e. the expected number of outplayed opponents and the value of controlled pitch (sec. 3.3.3);

4. the summed value of the future scenarios at the end of the pass is subtracted by the same at the start of the pass, resulting in the final score of a pass

The resulting behaviour of the future scenario models compared to the simple pass quantification models will be particularly of interest, which will be examined in the next chapter.

3.3.1 Pass options

In the extent of Fournier-Viger et al. [4], passing options were taken to be driven by coverage, pitch value and distance to the ball. A simple view sets the base: players have the objective to pass the ball to the most valuable region controlled by the team. This already outlines the dilemma on risk in passing for players. Value and control balance each other and decisions are made between them. Multiplying the models for these two variables allows for an investigation on passing options.

It must be noted that the Pitch Control model (section 3.2.2) does not account for interception probabilities of a pass. A region on the other side of the pitch might be highly controlled, but require exceptional passing skill. Based on the research of Fournier-Viger et al. [4], executability was modelled by a distance gradient. In figure 3.10, the observed passing distances are fitted for a Gaussian distribution and a Gamma distribution. The *Rooted Mean Square Error (RMSE)* was evaluated at respectively $9.6 \cdot 10^{-2}$ and $4.2 \cdot 10^{-2}$. The Gamma fit appears to corresponded significantly better to the data and was used to design the gradient.

The distance gradient was used to implement a penalty-reward function P by the use of two parameters. The amplitude of the distance gradient could be amplified or reduced, which determines the maximum penalty or reward. In addition, the distribution can be shifted in the x direction, in order to extra stimulate particular passing distances, while preserving the proportions of the distribution. Considering the hyper-parameters

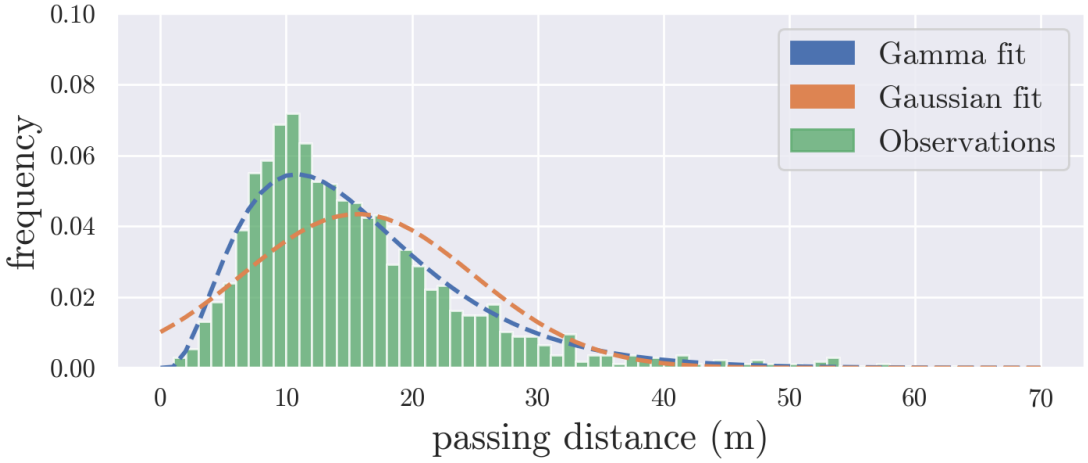


FIGURE 3.10: The obtained Gamma- and Gaussian fit to the observed passing distances. The *RMSE* evaluated respectively at $4.2 \cdot 10^{-2}$ and $9.6 \cdot 10^{-2}$. Based on this, the Gamma distribution ($\alpha = 3.06, \beta^{-1} = 4.90$) is used as distance gradient.

amplitude A_0 and shift x_0, the distance gradient was computed as follows

$$A = \frac{(2 \cdot A_0)}{\Gamma_{\alpha,\beta}(\alpha/\beta)} \tag{3.15}$$

$$P(\delta_{x,y}) = \Gamma_{\alpha,\beta}(\delta_{x,y} - x_0) \cdot A - A_0 \tag{3.16}$$

where $\Gamma_{\alpha,\beta}$ is the given Gamma distribution, α/β the corresponding mode and P the penalty-reward function for distance $(\delta_{x,y})$ to the ball. Note that amplitude A_0 will be the maximum reward at the corresponding mode and $-A_0$ will be the maximum penalty, approached by distances with a low probability. The hyper-parameters were optimised by parameter tuning, see section 4.1.1

The passing options were generated by taking the 5 maxima after applying the distance gradient. Each time after choosing the next maxima, all points within a range of that point were neglected. This range was taken to grow linear with passing distances, modelled by factor p, also optimised in the parameter tuning. The range ensures that the passing options are distinct from each other. The pseudo-code of the passing option model is given in algorithm 2.

3.3.2 Movement prediction

Subsequently, the movement of players was studied. Alguacil et al. [6] hypothesised that players try to optimise a combination of those three in their positioning. The obtained results (fig. 2.7) debunked that hypothesis, or at least the approach. The position of players at the start of the pass proved to be better a predictor than any combination of the three models. This appears to be an invaluable predictive model.

Algorithm 2: Generating passing options

Result: Possible passing options
initialise spatial information $R_{att}, R_{def}, R_{ball}$;
PC = $PitchControl(R_{att}, R_{def}, R_{ball})$;
PV = $PitchValue(R_{att})$;
DG = $DistanceGradient(R_{ball})$;
for $i \leftarrow 1$ **to** 5 **do**
  select max x^i_{max} from $PC \cdot PV + DG$;
  neglect all x with $\delta(x - x^i_{max}) < p \cdot \delta(x_{ball} - x^i_{max})$;
end

Predicting that players never move during a pass is an extremely basic idea and it should be elementary to improve it. The experiment was repeated here for current position, optimal pitch control and optimal pass probabilities against the new baseline *dead reckoning*. In addition, an extended adaptation was developed, incorporating ball orientation of the players.

3.3.2.1 Dead reckoning and ball orientation

Dead reckoning is a simple principle, given by

$$\vec{x}_t = \vec{x}_0 + \vec{v}_0 \cdot t \tag{3.17}$$

with the predicted position $\vec{x}_t$ computed from the initial position and velocity $\vec{x}_0, \vec{v}_0$ for time t. This concept will serve as baseline for the coming experimentation. It is hypothesised that this baseline can be improved by incorporating simple game dynamics, e.g. adjusting predictions towards the passing destination. As such, this thesis proposes a more sophisticated algorithm, named *ball oriented dead reckoning*, given in pseudo-code in algorithm 3. Briefly, three adjustments are implemented:

1. teams re-orient at the pass destination, even players faraway
2. movement is limited due to the bounds of the pitch
3. a strong acceleration of the intended receiver and covering defender towards the pass destination

The concept of team orientation on the ball is natural as the game revolves around the ball itself and movement within a team correlates strongly [61]. It is modelled by the orientation of the player and the direction from the player to the pass destination. If the angle between the two exceeds threshold θ_{ball}, the prediction is updated by adjusting

the player's velocity. This is done by the use of linear factor f_{ball} and maximum velocity towards the ball $\vec{v}_{to\ ball}$

$$f_{ball}(t) = f(- t_{react}) \quad \text{for } t > t_{react} \tag{3.18}$$

$$\vec{v}_{pl,t} = f_{ball}(t) \cdot \vec{v}_{to\ ball} + (1 - f_{ball}(t)) \cdot \vec{v}_{pl,0} \tag{3.19}$$

where t is the pass duration and t_{react} is the reaction time of a player. The value for slope f_0 and threshold θ_{ball} were tuned in a grid search, discussed in section 4.2.1.

The consideration of the boundaries of the pitch is relatively simple. Whenever predictions were located outside of the pitch, the predictions were adjusted. The new prediction was taken to be halfway between the boundary and the original position, maintaining the direction of the player. As players rarely position themselves at the absolute edge of the pitch, the prediction is taken to be halfway. It simulates the fact that players are located in the pitch, not on the boundary.

Lastly, if the ball is passed to an open space, e.g. a ball deep, one or two attackers will need to run towards it. In turn, one or two defenders will cover the ball as well. The decision dynamics are modelled by three thresholds: θ_{att}, θ_{def} and θ_{second}. If the distance of the closest attacker (or defender) to the ball exceeds θ_{att} (or $\theta_{att} + \theta_{def}$), the closest attacker (or defender) is moving towards the ball. The idea is that defenders will allow more space between them and the ball as they also need to consider coverage, therefore this modelled by the addition of the two threshold parameters. If $\theta_{att} + \theta_{second}$ (or $\theta_{att} + \theta_{def} + \theta_{second}$) is exceeded, the second-closest attacker (or defender) is moving towards the ball. The choice for the closest player is justified by the following statistic: in 85% of the passes, the observed receiver of the pass was also the closest to the pass destination based on dead reckoning.

The strong acceleration towards the ball is incorporated by second factor and maximum velocity towards the ball $\vec{v}_{to\ ball}$

$$g_{ball}(t) = g_0 \cdot t \quad \text{for } t > t_{react} \tag{3.20}$$

$$\vec{v}_{pl,t} = g_{ball}(t) \cdot \vec{v}_{to\ ball} + (1 - g_{ball}(t)) \cdot \vec{v}_{pl,0} \tag{3.21}$$

Again, $\theta_{att}, \theta_{def}, \theta_{second}$ and f_2 were tuned in a grid search (sec. 4.2.1).

Algorithm 3: Ball oriented dead reckoning

Result: Position prediction of player

initialise t_{pass}, $\vec{x}_{pl,0}$, $\vec{v}_{pl,0}$, $\vec{x}_{ball,0}$, $\vec{x}_{ball,t}$, $\mathbb{R}^2_{pitch}$, $team$, $\theta_{closest} = \theta_{team}$;

optimal velocity to ball $\vec{v}_{to\,ball} = \frac{(\vec{x}_{ball,t} - \vec{x}_{pl,0})}{\|\vec{x}_{ball,t} - \vec{x}_{pl,0}\|} \cdot v_{max}$;

direction difference $\hat{v}_{diff} = \frac{\vec{v}_{pl} - \vec{v}_{to\,ball}}{\|\vec{v}_{pl} - \vec{v}_{to\,ball}\|}$;

predict $\vec{x}_{pl,t} = \vec{x}_{pl,0} + \vec{v}_{pl,0} \cdot t_{pass}$;

if direction not to pass destination $\hat{v}_{diff} > \theta_{ball}$ **then**

  $\vec{v}_{pl,t} = (f_{ball}(t) \cdot \vec{v}_{to\,ball} + (1 - f_{ball}(t)) \cdot \vec{v}_{pl,0})$;

  $\vec{x}_{pl,1} = \vec{x}_{pl,0} + \vec{v}_{pl,t} \cdot t_{pass}$

if out of bounds $\vec{x}_{pl,t} \notin \mathbb{R}^2_{pitch}$ **then**

  $\vec{x}_{pl,t} = (\vec{x}_{pl,0} + \text{clipped}(\vec{x}_{pl,t}))/2$;

sorted team distances to pass $\delta_{team} = \text{sort}(\{\|\vec{x}_{pl',t} - \vec{x}_{ball}\| | pl' \in \text{team}\})$;

if player is closest $\|\vec{x}_{pl,t} - \vec{x}_{ball}\| = \delta_{team}[0]$ **then**

  if player too far $\|\vec{x}_{pl,t} - \vec{x}_{ball}\| > \theta_{closest}$ **then**

    $\vec{v}_{pl,t} = \vec{v}_{pl,0} \cdot (1 - g_{ball}) + \vec{v}_{pl-ball} \cdot g_{ball}$;

    $\vec{x}_{pl,t} = \vec{x}_{pl,0} + \vec{v}_{pl,t} \cdot t_{pass}$

else if player second-closest $\|\vec{x}_{pl,t} - \vec{x}_{ball}\| = \delta_{team}[1]$ **then**

  if player too far $\|\vec{x}_{pl,t} - \vec{x}_{ball}\| > \theta_{closest} + \theta_{second}$ **then**

    $\vec{v}_{pl,t} = \vec{v}_{pl,0} \cdot (1 - g_{bal}) + \vec{v}_{pl-ball} \cdot g_{ball}$;

    $\vec{x}_{pl,t} = \vec{x}_{pl,0} + \vec{v}_{pl,t} \cdot t_{pass}$

end

return predicted location $\vec{x}_{pl,t}$

3.3.2.2 Optimal pass probability and -pitch control

In order to relate the results presented to those of Alguacil et al. [6] (figure 2.7), a part of their algorithms were analysed as well, i.e. current position, optimal pass probability and optimal pitch control. Postulating that players optimise their recievability and control, one might expect prediction based on this to be effective. Nonetheless, this proved not to be the case in the study of Alguacil et al. and is not expected to be here.

As there is an seemingly infinite amount of possible positions for the players, one must carefully attempt to reduce the search space. In order to do so, a reachable area was computed based on eqs. 3.2 – 3.7 and the passing time. For each player, 200 random points were sampled of this area. For each point, the change in pass probability and pitch control compared to the current position were computed. The position with the most optimal change was accepted for each of the two.

3.3.3 Situation evaluation

Finally, an evaluative model is needed in order to conduct pass quantification through future scenarios. Two separate models were combined in order to assess a situation, in

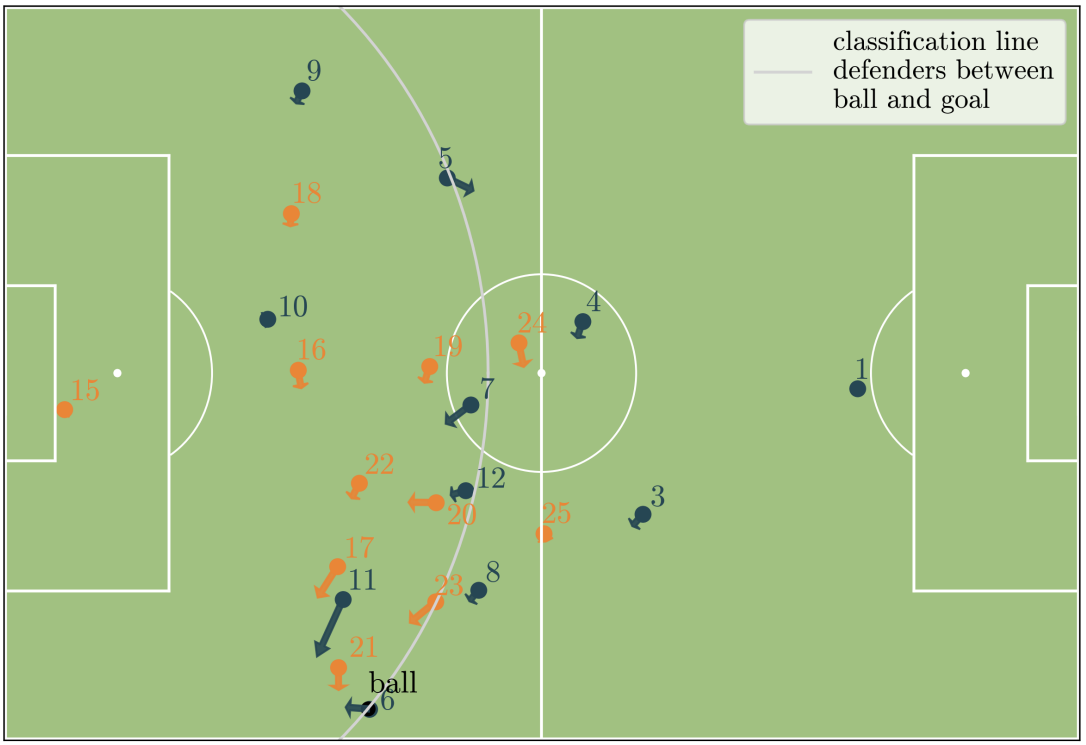


FIGURE 3.11: Visualisation of defenders in between the goal and the ball, in this case 9 players. In contrast to other models known, the classification is done by a circle from the middle of the goal and the ball. Only players 24 and 25 are not within this circle.

line with the research of Rein et al. [7]. One is the amount of opponents between the ball and the goal, the second is an evaluation of the value of the controlled pitch.

3.3.3.1 Outplayed opponents

As mentioned in the previous chapter, this concept (also known as packing) has grown popular recently in football analytics. The model is extremely simple and the only discussing point is the method to decide which opponents are between the ball and the goal. Scott [33] and Rein et al. [7] consider all opponents closer to the goal based solely on the x-direction. In our opinion, ignoring the y-direction is not justified. This can be illustrated by the following example: if the attacker is at the edge of the penalty box in the center and a defender is on the same horizontal line, but far out on the left flank. The defender will never be on time to block a shot. Therefore, it was chosen to draw a circle from the middle of the goal. It represent a rough estimate of the time needed to sprint towards the goal. The distance from the ball to the middle of the goal was taken as the radius. The number of opponents inside the circle are considered to be capable to block the ball, if the player in possession dribbles towards the goal. The model is shown in figure 3.11.

3.3.3.2 Change in controlled pitch

Having a model for both pitch control and pitch value, a simple combination of the two provides insight in the value of controlled regions. Considering that Pitch Control is given by probabilities and Pitch Value is based on goal-scoring probabilities, the models are combined through multiplication.

It remains the question what values are considered of PC x PV, as it is a grid of values across the field. One might take the maximum value or the average value of the whole grid. For this study, it was chosen to primarily consider the two regions that are highest in control and value. The three next highest value are also taken into account. This was done by applying the weights $\frac{1}{4}, \frac{1}{4}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}$ to the 5 maxima (highest first). Via the weighted summation, a quantified value of controlled pitch is found.

Chapter 4

Experiments and Results

The models have been tested against the data of two games as published by Metrica Sports [8]. First, the parameter tuning of the pass option model and ball oriented dead reckoning are discussed. Both were done via a simple grid search. Subsequently, the obtained behaviour in identifying passing options and achieved precision in predicting the movement of players are shown. Based on these results, the substantiality of the generated future scenarios can be assessed and justified.

Next, the results of pass quantification models are discussed. Due to the limited data set and a computational heavy algorithm, it was not possible to link the models to observed wins as done by others [7, 62]. Any conclusion on win-probabilities based on two games is not statistically justified. Therefore, a different statistical approach was needed to relate the model to the dynamics of the game. In the philosophy of the pitch value model (considering primarily shots inside the penalty area as valuable [24, 42, 60]), the time of the ball in the opponent's penalty area was taken as descriptive statistic. The relation of the pass evaluations to this metric was studied per 5 minutes of a game in order to generate sufficient data points. The behaviour of the standard pass quantification serves as a stepping stone to the future scenario pass quantification.

4.1 Pass options

4.1.1 Parameter tuning

The hyper-parameters amplitude A_0, shift x_0 and range factor p (sec. 3.3.1) were optimised by the means of a grid search. The resulting pass options were tested on two properties: the predictive value and the resilience to the real data. The former is quantified by the frequency that the actual pass was identified, i.e. one of the pass options is

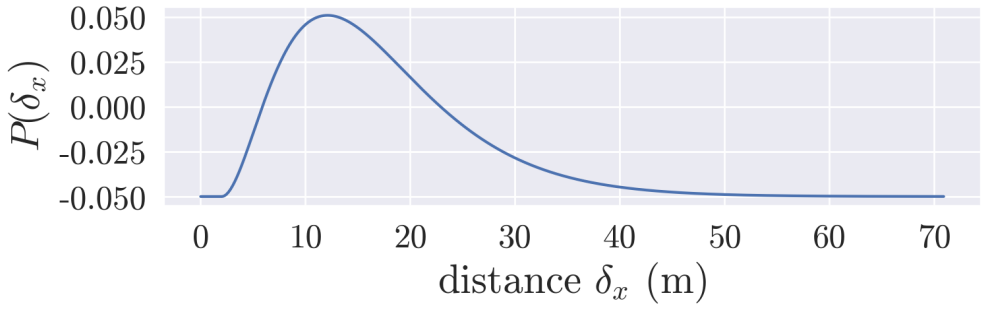


FIGURE 4.1: The optimised penalty-reward function P for distance to the ball δ_x, stimulating passes of 6 - 23 meters.

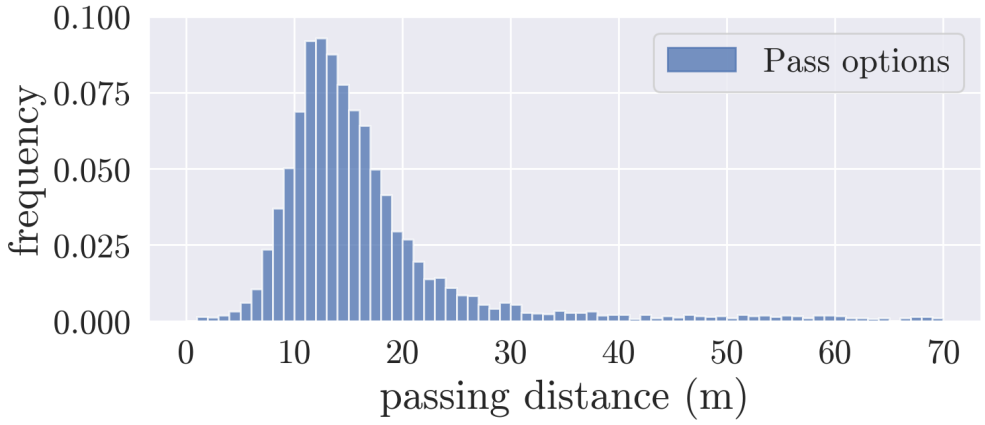


FIGURE 4.2: The optimised distribution in pass option distances. Despite a sharper peak around 10 m and longer right-tail, it resembles the observed passing distances (fig 3.10).

close to the actual pass. The range was taken to grow linear with the pass distance by a factor of 0.2. The latter is quantified by the RMSE between the observed distribution of actual passes (see fig. 3.10) and the resulting distribution in pass options, binned per meter. The results are shown included in Appendix B (fig. B.1). The values $A_0 = 0.05$, $x_0 = 2.0\ m$ and $p = 0.25$ were found maximise the predictive value to 30% and minimise the RMSE to 0.012. The value of $A_0 = 0.05$ appears reasonable, given that Pitch Control multiplied by Pitch Value scales in general from 0.1 to 0.3. The resulting penalty-reward function is plotted in figure 4.1.

4.1.2 Resulting behaviour

The resulting distribution in pass option distances is depicted in figure 4.2. Comparing it to figure 3.10, we recognise a resembling pattern. This substantiates the pass options model. The shorter passes are favoured in the pass option model, which is a positive effect for the model; the movement prediction model is more accurate on shorter periods of time is more accurate.

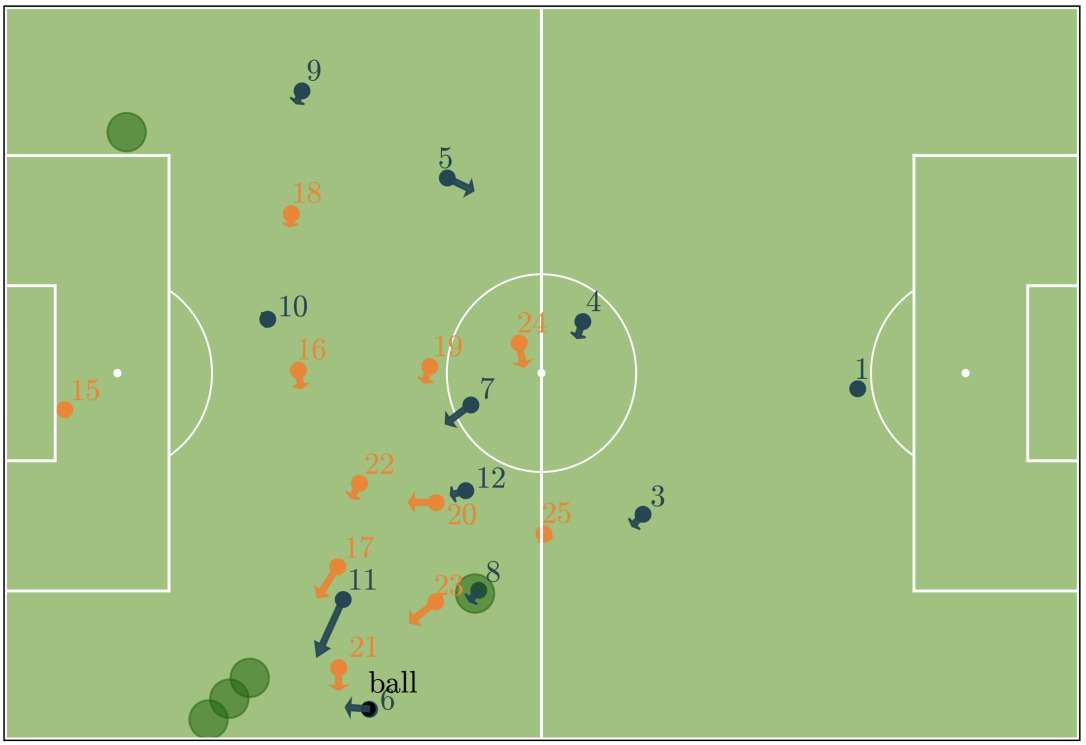
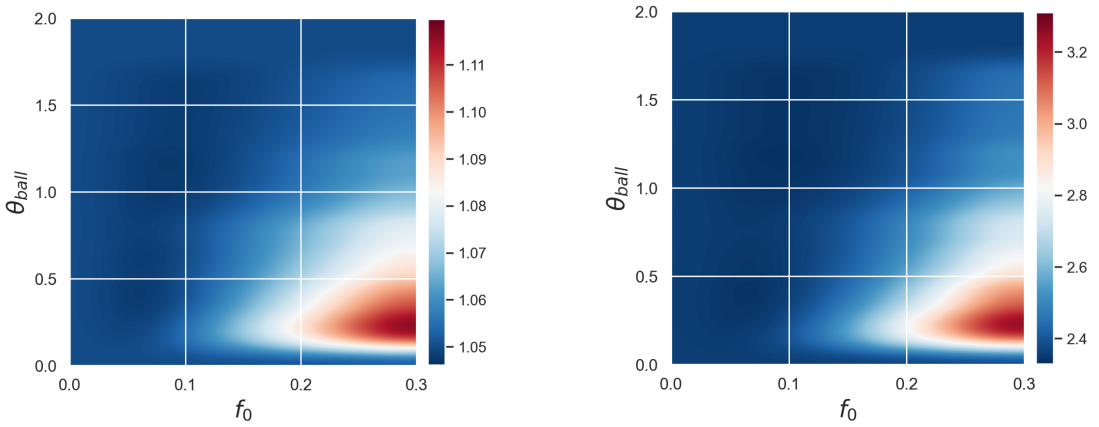


FIGURE 4.3: The resulting pass option model applied to the same scenario as figure 1.4. Remarkably, the algorithm identified passing options very similar to those recognised by eye, i.e. to players 8, 9 and 11.

Note that the corresponding predictive value is approximately 30%, i.e. the model does not identify the observed pass for 70% of the studied passes. Naturally, some situations result in suboptimal behaviour as players do not always make good decisions, but the model appears unable to fully capture the game dynamics. This is further investigated in the Discussion (ch. 5). The model is still considered usable for future scenario analysis, which focuses on optimal continuations of the game.

Lastly, to illustrate the functioning of the model, the generated pass options are plotted in figure 4.3 for the scenario of figure 1.4. Interestingly, the generated passing options of the model match some of the hand indicated passing options. This result cannot be generalised to all scenarios, but it illustrates the functioning of the model. The ball deep to player 11, which was the observed pass, is proposed in three formats. The range of neglecting points generated by factor $p = 0.25$ did not anticipate this, which could be interpreted as this pass option being the most valuable option. The generation of distinct pass options is further discussed in section 5.2.



(A) The mean prediction.

(B) Variance in the predictions.

FIGURE 4.4: Grid search on the reorientation of the whole team due to the pass. This is modelled by the threshold in direction difference θ_{ball} and a factor f_0 of adapting to a velocity towards the pass destination. The values $f_0 = 0.05$, $\theta_{ball} = 0.4$ rad were found to minimise the mean (A) and variance (B).

4.2 Movement prediction

4.2.1 Parameter tuning ball oriented dead reckoning

For the proposed ball oriented dead reckoning (pseudo-code 3), parameter tuning was done by two grid searches. First, the hyper-parameters θ_{ball}, f_0 on ball orientation of the whole team were tuned, see figure 4.4. The found values were $f_0 = 0.05$ and $\theta_{ball} = 0.4$ rad, which minimised the mean (1.046 m) and variance (2.328 m) of the predictions. Second, the hyper-parameters g_0, θ_{att}, θ_{def}, θ_{second} on runs towards passes into empty spaces were tuned. Due to the extent of the figures, the results are included in appendix B (figs. B.2 & B.3). The found values were $g_0 = 0.1$, $\theta_{att} = 3$ m, $\theta_{def} = 5$ m and $\theta_{second} = 3$ m were found to minimise variance (2.040 m). The resulting mean (1.009 m) is close to the minimised mean (1.006 m).

Interestingly, the resulting variance varies on a significantly larger scale than the resulting mean. It also stands out that the values for f_0 and g_0 implicate weak links between modelled ball orientation and the observed movement. The high variance and these weak links will both become more understandable in the next section. The impact of the extension to dead reckoning turns out to still be visible.

4.2.2 Prediction precision

In order to draw a proper comparison with other field-related results, the movement predictions are depicted exactly as done by Alguacil et al. [6] (fig. 2.7). Unfortunately,

it was not possible to sort the results per position as the data [8] was anonymised. The proposed models are compared to their models in the ordinal plot 4.5. The optimal pitch control, optimal pass probability and current position of Alguacil et al., see figure 4.5a, perform similar to the publication, substantiating the results found.

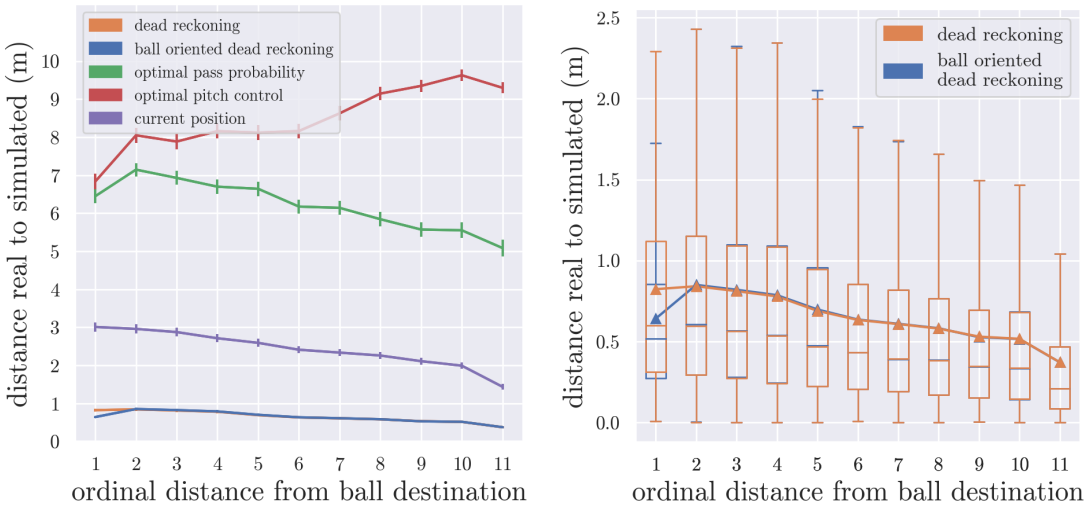
The predictions are clustered per ordinal distance, i.e. seperated into the receiver of the ball, the second-closest to the ball, etc. Per i^{th}-closest player, the distances of the predictions to the observations are analysed. The means are plotted alongside 95%-confidence intervals or box plots; showing both average performance as well as the performance range. In line with Alguacil et al., only passes of less than 2 seconds (86% of the data set) and the attacking team were considered.

Remarkably, dead reckoning proves to be an extremely robust predictor of the movement of players and no vast improvements were found. Comparing the two figures, it becomes apparent that the proposed models are produce improved predictions. The average distance from the prediction to the observation decreases over all positions by at least 2 m. Recalling the AI-based research of Le et al. [5] of figure 2.6, the model appears to achieve innovative precision.

Figure 4.5b might raise the expectation that the developed ball oriented dead reckoning only achieves improvements in predicting the position of the receiver. Although the prediction of the receiver might be viewed as most important, there are more improvements. The difference is better highlighted in figure 4.6, where the outliers are plotted. The outliers decrease both in value and amount. The same applies to extreme mispredictions over 10 meters. The fact that there are still exceptions leading to disastrous mispredictions forms a critical point. The model is still not fully able to capture the game dynamics.

4.3 Standard pass quantification

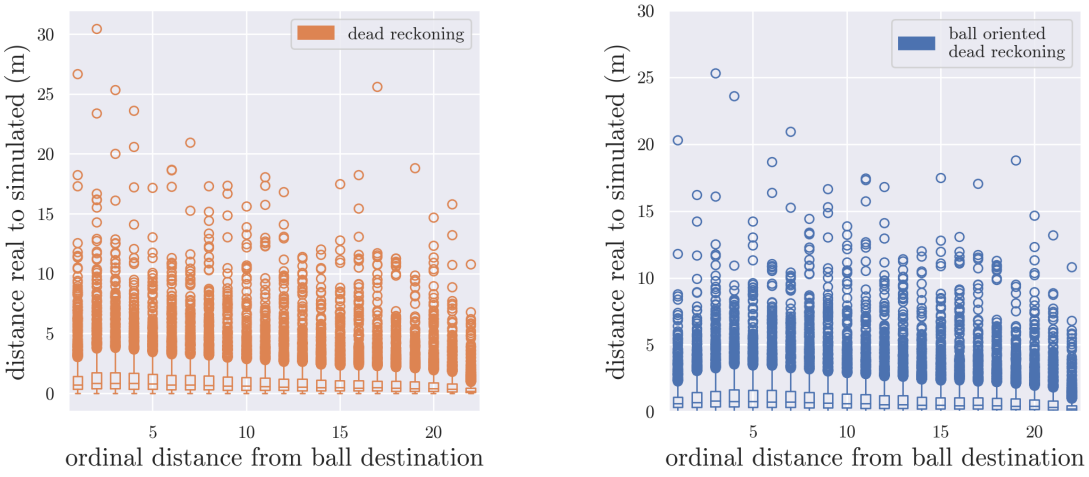
In order to understand observations on the future scenario pass quantification model, the standard pass quantification models were studied. The average pass value was related to three performance indicators per 5 minutes of a game, i.e. *shots*, *shots on target* and *the time of the ball in the opponents penalty box*. The results for the outplayed opponents pass quantification are shown in figure 4.7 and for the change in controlled pitch pass quantification in figure 4.8. The former appears to have no correlation to any of the metrics, i.e. Pearson product-moment correlation coefficient $0 \leq r \leq 0.01$. This is somewhat contrasting with the work of Rein et al. [7] (see fig. 2.5). The latter only appears to have a weak correlation with the number of shots and shots on target, i.e.



(A) Published and proposed models

(B) Close-up proposed models

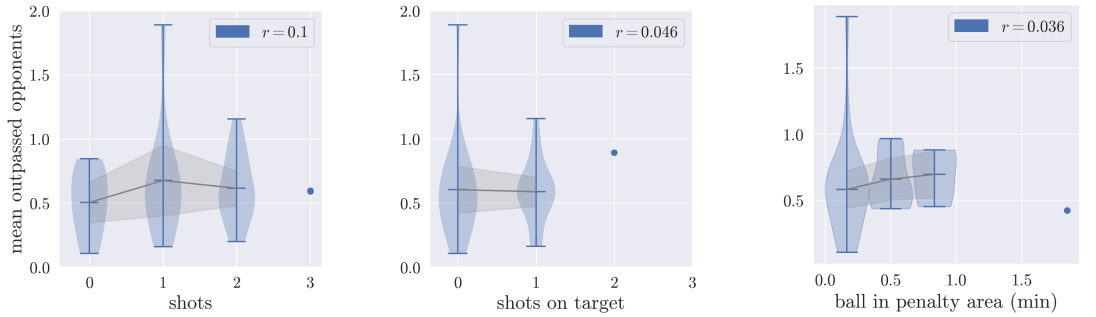
FIGURE 4.5: The predictions of the position of attacking players after passes within 2 seconds. The results are ordered per i^{th} closest player to the pass destination. The left plot (A) shows the means and 95%-confidence intervals of the proposed models, which outperform the published models. The right plot (B) shows a close-up of the proposed models in the form of box plots and mean values (triangles). The box plots contain 93.7% of the data.



(A) Proposed baseline

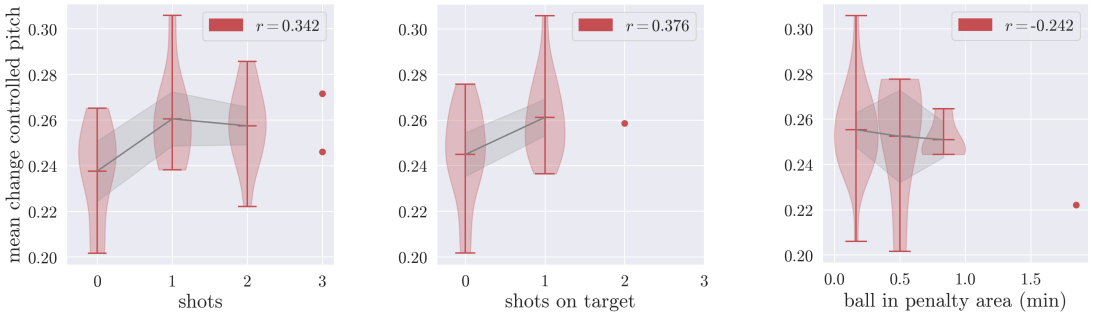
(B) Developed algorithm

FIGURE 4.6: The predictions of the position of attacking & defending players after all passes, including the identified outliers. It can be seen that the extreme values are reduced by the developed ball oriented dead reckoning (B). Compared to dead reckoning (A), the mean outlier value is reduced from 4.80 m to 4.47 m and the amount of outliers is reduced from 6.40% to 6.33%. The amount of extreme mispredictions (>10 meters) is reduced from 0.34% to 0.23%.



(A) Number of shots (B) Shots on target (C) Time ball in penalty area

FIGURE 4.7: The relation between the mean number of outplayed opponents in a pass and three performance indicators. This was studied per 5 minutes of the game. The Pearson correlation coefficient r appears to indicate little correlation between the pass quantification and shots (A), shots on target (B) or ball-time in the penalty area (C). This is somewhat contrasting with the results of Rein et al. [7] (fig. 2.5).



(A) Number of shots (B) Shots on target (C) Time ball in penalty area

FIGURE 4.8: The relation between change in controlled pitch and three performance indicators. This was studied per 5 minutes of the two games of the data set. In contrast to the number of outplayed opponents (fig. 4.7), the Pearson correlation coefficient r is significantly larger and indicates a small correlation with the amount of shots (A) and shots on target (B). Surprisingly it also indicates small negative correlation with ball-time in the penalty area (C).

$r = 0.342$ and respectively $r = 0.376$, which is more in line with the work of Rein et al. [7]. Surprisingly, $r = -0.242$ for the ball-time in the penalty area. This indicates either that the pass quantification is inconsistent or the ball-time in the penalty area is a weak performance indicator. Although shots and shots on target are known not to be strong indicators of effective play, this result indicated that the situation evaluation methods leave room for improvement via future scenario analysis.

4.4 Pass quantification through future scenarios

Finally, let us turn to the identical analysis for the future scenario quantification model. The pass quantification is repeated for each of the situation evaluation models. Figure 4.9 shows the results for the number of outplayed opponents in future scenarios and figure 4.10 shows the results for change in controlled pitch in future scenarios. Evaluating

(A) Number of shots (B) Shots on target (C) Time ball in penalty area

FIGURE 4.9: The relation between the mean number of outplayed opponents in future scenarios with three game-related metrics. This was studied per 5 minutes of the two games of the data set. In comparison to figure 4.7, the Pearson correlation coefficient r has increase. On the other hand, the it is still smaller than in figure 4.8.

(A) Number of shots (B) Shots on target (C) Time ball in penalty area

FIGURE 4.10: The relation between the mean controlled pitch value in future scenarios with three game-related metrics. This was studied per 5 minutes of the two games of the data set. In comparison to figures 4.7, 4.8 and 4.9, the Pearson correlation coefficient r is the smallest. For the left figure, r is even negative.

future scenarios instead of the pass itself strengthens the correlation between the number of outplayed opponents and the shots on target ($r = 0.202$). However, there is no correlation with the number of shots ($r = 0.016$) or the ball-time in the penalty area ($r = -0.085$). For the change in controlled pitch, the correlation increase with shots to $r = 0.378$ and with ball-time in penalty area to (0.048). The correlation with the shots on target appears to have decreased ($r = 0.234$). For both evaluative methods, no consistent change is found by applying an analysis on future scenarios. As such, the developed future scenario analysis does not produce new insights in the number of shots, the number of shots on target or the time of the ball in the penalty area. Further remarks on these results are discussed in the following chapter.

Chapter 5

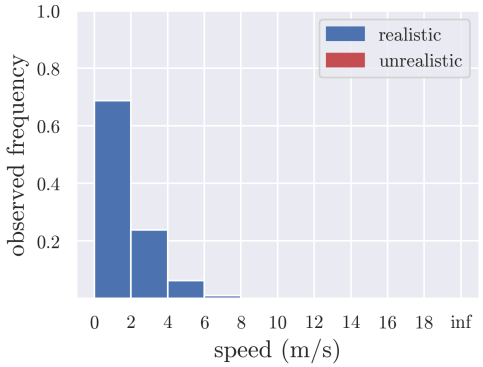
Discussion

The obtained results for different models are surprising. Where the results on movement prediction and passing options have been positive, the effect of the future scenario analysis was less so. Both will be discussed below. The models are placed in a broader context in order to better understand what we have seen.

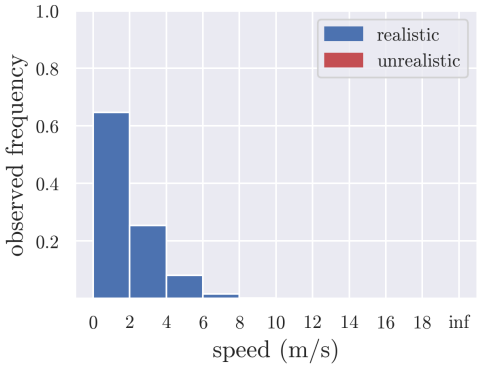
5.1 Data

First, let us turn to the data, of which the effects of the Kalman filter were yet to be further discussed. Due to the fact that the data is generated by a third-party, it is important to understand the data quality. In figure 3.4, the effects of the Kalman filter were visible. The strong reduction in unrealistic accelerations was in line with the expectations. Smoothing the tracking data of a player, by using the expected position based on the previous position and velocity, resulted in a more realistic acceleration pattern. However, the smoothing resulted in only a small decrease in unrealistic observed speeds, a rather counterintuitive result. This is can be explained by the small fraction (0.036%) it makes up. At these data points, it is possible that the observational is too large, which also limits the reduction in unrealistic accelerations.

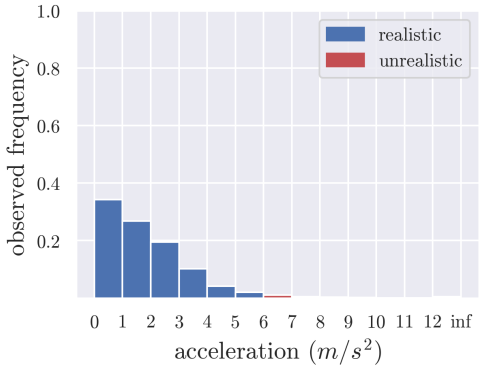
Once the results of the Kalman filter are plotted on a linear scale, see figure 5.1, the amount of unrealistic data is better understood. Generally, the open data-set of Metrica sports [8] appears to be consistent on speeds; it can be justified to ignore the small fraction of unrealistic speeds. The fraction of unrealistic accelerations makes up 2.9% prior to the Kalman filter, which is harder to justify to ignore. After smoothing, this fraction only consist of 0.050% of the data. However, the frequency of low accelerations ($1 - 2\ m/s^2$) has almost tripled. As the distribution on observed speeds is nearly



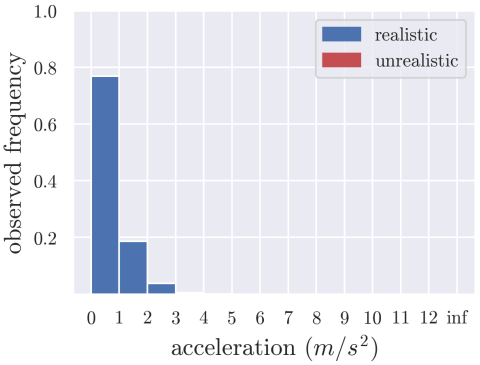
(A) Observed speeds in raw data



(B) Observed speeds after Kalman filter



(C) Observed accelerations in raw data



(D) Observed accelerations after Kalman filter

FIGURE 5.1: The results of figures 3.3 and 3.4, depicted on a linear scale instead of a log scale. It highlights the general absence of the unrealistic data.

unchanged, it can be argued that the increase in low accelerations does not indicate over-smoothing. In conclusion, serious errors are visible in the data and the Kalman filter was able to anticipate a part of it. Although the viability of the data is increased, more investigation and understanding of the data would ideally be available to properly conduct the study. Possibly, this can be provided by Ajax in the future.

5.2 Pass options model

The first model developed for future scenario analysis regarded pass options. The results of the parameter tuning (app. B.1) and figures 4.2 & 4.3 showed the behaviour of the model. Four aspects of the model are further explored here. The model is first related to the literature discussed in section 2.4.2 in order to understand its performance. Subsequently, the trade-off between generating distinct pass options and not skipping pass options, due to neglecting points in choosing the next pass option, is zoomed into. Lastly, the underlying models, Pitch Value and the distance gradient, are reviewed. As the Pitch Control model is reproduced work, it is left undiscussed.

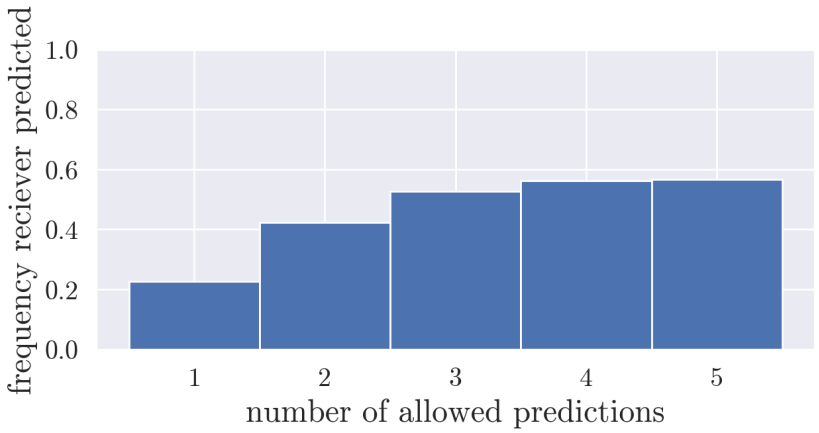


FIGURE 5.2: The frequency of correct predictions on the receiver by the pass option model. The first prediction is correct for 22.6% of the passes. By allowing two guesses, this number becomes 42.2%. Both are 10% below the performance of Fournier-Viger et al. [4].

5.2.1 Predicting the receiver

In order to relate this model to the research of Fournier-Viger et al. [4] (sec. 2.4.2), the proposed model needs to be translated from locations to players. This was done by selecting the player closest to the pass option. Figure 5.2 shows the correctness of the prediction for a varying amount of guesses. The primary (or optimal) pass choice according to the model corresponds 22.6% to the actual pass choice. This is 42.2% if the second best pass choice is included. Both are approximately 10% lower than the accuracies achieved by Fournier-Viger et al. [4]. The hypothesis that their model could be improved by considering Pitch Control and Pitch Value is debunked. However, it must be noted that this model is able to produce the locations of pass options, which is a more direct approach than only considering what player can be passed to.

5.2.2 Distinct pass options

As visible in figure 4.3, the model currently is not always able to generate distinct pass options. The driving force in the model is the amount of points being neglected in choosing the next pass options, modelled by parameter p (alg. 2). By increasing p the distance between generated pass options is forced to be bigger; it leads to more distinct pass options. However, when teammates positioned close to each other, the model might neglect particular pass options. The approach of this models focuses on spatial aspects, as a pass in front of a player is different from a pass behind the same player. Taking the superior results of Fournier-Viger et al. [4] into account, the model is possibly improved by incorporating a focus on receiver options.

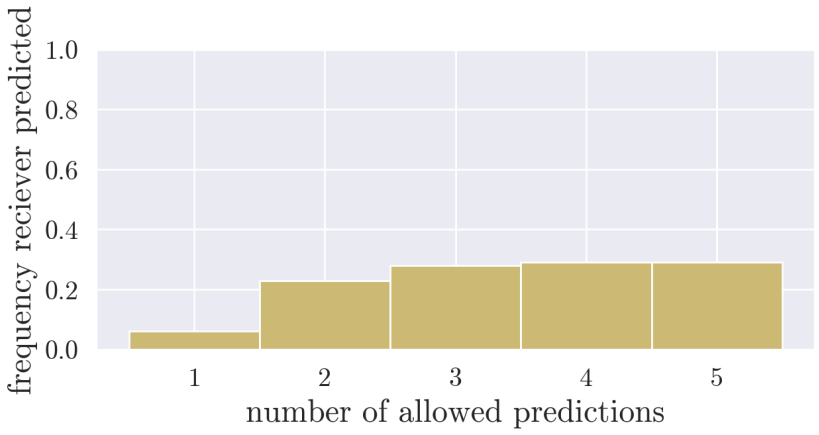


FIGURE 5.3: The frequency of correct predictions on the receiver by the pass option model *without Pitch Value*. Compared to figure 5.2, it can be seen that the performance drastically decreases.

5.2.3 Pitch Value model

The postulated focus on the penalty box for the modelled Pitch Value can be seen as a tactical vision. One team might decide that the optimal playing style is not focused on getting the ball to the penalty box. This argument can be countered by figure 5.3, which shows the receiver predictions of the model without Pitch Value. The success frequency is less than half the success frequency of the original model.

Still, the model cannot be claimed to be universal and might not be applicable to all teams and games. For example, a team might decide to focus maintaining possession during a game in order to prevent the other team from scoring. Then, passes towards the penalty area are actually less valuable, due to the increased risk of interception. The bias of the model should be taken into account when applying it. This bias could be anticipated by considering multiple value distributions on the pitch and making them dependent on team tactics.

5.2.4 Distance gradient

The applied distance gradient is based on a fit to the same data as the eventual pass choice model is tested. This produces a strong bias in the model towards the data used. There is little research published on the distribution of passing distances. It is hard to relate the found gamma distribution to known publications. The concept of the gamma distribution is not unknown to the field of football analytics. Mendes et al. [63] found the distribution in inter-touch times and Narizuka et al. [64] found it to be the degree distribution of passing networks. Nonetheless, it is hard to underpin the modelling choice properly and an extensive data analysis should be done to substantiate the distribution.

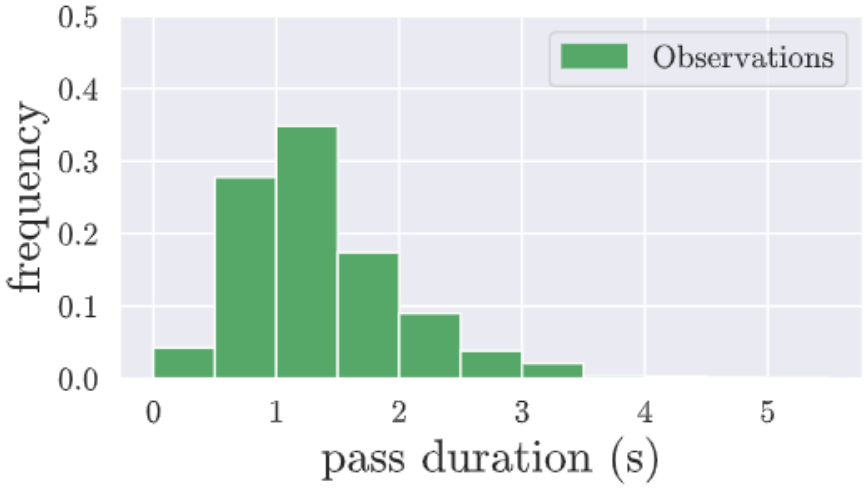


FIGURE 5.4: The observed pass durations.

In conclusion, it becomes apparent that the proposed pass option model is unable to fully capture game dynamics. In spirit of understanding these dynamics, improving this pass option model is fundamental to the future scenario analysis. Possibly, the impact of the application of future scenario to the pass quantification models changes for an improved pass option model.

5.3 Movement prediction

5.3.1 Dead reckoning

Looking at the results of figure 4.5, the observations are both intuitive as well as counterintuitive. As hypothesised, the baseline for movement prediction is significantly improved compared to the research of Alguacil et al. [6]. Incorporating the velocity of a player in addition to its location at the start of the pass has decreased the average prediction by more than 3 times compared to only the location. As such, it has been proven that dead reckoning serves as a proper baseline for current research. In addition, the decrease in prediction error for players further from the ball is intuitive. Players further away are less likely to participate in the current play and therefore less likely to accelerate drastically.

Moreover, dead reckoning proves to be an extremely robust predictor on movement in the majority of the passes. It might have produced predictions more accurate than anticipated, as one can imagine that players do not move in straight lines. The low intelligence of the model creates some expectation of it failing in many cases as well. Although this was highlighted by the plotted outliers (fig. 4.6a), the majority of the

passes were accurately predicted. This can be explained by contemplating on the range of reaction times of players, i.e. 0.20 - 0.35 s [58], and figures 5.1 and 5.4. The majority of the passes (85%) last less than 2 seconds, leaving (on average) less than 1.75 seconds for the players to accelerate. As the players rarely accelerate with more than 2 m/s^2, their velocities will often have changed no more than 2 $m/s^2 \cdot 1.75\ s = 3.5\ m/s$. Comparing the second order kinematic equation $x(t) = x(0) + v \cdot t + \frac{a}{2} \cdot t^2$ to the first order $x(t) = x(0) + v \cdot t$, we expect to observe maximum deviations of order $\frac{2m/s^2}{2} \cdot 1.75^2 s^2 \approx 3\ m$, which close to the upper neighbourhood of the observed variations of 2.5 m in figure 4.5b.

5.3.2 Effects of ball orientation

The developed ball oriented dead reckoning did produce more accurate predictions than the dead reckoning baseline, as hypothesised. In particular the positional prediction of the receiver was improved. Figure 4.6b, highlighted the improvements on the worst-case predictions of dead reckoning. Unfortunately, the general improvement was moderate. The found value for $f_0 = 0.05$ from the parameter tuning indicates a very weak link between the direction towards the ball and the movement of the players across the team. This is possibly caused by the modelling method, as all players are shifted in velocity in the same way. The model does not account for a couple of imaginably important factors, for example: distance to the ball, as movement will primarily change for players close to the ball; current speed of the player, as players will change less drastically in full sprint; or, the coverage of the field, as players will avoid standing close to players of the own team.

The modelled sprints towards passes into empty spaces resulted in a stronger effect. The found value for $g_0 = 0.1$ appeared to indicated another moderate effect of the modelled ball orientation on the prediction precision. However, figure 4.5b indicates a significant improvement in predicting the movement of the receiver. Arguably, the movement of this particular player is most important to understand; it is vital to understand the receivers position first. This highlights the significant value of the modelled ball orientation.

5.3.3 Prediction precision

In the parameter tuning of the ball oriented dead reckoning (app. B.2) it became visible that the variance among the predictions scales larger than the mean. The driving force is made visible in figure 5.5, which shows the performance in movement prediction for long passes ($> 40\ m$). The bulk of the passes are short in both duration and distance (figs. 3.10 & 5.4), for which the dead reckoning based algorithms predict the movement well. However, for the small share of long passes (2.81% of the data), the accuracy is

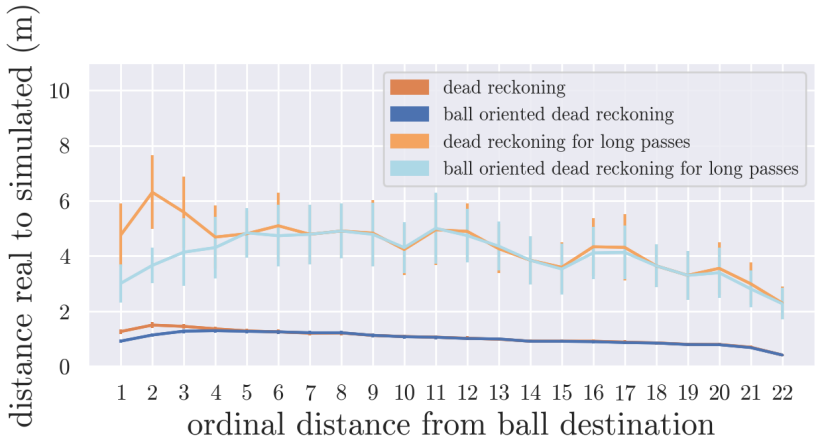


FIGURE 5.5: The positional predictions of the proposed models for all passes and long passes; all players are included. The mean predictions as well as the 95%-confidence intervals are significantly higher for longer passes (> 40 m), which is an explanation of the high variance seen in the parameter tuning (app. B.2).

significantly lower. The average error is 3 - 4 times higher; the corresponding confidence intervals scale around 2 - 3 m. To provide further support, the mean prediction error and the max prediction error were found to have the Pearson correlation coefficient $r = 0.672$ and respectively $r = 0.665$ with the passing distance. The prediction errors were not found to behave differently in the x or y direction. This illustrates that the direction of the pass is not very influential, but the distance of the pass is.

In order to fully understand the precision in movement prediction, it must be compared to the accuracy of the data. Unfortunately, little is to be found on the accuracy in the tracking data of Metrica Sports [8]. A recent study [65] elaborately studied the accuracy of Tracab[1], a company that provides similar data. The accuracy of their technology was found to be bound by a maximum error of 0.2 m.

Although this accuracy could be worse for the data used, this already highlights the current limit in achievable prediction accuracies. The proposed models already approach this limit. For the majority of players, their velocity will not change in 2 seconds, especially those far away from play. The largest improvements can be made by identifying the difference in the situations of outliers and change the approach for those scenarios.

5.4 Standard pass quantification

Before turning to the final developed model on future scenarios, let us briefly examine the results of the standard pass quantification models, i.e. outplayed opponents (fig. 4.7) and controlled pitch quantification (fig. 4.8). The latter was newly developed and

[1] https://tracab.com

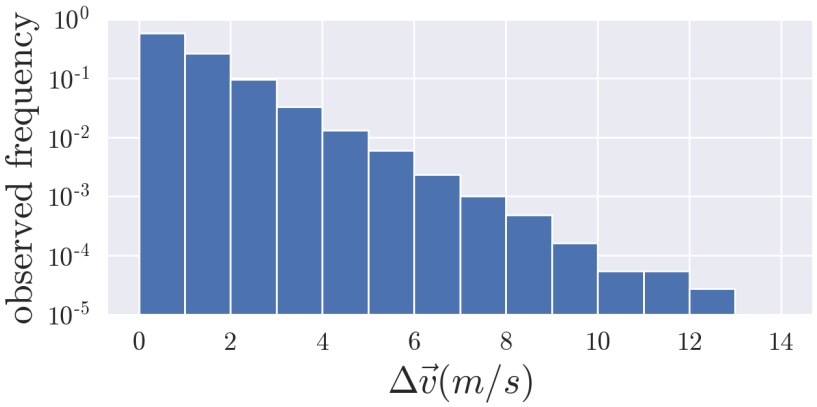


FIGURE 5.6: Observed changes in velocity during a pass. The change is 58.2% between 0 - 1 m/s, 26.6% between 1 - 2 m/s and 9.53% between 2 - 3 m/s.

behaves as expected. The quantification model only weakly correlates to the descriptive metrics. The former is known to be used extensively across football analytics to quantify passes. However, it appears to have little correlation to the performance indicators studied. This implicates one of the following three: good passes have little correlation to good play (unlikely); the pass quantification model is inaccurate; or, the performance indicators are inaccurate. It is known that the number of shots and shots on target are imperfect descriptors of a football game. The last metric, the ball-time in the penalty area, is devised for this study. Therefore, the analysis requires the use of additional metrics. This will probably require a more extensive data set, such as in the research of Power et al. [29], Link et al. [3] or Chawla et al. [30]. However, the pass quantification methods both appear insufficient to completely capture passing values, as hypothesised.

A final remark can be made on the time-windows that were used. In order to obtain enough data points, the models were analysed per 5 minutes of the games. This might not incorporate full match dynamics. It is possible that the playing style observed in 5 minutes of a game is hard to be described by statistical metrics, such as shots or shots on target, especially when these events are relatively rare within a game.

5.5 Pass quantification through future scenarios

In contrast with the hypothesis, the application of future scenarios did not provide additional insight in pass quantification. Comparing figures 4.9 and 4.10 to figures 4.7 and 4.8, the correlation to the descriptive metrics did not significantly improve. For future evaluation of change in controlled pitch, the Pearson correlation coefficient only increased for the number of shots from 0.342 to 0.378. For the future evaluation of number of outpassed opponents, the correlation only increased for the number of shots

from 0.046 to 0.202. The overall result appears to indicate that the future scenario model is able to be of addition to the pass quantification models.

5.5.1 Dead reckoning for hypothetical situations

One potential reason is the use of dead reckoning in order generate future scenarios. It can be viewed as conflicting, even the ball oriented model. Simply put, using the model suggests that passing the ball to different destinations has little effect on the movement of players within 2 seconds. Whether this is true is hard to study, as it requires an extensive data set of comparable situations. It can be reasoned that players anticipate their movement on the posture of the player in possession. Thus, dead reckoning might be a good predictor, because most players are able to anticipate the pass destination. Modelling alternative passes via dead reckoning possibly does not capture the dynamics of player movement.

A counter-argument can be made based on figure 5.6, this is covered to some extent. It can be seen that the change in velocity during a pass is not often, i.e. 5.67%, more than 3 m/s. This indicates that players in most scenarios do not strongly change their velocity. It is interesting to note that the duration and length of passes is highly variable through different sports. Studying the usability of dead reckoning through different sports will provide a better understanding of its application in football.

In particular, for the future evaluation of the change in controlled pitch, the modelled pitch control can be unrealistic, as (ball reckoning) dead reckoning does not model the drastic changes in velocity by players. In figure 5.6, we fortunately see that this effect is minimal, as strong changes in velocity are rare.

5.5.2 Computation method on future scenario evaluations

Due to limited time, this study was not able to explore different methods of using the values of future scenarios. The current models simply selects the 5 (identified as) best scenarios and evaluates those scenarios. Via a weighted sum, the value of the current scenarios is found. The final quantification of a pass is computed by subtracting the future scenarios evaluation of the start of the pass from the future scenarios evaluation of the end of the pass. Naturally, this method is not trivial and many variants are possible. The influence of different computation scheme should be further studied.

Chapter 6

Conclusions and Future Work

This thesis started off with the research question: *How can the long-term impact of a single football pass be quantified? In particular, how do we evaluate all future scenarios after the pass and quantify their impact.* Inspired by the approach of chess-computers, the concept was to evaluate passes via the corresponding future possibilities. In order to maintain a tractable search space, only two types of action were considered: running and passing. In contrast to the hypothesis, the developed model was not able to produce additional insights in game dynamics to the standard pass quantification (secs. 4.3 & 4.4). However, models used for the future scenario analysis showed interesting results. The established pass option model was showed potential and was able to approach the performance of published models (secs. 4.1 & 5.2). The established model on movement prediction was able to outperform field standards (sec. 4.2).

6.1 Reproduced work

At the basis of the future scenario analysis are three reproduced models (sec. 3.2): Pass Probability, Pitch Control and Pitch Value. Some remarks can be made on the reproduced work. Despite their proven scientific value by preceding research, some opportunities for improvement can be addressed.

6.1.1 Pass Probability

As stated by Alguacil et al. [6] themselves, a proper study on the ball-trajectory in passing is needed. Considering published work on physical forces on the ball [66–68] and accuracy in kicking-speeds [69], a more elaborate passing model should be developed. If this includes air-passed, the Pass Probability model might become more valuable than

the Pitch Control model. With more realistic trajectory simulation, the Pass Probability model will allow for a more extensive use.

The modelled approximation of player movement towards passing locations would ideally be fundamentally different. The obstacle is computational power and time. Currently, extending this model to a realistic simulation of a player's reaction on a pass requires too much computation. It would the implication that the algorithm requires a drastic increase in time. Future research might be able to formulate an answer to this dilemma.

6.1.2 Pitch Control

Similar to the Pass Probability model, the approximation on ball trajectory is oversimplified in this study. The use of an average passing speed does not result in realistic modelling. This is not an insurmountable problem, for the Pitch Control model does not aim to model passing dynamics. Rather, it aims to provide insight on the control of teams on the pitch, which is more strongly related to the time of arrival of players. This also can be improved. Looking at the resulting figures 3.8, the resulting values appear to be realistic. Therefore, the fundamental approximation can be improved, but the resulting model is already able to capture the desired dynamics. It is debatable how realistic fringes are captured by the model, as it takes a considerable amount of time to get the ball there.

6.1.3 Pitch Value

The Pitch Value model was based on existing research [3, 24, 42, 60] indicating that the value of penalty area is guiding. This study incorporated the values of the final third of Link et al. [3] and added both a linear gradient in the x-direction and a Gaussian gradient in the y-direction for the rest of the pitch. No further empirical evidence was found for this Pitch Value model, as no suitable experimentation was known to be conducted. As indicated in the previous chapter, this model can be viewed as a tactical vision and should not be treated as universal. The results in passing options do argue for this model to some extent.

6.2 Passing Options

The first step to generating future scenarios was the identification of pass options. This study examined a passing option model, based on Pitch Control, Pitch Value and a distance gradient. The distance gradient was fitted to a Gaussian distribution, which

showed remarkable resemblance to the observed pass distances. The overall passing option model tuned by the means of a grid search to result in both a realistic distribution in passing distances as well as a resemblance to the observed pass destinations (fig. 4.3). The resulting model proved to generate realistic pass options distances (fig. 3.10 & 4.2). It did not consistently included the actual pass, only for 30% of the data; it produced a predictive value that was lower than field standards (fig. 5.2).

Future work can focus on examining the universality of the Guassian distribution. The model has to be tested against other data, in order understand whether there was over-fitting in this study. In addition, once the Pass Probability model is extend to air passes as well, the Pitch Control model and distance gradient might be replaced in order to produce a more realistic modelling of the conceptual passes. Lastly, the inclusion of fatigue will result in a more realistic model. Speed and precision of shooting as well as passing are found to decline over time in matches [70, 71].

6.3 Movement Prediction

This study explored the performance of two algorithms on movement prediction: *dead reckoning* and the developed *ball oriented dead reckoning*. The former was taken to form a new baseline in movement prediction, which proved itself a extremely robust predictor of movement during passes (fig. 4.5). The latter was expected to improve the baseline, as it incorporates some simple game-dynamics. This was observed (fig. 4.5 & 4.6). Most importantly, the predictions for the receiver were significantly improved; the predictions for the other players were not improved to the extent as expected. Both algorithms outperform the field standards [5, 6].

Future work can focus on improving the dead reckoning algorithm by further incorporation of game-dynamics, e.g. distance to the ball [4], possible accelerations due to current speed of the player, the coverage of the field or the effect of fatigue. The latter could be important as distance coverage declines with 20% in a match [62, 72] and possibly explains why more goals are scored towards the end of the game [73, 74]. Studying the effect of including these aspects will improve the scientific understanding of football.

6.4 Pass Quantification

6.4.1 Standard pass quantification

In extension to the research of [7], two standard methods in pass quantification were developed: number of outplayed opponents, computed on both x and y, and the change in controlled pitch, by the use of Pitch Control and Pitch Value. In contrast to Rein et al. [7] and unexpectedly, the former showed no correlation to the descriptive metrics (fig. 4.7). The latter showed little correlation to the descriptive metrics (fig. 4.8), which was expected. The explanation might be the choice of metrics, the short time-frames of 5 minutes could be imperfect, or the fact that the number of outplayed opponents does not quantify passes effectively. Further research can be done by investigating other metrics, which is a general problem in football analytics. Once a more extensive data set is available, the analysis can be extended and possibly lead to new insights.

6.4.2 Future scenario analysis

Lastly, the aforementioned models were combined in order to conduct an analysis on future scenarios at the start and the end of the pass. The two resulting pass quantification models (fig. 4.9 & 4.10) did not show improvements compared to the standard pass quantification models. The reason for this can be all sorts of things. Again, the chosen metric might not be justified and should be studied and/or varied. In addition, the dead reckoning model on movement prediction is possibly not optimal for hypothetical scenarios, which is difficult to study. Nevertheless, the concept of using future scenarios is not yet to be neglected.

Future research can examine the influence of different movement prediction models on the future scenario quantification. The computation method to combine the values of future scenarios could also be varied in order to understand the influence of the future scenarios on the whole model better. If this is all done in time, Frenkie de Jong will still be able to carry out that perfect pass and lead the Dutch team to victory; the World Cup.

Appendix A

Kalman Filter

A.1 Concept

Kalman filtering is a common technique to reduce noise in tracking data; it is used across applications of signal processing and navigation [75, 76]. The principle of this technique relies on estimating the real position based on noise magnitude and the preceding observation, schematically depicted in figure A.1. The concept is best illustrated by this example: if a car drives through a tunnel, its GPS signal will be distorted. Based on the velocity preceding the tunnel, the noise can by reduced by orienting at the expected position of the car. The estimated position is based on the variance of expectations and observations.

The technique works in an online manner: only observations of the previous and current iteration are used. The process (fig. A.1) is a constant repetition of predicting the next data point and updating/correcting it to measurements [75]. Mathematically, the predictive (or a-priori) variables are given by

$$\textit{Predicted state estimate} \qquad \hat{\boldsymbol{x}}_{k|k-1} = \boldsymbol{F}_k \hat{\boldsymbol{x}}_{k-1|k-1} + \boldsymbol{B}\hat{\boldsymbol{u}}_k \tag{A.1}$$

$$\textit{Predicted estimate covariance} \qquad \boldsymbol{P}_{k|k-1} = \boldsymbol{F}_k \boldsymbol{P}_{k-1|k-1} \boldsymbol{F}_k^T + \boldsymbol{Q}_k \tag{A.2}$$

where $\boldsymbol{F}_k$ is the state-transition model describing the system dynamics, $\hat{\boldsymbol{x}}_{k-1|k-1}$ is the previous a-posteriori estimate and $\boldsymbol{B}$ the control-input model that is applied to an optional control input $\hat{\boldsymbol{u}}_k$, which are both not used here. After calculating the a-priori estimates, the next observation $\boldsymbol{z}_k$ is used to update the estimate. This is done by the

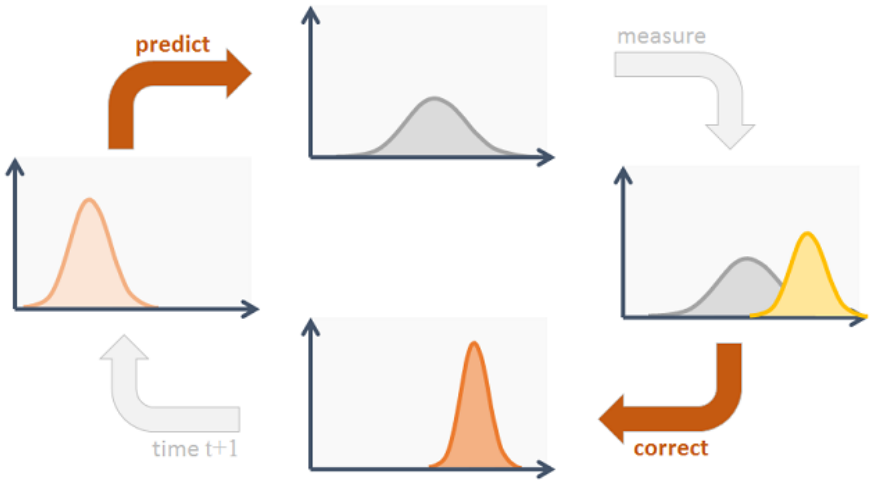


FIGURE A.1: Kalman filter schematically represented, taken from Jurić [77]. The process starts on the left with an observation and an estimated inaccuracy of the observation. Based on the previous observation(s) and these inaccuracies, a prediction for the next position is generated. Next, the actual measurement is observed. The prediction is corrected based on this observation to arrive at the final estimate. The final estimate for time t forms the start of the next iteration $t+1$.

following rules

$$\textit{Measurement pre-fit residual} \qquad \tilde{\boldsymbol{y}}_k = \boldsymbol{z}_k - \boldsymbol{H}_k \hat{\boldsymbol{x}}_{k|k-1} \tag{A.3}$$

$$\textit{Pre-fit residual covariance} \qquad \boldsymbol{S}_k = \boldsymbol{H}_k \boldsymbol{P}_{k|k-1} \boldsymbol{H}_k^\top + \boldsymbol{R}_k \tag{A.4}$$

$$\textit{Kalman gain} \qquad \boldsymbol{K}_k = \boldsymbol{P}_{k|k-1} \boldsymbol{H}_k^\top \boldsymbol{S}_k^{-1} \tag{A.5}$$

$$\textit{Updated state estimate} \qquad \hat{\boldsymbol{x}}_{k|k} = \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k \tilde{\boldsymbol{y}}_k \tag{A.6}$$

$$\textit{Updated estimate covariance} \qquad \boldsymbol{P}_{k|k} = (\boldsymbol{I} - \boldsymbol{K}_k \boldsymbol{H}_k)\, \boldsymbol{P}_{k|k-1} \tag{A.7}$$

$$\textit{Measurement post-fit residual} \qquad \tilde{\boldsymbol{y}}_{k|k} = \boldsymbol{z}_k - \boldsymbol{H}_k \hat{\boldsymbol{x}}_{k|k} \tag{A.8}$$

where we need to define the observation model $\boldsymbol{H}_k$, the state-transition model $\boldsymbol{F}_k$, the covariance of the process noise $\boldsymbol{Q}_k$ and the covariance of the observation noise $\boldsymbol{R}_k$.

A.2 Implementation

The tracking data consists of the positions of players, given in 2 dimension x, y. The corresponding observed velocities $\dot{x}, \dot{y}$ are used for prediction of the next position. Therefore, the estimate states are described by the vector

$$\boldsymbol{x} = \begin{bmatrix} x & y & \dot{x} & \dot{y} \end{bmatrix}. \tag{A.9}$$

The observational model translates these states back to the dimensions of the observations, i.e. a positional vector. This implies

$$\boldsymbol{H}_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{A.10}$$

The state-transition model $\boldsymbol{F}_k$ describes how the next position is estimated based on the estimate state (position and speed), which is basic kinematic physics: $x_1 = x_0 + v_0 * \Delta t$. Taking into account how the states are defined, this model is defined as

$$\boldsymbol{F}_k = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{A.11}$$

The covariance of the process noise $\boldsymbol{Q}_k$ describes how deviations in observations are processed and evolve through estimations. For positional variables, this is considered to be of order $\sigma_x = \sigma_y = \frac{\Delta t^2}{2}$ and for velocity variables of order $\sigma_{\dot{x}} = \sigma_{\dot{y}} = \Delta t$ [75, 78]. In addition, the magnitude of the standard deviation of the acceleration σ_a must be defined. For this study, this was set at Δt. The process noice matrix is then given by

$$\boldsymbol{Q}_k = \begin{bmatrix} \sigma_x^2 & 0 & \sigma_x\sigma_{\dot{x}} & 0 \\ 0 & \sigma_y^2 & 0 & \sigma_y\sigma_{\dot{y}} \\ \sigma_{\dot{x}}\sigma_x & 0 & \sigma_{\dot{x}}^2 & 0 \\ 0 & \sigma_{\dot{y}}\sigma_y & 0 & \sigma_{\dot{y}}^2 \end{bmatrix} \sigma_a^2 = \begin{bmatrix} \frac{\Delta t^4}{4} & 0 & \frac{\Delta t^3}{2} & 0 \\ 0 & \frac{\Delta t^4}{4} & 0 & \frac{\Delta t^3}{2} \\ \frac{\Delta t^3}{2} & 0 & \Delta t^2 & 0 \\ 0 & \frac{\Delta t^3}{2} & 0 & \Delta t^2 \end{bmatrix} \Delta t^2 \tag{A.12}$$

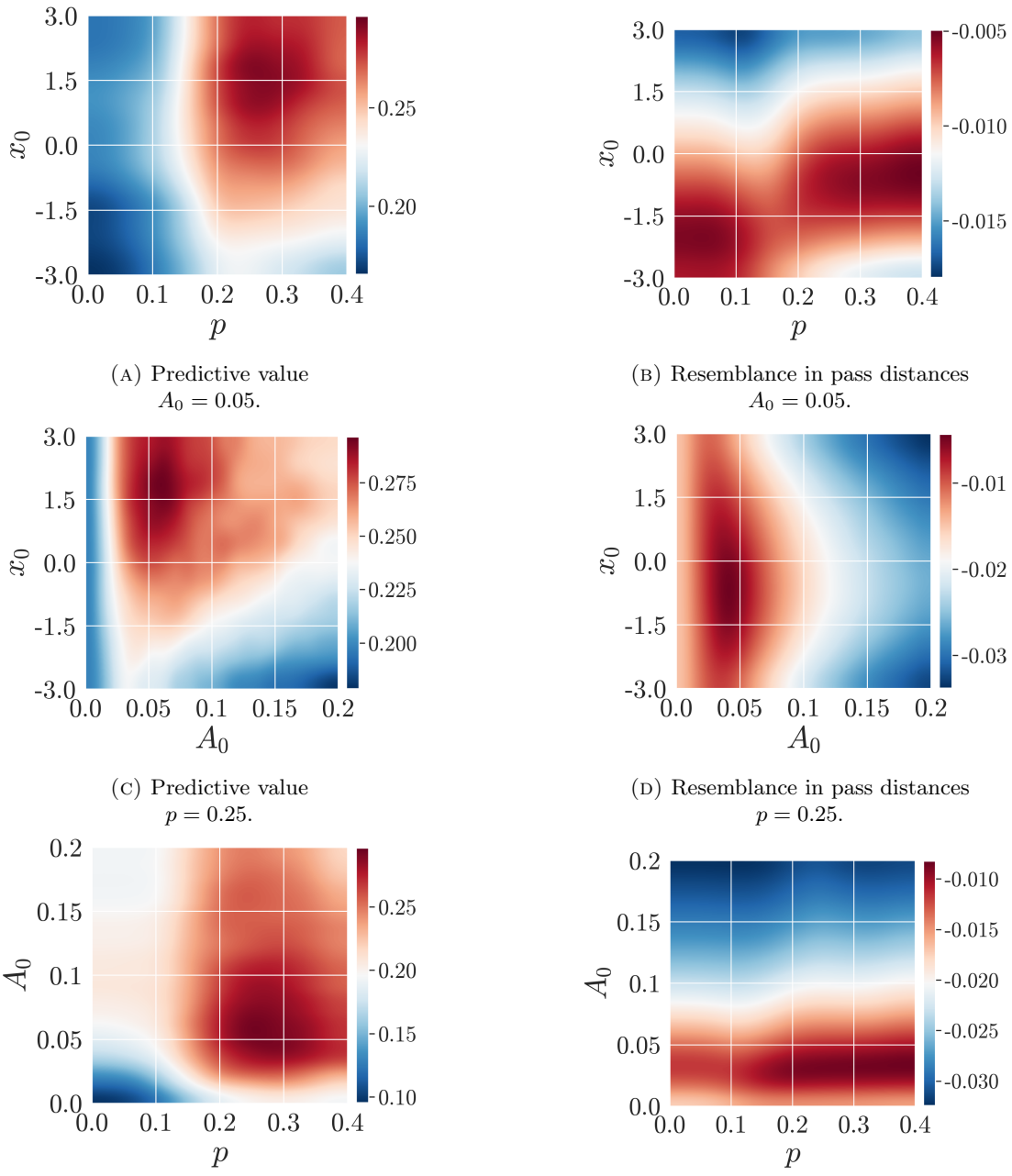
Lastly, the observation noise $\boldsymbol{R}_k$ describes the expected magnitude of noise in the observations. This was set at Δt for both positional variables and without covariance between the two:

$$\boldsymbol{R}_k = \begin{bmatrix} \Delta t & 0 \\ 0 & \Delta t \end{bmatrix} \tag{A.13}$$

Appendix B

Figures parameter tuning

B.1 Pass options model



(A) Predictive value $A_0 = 0.05$.

(B) Resemblance in pass distances $A_0 = 0.05$.

(C) Predictive value $p = 0.25$.

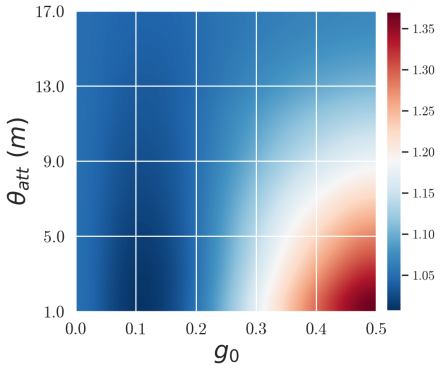
(D) Resemblance in pass distances $p = 0.25$.

(E) Predictive value $x_0 = 2.0$.

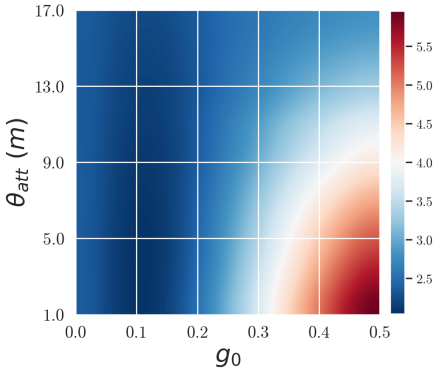
(F) Resemblance in pass distances $x_0 = 2.0$.

FIGURE B.1: Grid search on the pass option model. The values of amplitude $A_0 = 0.05$, shift $x_0 = 2$ and range factor $p = 0.25$ were found to optimise the predictive value and the resemblance to the pass distance distribution (fig. 3.10). The two were computed respectively by the frequency of identifying the actual pass, evaluated at 0.296, and the RMSE to the pass distance distribution, evaluated at 0.012. Given these value, heat maps are shown per duo of parameters for the tested ranges.

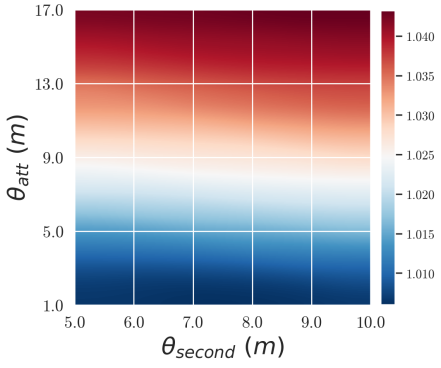
B.2 Ball oriented dead reckoning



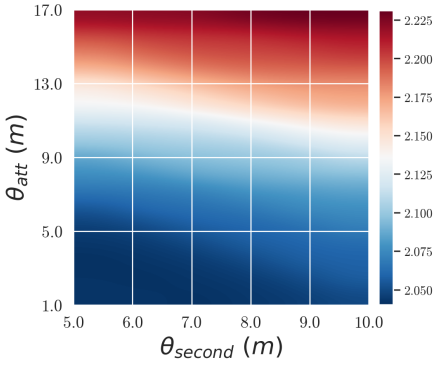
(A) Mean
$\theta_{second} = 5$, $\theta_{def} = 3$.



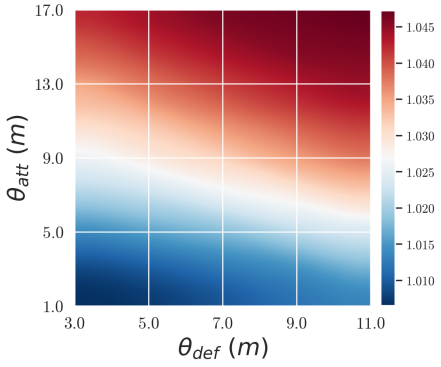
(B) Variance
$\theta_{second} = 5$, $\theta_{def} = 3$.



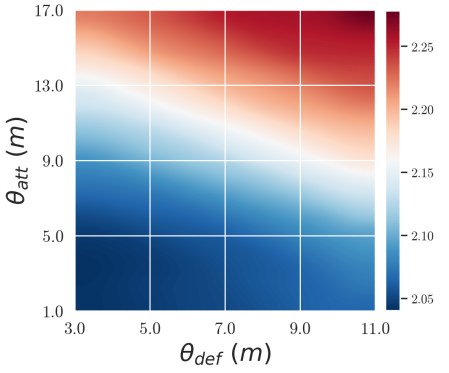
(C) Mean
$g_0 = 0.1$, $\theta_{def} = 3$.



(D) Variance
$g_0 = 0.1$, $\theta_{def} = 3$.

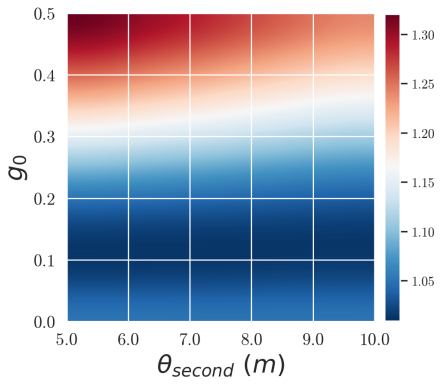


(E) Mean
$g_0 = 0.1$, $\theta_{second} = 5$.

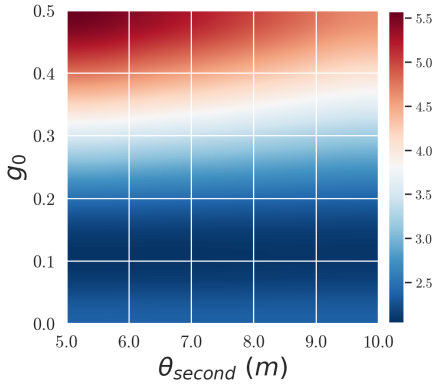


(F) Variance
$g_0 = 0.1$, $\theta_{second} = 5$.

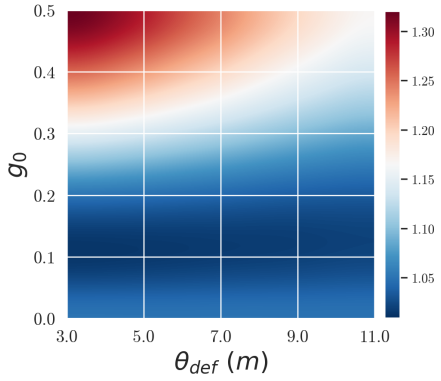
FIGURE B.2: Grid search on the strong acceleration towards the pass destination, if no player is near. The values found to minimise both mean and variance are the distance thresholds $\theta_{att} = 3$, $\theta_{second} = 5$, $\theta_{def} = 3$ and factor $g_0 = 0.1$ of adapting to a velocity towards the pass destination. Given these value, heat maps are shown per duo of parameters for the tested ranges.



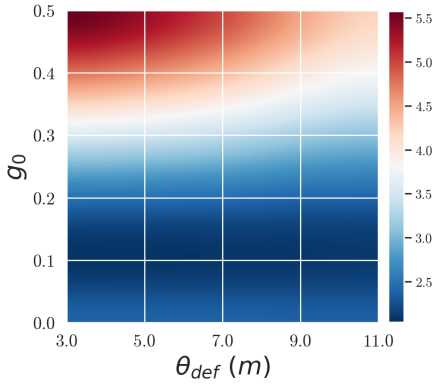
(A) Mean
$\theta_{att} = 3$, $\theta_{def} = 3$.



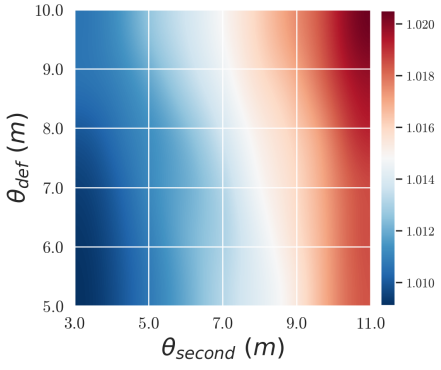
(B) Variance
$\theta_{att} = 3$, $\theta_{def} = 3$.



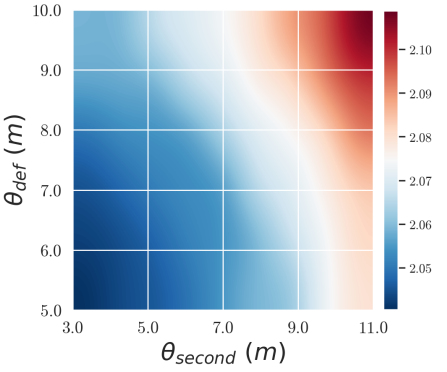
(C) Mean
$\theta_{att} = 3$, $\theta_{second} = 5$.



(D) Variance
$\theta_{att} = 3$, $\theta_{second} = 5$.



(E) Mean
$\theta_{att} = 3$, $g_0 = 0.1$.



(F) Variance
$\theta_{att} = 3$, $g_0 = 0.1$.

FIGURE B.3: Grid search on the strong acceleration towards the pass destination, in case no player of a team near the location. The values found to minimise both mean and variance are the distance thresholds $\theta_{att} = 3$, $\theta_{second} = 5$, $\theta_{def} = 3$ and a factor $g_0 = 0.1$ of adapting to a velocity towards the pass destination. Given these value, heat maps are shown per duo of parameters for the tested ranges.

Bibliography

[1] William Spearman, Austin Basye, Greg Dick, Ryan Hotovy, and Paul Pop. Physics-Based Modeling of Pass Probabilities in Soccer. *Research Papers Competition*, pages 1–14, March 2017.

[2] William Spearman. Beyond Expected Goals. *12th Annual MIT Sloan Sports Analytics Conference*, pages 1–17, August 2018.

[3] Daniel Link, Steffen Lang, and Philipp Seidenschwarz. Real time quantification of dangerousity in football using spatiotemporal tracking data. *PLoS ONE*, 11(12): 1–16, 2016. ISSN 19326203. doi: 10.1371/journal.pone.0168768.

[4] Philippe Fournier-Viger, Tianbiao Liu, and Jerry Chun-Wei Lin. Football pass prediction using player locations. In Ulf Brefeld, Jesse Davis, Jan Van Haaren, and Albrecht Zimmermann, editors, *Machine Learning and Data Mining for Sports Analytics*, pages 152–158, Cham, 2019. Springer International Publishing. ISBN 978-3-030-17274-9.

[5] Hoang M. Le, Yisong Yue, Peter Carr, and Patrick Lucey. Coordinated multi-agent imitation learning. *34th International Conference on Machine Learning, ICML 2017*, 4:3140–3152, 2017. ISSN 1938-7228.

[6] Francisco Peralta Alguacil, Javier Fernandez, and Pablo Piñones Arce. Seeing in to the future : using self-propelled particle models to aid player decision-making in soccer. *MIT Sloan Sports Analytics Conference*, pages 1–23, 2020.

[7] Robert Rein, Dominik Raabe, and Daniel Memmert. "Which pass is better?" Novel approaches to assess passing effectiveness in elite soccer. *Human Movement Science*, 55(August):172–181, 2017. ISSN 18727646. doi: 10.1016/j.humov.2017.07.010.

[8] B. Dagnino. Metrica sports sample data. https://github.com/metrica-sports/sample-data, 2021. Accessed: 2021-04-12.

[9] Jim Albert. Streakiness in team performance. *Chance*, 17(3):37–43, 2004.

[10] T. Taki, J. Hasegawa, and T. Fukumura. Development of motion analysis system for quantitative evaluation of teamwork in soccer games. In *Proceedings of 3rd IEEE International Conference on Image Processing*, volume 3, pages 815–818 vol.3, 1996. doi: 10.1109/ICIP.1996.560865.

[11] Alen Rajšp and Iztok Fister. A systematic literature review of intelligent data analysis methods for smart sport training. *Applied Sciences*, 10(9), 2020. ISSN 2076-3417. doi: 10.3390/app10093013.

[12] Sian Barris and Chris Button. A review of vision-based motion analysis in sport. *Sports Medicine*, 38(12):1025–1043, 2008.

[13] Luke Bornn, Dan Cervone, and Javier Fernandez. Soccer analytics: Unravelling the complexity of "the beautiful game". *Significance*, 15(3):26–29, 2018. ISSN 17409713. doi: 10.1111/j.1740-9713.2018.01146.x.

[14] Sofia Fonseca, João Milho, Bruno Travassos, and Duarte Araújo. Spatial dynamics of team sports exposed by voronoi diagrams. *Human Movement Science*, 31(6):1652 – 1659, 2012. ISSN 0167-9457. doi: https://doi.org/10.1016/j.humov.2012.04.006.

[15] Joachim Gudmundsson and Thomas Wolle. Football analysis using spatio-temporal tools. *Computers, Environment and Urban Systems*, 47:16–27, 2014. ISSN 01989715. doi: 10.1016/j.compenvurbsys.2013.09.004.

[16] Javier Fernandez and Luke Bornn. Wide Open Spaces : A statistical technique for measuring space creation in professional soccer. *MIT Sloan Sports Analytics Conference*, pages 1–19, 2018. URL http://www.lukebornn.com/sloan/space{_}occupation{_}1.mp4.

[17] Floris R. Goes, Matthias Kempe, Laurentius A. Meerhoff, and Koen A.P.M. Lemmink. Not Every Pass Can Be an Assist: A Data-Driven Model to Measure Pass Effectiveness in Professional Soccer Matches. *Big Data*, 7(1):57–70, 2019. ISSN 2167647X. doi: 10.1089/big.2018.0067.

[18] Luca Pappalardo, Paolo Cintia, Paolo Ferragina, Emanuele Massucco, Dino Pedreschi, and Fosca Giannotti. Playerank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5):1–27, 2019.

[19] Ronald Yurko, Francesca Matano, Lee F Richardson, Nicholas Granered, Taylor Pospisil, Konstantinos Pelechrinis, and Samuel L Ventura. Going deep: models for continuous-time within-play valuation of game outcomes in american football with tracking data. *Journal of Quantitative Analysis in Sports*, 1(ahead-of-print), 2020.

[20] Rajiv Maheswaran, Yu-Han Chang, Aaron Henehan, and Samantha Danesis. Deconstructing the rebound with optical tracking data. In *Proceedings of the 6th annual MIT SLOAN sports analytics conference*, 2012.

[21] Javier Fernández, Luke Bornn, and Dan Cervone. Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer. *MIT Sloan Sports Analytics Conference*, pages 1–18, 2019.

[22] Hoang M. Le, Carr Peter, and Yisong Yue. Data-Driven Ghosting using Deep Imitation Learning. *MIT SLoan Sports Analytics Conference*, pages 1–15, 2017.

[23] Paolo Cintia, Salvatore Rinzivillo, and Luca Pappalardo. A network-based approach to evaluate the performance of football teams. 09 2015.

[24] I. Gómez. Fitting your own football xg model. https://www.datofutbol.cl/xg-model/, 2020. Accessed: 2021-06-04.

[25] Tom Decroos, Jan van Haaren, Lotte Bransen, and Jesse Davis. Actions speak louder than goals: Valuing player actions in soccer. *arXiv*, pages 1851–1861, 2018. ISSN 23318422.

[26] Xiangyu Sun, Jack Davis, Oliver Schulte, and Guiliang Liu. Cracking the black box: Distilling deep sports analytics. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '20, page 3154–3162, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403367.

[27] Claude E Shannon. Xxii. programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 41(314): 256–275, 1950.

[28] J. Tromp. John's chess playground. https://tromp.github.io/chess/chess.html, 2010. Accessed: 2021-07-29.

[29] Paul Power, Hector Ruiz, Xinyu Wei, and Patrick Lucey. "Not all passes are created equal:" Objectively measuring the risk and reward of passes in soccer from tracking data. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Part F1296:1605–1613, 2017. doi: 10.1145/3097983.3098051.

[30] Sanjay Chawla, Joël Estephan, Joachim Gudmundsson, and Michael Horton. Classification of passes in football matches using spatiotemporal data. *ACM Transactions on Spatial Algorithms and Systems*, 3(2), 2017. ISSN 23740361. doi: 10.1145/3105576.

[31] Ian G. McHale and Samuel D. Relton. Identifying key players in soccer teams using network analysis and pass difficulty. *European Journal of Operational Research*, 268(1):339–347, 2018. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2018.01.018.

[32] Else Marie Haland, Astrid Salte Wiig, Magnus Stalhane, and Lars Magnus Hvattum. Evaluating passing ability in association football. *IMA journal of management mathematics*, 31(1):91–116, 2020. ISSN 1471-678X.

[33] L. Scott. Packing in the bundesliga – data analysis. https://totalfootballanalysis.com/data-analysis/packing-in-the-bundesliga-data-analysis, 2020. Accessed: 2021-05-20.

[34] Panna Felsen, Patrick Lucey, and Sujoy Ganguly. Where Will They Go? Predicting Fine-Grained Adversarial Multi-agent Motion Using Conditional Variational Autoencoders. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11215 LNCS: 761–776, 2018. ISSN 16113349. doi: 10.1007/978-3-030-01252-6_45.

[35] Silvan Steiner, Stephan Rauh, Martin Rumo, Karin Sonderegger, and Roland Seiler. Using position data to estimate effects of perceptual features of play on passing decisions in soccer. *Current Issues in Sport Science (CISS)*, 3:1–9, 2018. ISSN 2414-6641. doi: 10.15203/ciss_2018.009.

[36] Andreas Glöckner, Thomas Heinen, Joseph G. Johnson, and Markus Raab. Network approaches for expert decisions in sports. *Human Movement Science*, 31(2):318–333, 2012. ISSN 01679457. doi: 10.1016/j.humov.2010.11.002.

[37] Juan Carlos Núñez and Bruno Dagnino. Exploring the application of soccer mathematical models to game generation on a simulated environment. *Sports Tomorrow FC Barcelona*, pages 1–10, 2020.

[38] Karol Kurach, Anton Raichuk, Piotr Stanczyk, Michał Zajac, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly. Google Research Football: A novel reinforcement learning environment. *arXiv*, 2019. ISSN 23318422. doi: 10.1609/aaai.v34i04.5878.

[39] A. Chacoma, N. Almeira, J. I. Perotti, and O. V. Billoni. Modeling ball possession dynamics in the game of football. *Physical Review E*, 102(4):1–8, 2020. ISSN 24700053. doi: 10.1103/PhysRevE.102.042120.

[40] Matthew Oldham and Andrew T. Crooks. Drafting agent-based modeling into basketball analytics. *Simulation Series*, 51(1), 2019. ISSN 07359276. doi: 10.23919/SpringSim.2019.8732893.

[41] L. Shaw. Laurieontracking. https://github.com/Friends-of-Tracking-Data-FoTD/LaurieOnTracking, 2020. Accessed: 2021-02-28.

[42] Andrew Rowlinson. Football Shot Quality: Visualizing the Quality of Soccer/Football Shots. Master's thesis, Aalto University. School of Business, 2020. URL http://urn.fi/URN:NBN:fi:aalto-202008234885.

[43] Amir Yahyavi, Kévin Huguenin, and Bettina Kemme. Interest modeling in games: the case of dead reckoning. *Multimedia systems*, 19(3):255–270, 2013.

[44] Cliff Randell, Chris Djiallis, and Henk Muller. Personal position measurement using dead reckoning. In *Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings.*, pages 166–166. IEEE Computer Society, 2003.

[45] Wei Chen, Ruizhi Chen, Yuwei Chen, Heidi Kuusniemi, and Jianyu Wang. An effective pedestrian dead reckoning algorithm using a unified heading error model. In *IEEE/ION Position, Location and Navigation Symposium*, pages 340–347. IEEE, 2010.

[46] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

[47] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2.

[48] Siu Kwan Lam, Antoine Pitrou, and Stanley Seibert. Numba: A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pages 1–6, 2015.

[49] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. Apache spark: A unified engine for big data processing. *Commun. ACM*, 59(11):56–65, October 2016. ISSN 0001-0782. doi: 10.1145/2934664.

[50] Xinyi Zhou, Wei Gong, WenLong Fu, and Fengtong Du. Application of deep learning in object detection. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pages 631–634. IEEE, 2017.

[51] Liyu Liu and Moeness G. Amin. Tracking performance and average error analysis of gps discriminators in multipath. *Signal Processing*, 89(6):1224–1239, 2009. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2009.01.007.

[52] Pramod R. Gunjal, Bhagyashri R. Gunjal, Haribhau A. Shinde, Swapnil M. Vanam, and Sachin S. Aher. Moving object tracking using kalman filter. In *2018 International Conference On Advances in Communication and Computing Technology (ICACCT)*, pages 544–547, 2018. doi: 10.1109/ICACCT.2018.8529402.

[53] A. Mathisen, S. Krogh Sørensen, A. Stisen, H. Blunck, and K. Grønbæk. A comparative analysis of indoor wifi positioning at a large building complex. In *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pages 1–8, 2016. doi: 10.1109/IPIN.2016.7743666.

[54] C. Carroll. Who are the fastest players at euro 2020? https://statsports.com/who-are-the-fastest-players-at-euro-2020/, 2019. Accessed: 2021-05-25.

[55] A. Smyth. Top 10 fastest football players in the world in 2020. https://acefootball.com/football-news/top-10-fastest-football-players-in-the-world-in-2020/, 2020. Accessed: 2021-05-25.

[56] Thomas Little and Alun Williams. Specificity of acceleration, maximum speed, and agility in professional soccer players. *Journal of strength and conditioning research / National Strength & Conditioning Association*, 19:76–8, 03 2005. doi: 10.1519/14253.1.

[57] F.J.P. Alguacil. Modelling the collective movement of football players. Master's thesis, Uppsala University, Department of Information Technology, 2019.

[58] Tara Thakur. A study on variation of reaction time with respect to playing positions of football players. *IOSR-JOURNAL OF SPORTS AND PHYSICAL EDUCATION*, 3:2347–6745, 02 2016. doi: 10.9790/6737-0313032.

[59] Emre Külah and Hande Alemdar. Quantifying the value of sprints in elite football using spatial cohesive networks. *Chaos, Solitons & Fractals*, 139:110306, 2020. ISSN 0960-0779. doi: https://doi.org/10.1016/j.chaos.2020.110306.

[60] Alex Rathke. An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise*, 12(2):514–529, 2017.

[61] Ricardo Duarte, Duarte Araújo, Vanda Correia, Keith Davids, Pedro Marques, and Michael J. Richardson. Competing together: Assessing the dynamics of team–team and player–team synchrony in professional association football. *Human Movement Science*, 32(4):555–566, 2013. ISSN 0167-9457. doi: https://doi.org/10.1016/j.humov.2013.01.011.

[62] Daniel Linke, Daniel Link, Hendrik Weber, and Martin Lames. Decline in match running performance in football is affected by an increase in game interruptions. *Journal of sports science & medicine*, 17(4):662, 2018.

[63] R. S. Mendes, L. C. Malacarne, and C. Anteneodo. Statistics of football dynamics. *European Physical Journal B*, 57(3):357–363, 2007. ISSN 14346028. doi: 10.1140/epjb/e2007-00177-4.

[64] Takuma Narizuka, Ken Yamamoto, and Yoshihiro Yamazaki. Statistical properties of position-dependent ball-passing networks in football games. *Physica A: Statistical Mechanics and its Applications*, 412:157–168, 2014. ISSN 0378-4371. doi: https://doi.org/10.1016/j.physa.2014.06.037.

[65] Daniel Linke, Daniel Link, and Martin Lames. Football-specific validity of tracab's optical video tracking systems. *PLOS ONE*, 15(3):1–17, 03 2020. doi: 10.1371/journal.pone.0230179.

[66] MJ Carré, T Asai, T Akatsuka, and SJ Haake. The curve kick of a football ii: flight through the air. *Sports Engineering*, 5(4):193–200, 2002.

[67] Athanasios Katis, Emmanouil Giannadakis, Theodoros Kannas, Ioannis Amiridis, Eleftherios Kellis, and Adrian Lees. Mechanisms that influence accuracy of the soccer kick. *Journal of Electromyography and Kinesiology*, 23(1):125–131, 2013.

[68] John Eric Goff and Matt J Carré. Trajectory analysis of a soccer ball. *American Journal of Physics*, 77(11):1020–1027, 2009.

[69] TB Andersen and HC Dörge. The influence of speed of approach and accuracy constraint on the maximal speed of the ball in soccer kicking. *Scandinavian journal of medicine & science in sports*, 21(1):79–84, 2011.

[70] Mark Russell, David Benton, and Michael Kingsley. The effects of fatigue on soccer skills performed during a soccer match simulation. *International journal of sports physiology and performance*, 6(2):221–233, 2011.

[71] R Izzo, U Rossini, G Raiola, A Cejudo Palomo, and C Hosseini Varde'i. Insurgence of fatigue and its implications in the selection and accuracy of passes in football. a case study. *Journal of Physical Education and Sport*, 20(4):1996–2002, 2020.

[72] Ricardo Izzo, Tiziana D'isanto, Gaetano Raiola, Antonio Cejudo, Nasar Ponsano, and Ciro Hosseini Varde'i. The role of fatigue in football matches, performance model analysis and evaluation during quarters using live global positioning system technology at 50hz. *Sport Science*, 13(1):30–35, 2020.

[73] Vasilis Armatas and Michail Mitrotasios. Analysis of goal scoring patterns in the 2012 european football championship. *The Sport Journal*, 01 2014.

[74] F. Bohrmann. Goal time analysis. http://www.soccerstatistically.com/blog/2013/7/15/goal-time-analysis.html, 2013. Accessed: 2021-02-09.

[75] Y Kim and H Bang. *Introduction and Implementations of the Kalman Filter*. IntechOpen, 2019. ISBN 1-83880-739-X.

[76] Howard Musoff and Paul Zarchan. *Fundamentals of Kalman filtering: a practical approach*. American Institute of Aeronautics and Astronautics, 2009.

[77] Darko Jurić. Object tracking: Kalman filter with ease, 2015.

[78] R. Sadli. Object tracking: 2-d object tracking using kalman filter in python. https://machinelearningspace.com/2d-object-tracking-using-kalman-filter/, 2020. Accessed: 2021-05-03.