Modeling and Analyzing Pass Values in Football

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Business Analytics and Operations Research

Tilburg School of Economics and Management
Tilburg University

Company supervisors:
Jon Daal MSc
(Pipple)
Maarten de Bruijn
(FC Eindhoven)

Author:
Annemarijn Blom
SNR: 2088078

Supervisor Tilburg University
dr.mr. Sven Polak

Second reader:
dr.ir. Pieter Kleer

May 23, 2024

Abstract

This thesis introduces a model to value passes in football, based on the expert opinion of the staff at FC Eindhoven. This model is constructed by combining Expected Threat, Packing and Pitch Control together in a linear model. The parameters are estimated based on experts that have rated 34 passes, and the model is selected based on different measures, the (Root) Mean Squared Error, the Mean Absolute Error and R^2. With the constructed model, we looked into several possible relationships of the pass value with other metrics, such as pass distance, player position, pass accuracy, ball possession and the number of goals. We do this with the Ordinary Least Squares method and by computing the correlation coefficient. For the pass accuracy and ball possession, we find small negative relationships which can be explained due to the fact that we only use successful passes, and for the number of goals we get a small positive relationship. Furthermore, a promising result is that the subtypes that are labeled in the event data, are rated on average higher than the normal event type, and key-passes and assists get the highest rating. Finally, we look into the potential practical applications the pass value model could have. We look into alternative passes: was the executed pass the best option or were there way better alternatives? We also look into rating players on their pass value score and compare this with real ratings. Both of these applications can be interesting for FC Eindhoven to use in practice.

Contents

Acknowledgements **4**

Management Summary **5**

1 Introduction **6**
1.1 Pipple and FC Eindhoven . . . 7
1.2 Outline of the report . . . 7

2 Related Work **8**
2.1 Expected Goals . . . 8
2.2 Expected Threat . . . 8
2.3 Packing . . . 9
2.4 Pitch Control . . . 9
2.5 Valuing passes . . . 10
2.5.1 Valuing passes with event data . . . 10
2.5.2 Physics-based approach for pass probabilities . . . 11
2.5.3 Expected passes . . . 11
2.5.4 Pass quality in football . . . 11

3 Data Description **12**
3.1 Tracking data . . . 12
3.1.1 Metadata . . . 12
3.1.2 Preprocessing . . . 13
3.2 Event data . . . 13
3.2.1 Preprocessing . . . 14
3.3 Linking tracking and event data . . . 14
3.4 Expert data . . . 15

4 Methodology **17**
4.1 Valuing passes . . . 17
4.1.1 Start and end coordinates . . . 17
4.1.2 Expected Threat . . . 19
4.1.3 Packing . . . 21
4.1.4 Pitch Control . . . 21
4.1.5 Pass value . . . 24
4.1.6 Parameter estimation and model selection . . . 24
4.2 Relationship analysis . . . 27
4.2.1 Correlation and regression analysis . . . 27
4.2.2 Comparative analysis . . . 28
4.3 Applications . . . 29
4.3.1 Considering alternative passes . . . 29
4.3.2 Player rating . . . 29
4.3.3 Relationship analysis . . . 30

5 Results **31**
5.1 Pass value . . . 31
5.1.1 Parameter estimation . . . 31
5.1.2 Model selection . . . 32
5.2 Relationship Analysis . . . 34
5.2.1 Correlation and regression analysis . . . 34
5.2.2 Comparative analysis . . . 38

5.3 Applications 39
5.3.1 Considering alternative passes 39
5.3.2 Player rating 41
5.3.3 Relationship analysis 43

6 Conclusion **48**
6.1 Summary 48
6.2 Conclusions 48
6.3 Future research 49

References **51**

A Appendix **53**
A.1 Expert data 53
A.2 Alternative passes 54
A.3 Player rating 55

Acknowledgements

For the past few months, I have been working on this thesis, and I could not have done this without the help of some people, who I want to thank for their help and guidance during this process.

First of all, I want to thank my supervisor from the university, dr.mr. Sven Polak. During the process, you regularly gave constructive feedback to my work, which helped me making every version a better one. Additionally, I want to thank dr.ir. Pieter Kleer for taking the time to be the second reader of my work.

Secondly, I want to thank my supervisor from Pipple, Jon Daal, for guiding me through this process. Every week you took the time to help me with whichever part I needed help on, and you were very enthusiastic about the subject. When I was a bit stuck, you were able to get me out of that. And if I needed a bit of a distraction, you were always willing to make the weekly crossword puzzle or play some foosball. Next to that, I want to thank Vera van der Lelij for our monthly chats about the not content related things, like how I was enjoying my time at Pipple or to vent whenever I needed that. Also, I want to thank all the other Pipple colleagues who made my time at the office very fun. I am very glad I do not have to leave you yet!

I also want to thank FC Eindhoven for the subject of this thesis, and the opportunity to use their data. It was also very nice to have a session with the staff, to get some more information on what the relevant people in such a club find important.

Last, but not least, I want to thank all my friends and family that have supported my during this process: for listening to me talking about football again and again, supporting me through the tougher times but also for all the times they gave me fun distractions by doing fun stuff together. I want to end with a special thanks to my boyfriend Martijn, who supported me in every way possible during my thesis time. From cooking dinner so many times, to helping me when I could not find the error in my code, and for proofreading my thesis as well.

Management Summary

In this thesis, we focus on constructing a representative pass value model for the football club FC Eindhoven. They provided us with tracking and event data from the matches of this season. There are a lot of potential factors to take into account when valuing a pass, and we have chosen three of these based on the suggestions of staff from FC Eindhoven. First, we consider Expected Threat. This is a metric that assigns a threat value to every location on the pitch, and can be used to measure the contribution a pass has made to the buildup play towards a shot on goal. Secondly, we take into account the Packing, which is the amount of opponents that are bypassed during a pass, which is useful because it is more likely that you are in a scoring position when there are fewer opponents between you and the goal. Lastly, we consider Pitch Control, which is the probability that a team will get possession of the ball, if it moves to a certain location on the field. In contrary to the other two metrics, this metric is less about the potential impact of a pass towards scoring and more about the risk if the pass will succeed or not.

Combining these three metrics, we construct two potential pass value models. To determine the parameters in these models, we use the data we have collected from experts that considered 34 schematic overviews of passes and rated these passes. Furthermore, we measure which of these models performs best and choose this as the representative pass model.

After the model has been constructed and chosen, we look into potential relationships with traditional statistics. We start with assessing individual passes and look at the distance a pass travels and find that on average, longer passes get rated higher than short passes. We also look into the relationship between the position of a player on the field and the pass value but it seems there is no relationship there. Then we continue and look into statistics about a team in a match, such as the pass accuracy, the ball possession and the number of the goals. For the first two of these, we find a very small negative relationship, which can be explained due to the fact we only consider successful passes, and for the number of goals it is a small positive relationship. Furthermore, we look into the average pass value for the different subtypes in the event data. We find a promising result that the 'standard' successful pass is rated the lowest, and the subtypes of greater impact, such as key passes and assists, are rated the highest.

Finally, we look into the potential practical applications that follow from the pass value model. For this we suggest two things: first, we consider alternative passes. Because it can happen that a does not get a high score, while it was the best option in that situation. For this, we construct a way to look at the alternatives and see how many better alternatives there were, and the difference between the pass value of the actual pass and the pass value of the best alternative. Secondly, we look at a way to rate players based on their pass value, which can be used by trainers to see who performs well and compare players with each other. We also compare this rating to real ratings that are available, and we see that if we look at the different types of player separately, there is a positive correlation between the rankings.

1 Introduction

In football, winning is the ultimate goal, and scoring goals is key to achieving it. That is why goals and assists are often used as important measures of the impact of a player. However, these events occur only a few times per match.

The event that occurs most frequently during a football match, according to our data on average every 6 seconds, is a pass. While most of the times not leading directly to a goal, passes are the building blocks of every successful offensive move, which eventually might result in goals. Passes create space and enable players to advance the ball toward the opponent's goal.

Despite the importance of passes, it is usually judged simply as successful or not, without considering the difficulty and risk of the pass. A pass can be influenced by many factors, like the amount of opponents nearby, where on the field it is happening and the distance the ball needs to travel. Traditional statistics do not capture these factors, so we will explore a method to value passes beyond the simple binary measure of success.

To do so, we first dive into related research, because this is not the first time that a pass value has been constructed. Authors we consider have used different approaches for the pass value: while one only uses event data and no tracking data, another has a more physics-based approach. Since we execute this research at the request of FC Eindhoven, we have discussed with their experts what they find important in a pass, such that the pass value model represents the importance they are looking for.

The components that they think are important are Expected Threat, Packing and Pitch Control. Expected Threat measures the probability that a goal will be made within five actions from the starting position, and with this metric we can evaluate passes during buildup play. Packing also contributes to the impact of the pass by measuring the number of opponents that are bypassed during a pass. Lastly, Pitch Control is used to assess the risk of the pass by measuring the probability that your team will maintain control of the ball if it arrives at a particular location on the field. By combining these components we will develop two pass value models. We will use expert data as our 'real' pass values to estimate our parameters, and we will select our model based on evaluation metrics.

When we have found a representative pass value model, we aim to understand if there is a relationship between the pass value and several factors such as the pass completion rate, pass accuracy, ball possession and more. Furthermore, we compare the average pass value of subtypes available in event data: a pass that is labeled in the data as a key pass should probably be ranked well by our model.

Last but not least, we also explore real world applications of the pass value model. What can a football club do with these insights to improve on their play? First, we consider alternative passes: if your pass got a certain value, was that the best possible move from that position, or would another pass have resulted in a better value? Finally, we consider rating players based on their average pass value during their time played in a match.

In conclusion, we aim to make passing analysis in football more accessible and insightful. By diving into the details of passing dynamics, we hope to find valuable insights that can enhance the post-match analysis of FC Eindhoven. To get there, we want to answer the following research questions:

1. How can the expert-identified factors be incorporated into a representative pass value model?
2. What is the relationship between our pass value model and various statistics regarding passes and matches?
3. How can this model be applied in practice to optimize team performance?

1.1 Pipple and FC Eindhoven

Pipple is a data science consultancy located in Eindhoven, which specializes in using mathematics and AI to create positive impacts across various sectors. They have clients in different sectors such as supply chain and logistics, high tech, financial support and healthcare. They help organizations make sense of their data by using advanced techniques to uncover valuable insights and develop practical solutions.

A small professional football club, FC Eindhoven, approached Pipple for assistance with their data. They have a lot of interesting data available, like tracking data (player positions) and event data (passes, shots, goals etc.) and they want to get more insights and value from this information. They are interested in using this data to improve their own play, but they do not have the internal expertise and knowledge related to data science.

In this collaboration, with FC Eindhoven's knowledge of football-related questions and their specific objectives, and the data science knowledge from Pipple, we have a good foundation to conduct this research. Together we aim to extract valuable insights from FC Eindhoven's data to enhance the club's play and decision-making processes.

1.2 Outline of the report

The structure of the rest of the thesis is as follows: we will kick-off in Chapter 2 with introducing relevant concepts that will be used later on the thesis, such as Expected Threat, Packing and Pitch Control. In addition, we also consider related research to valuing passes in football. We continue with Chapter 3, where we dive into the datasets that have been provided by FC Eindhoven. The relevant data we will consider is tracking data and event data, combined with metadata which is relevant to understand the data and make it readable and usable. Also, we dive into expert data that will be used for the parameter selection of the pass value model. Chapter 4 will expand on the methodology of this research. Starting off with elaborating on how we find the start and end coordinates of a pass by combining the data. We will explain our approach on how to extract the values of Expected Threat, Packing and Pitch Control from the data. Then we introduce the potential pass models and techniques we want to use to estimate parameters and choose the best performing model. Afterwards we will look at the relationship between certain variables, such as the pass accuracy or ball possession and the constructed pass value. We conclude this chapter with some practical applications of the pass value model, such as considering alternative passes or rating players by their average pass value. In Chapter 5, we will discuss the results of the proposed methodology from the previous chapter. We will conclude the thesis with Chapter 6 in which we will summarize, draw conclusions and evaluate our research. We will end with recommendations for future research.

2 Related Work

In the upcoming chapter, we will explore other articles, papers and blogs that discuss football analytics, and more specifically techniques that are relevant for evaluating passes. There already exist some papers that value passes, which we will introduce here. As mentioned in the introduction, valuing passes is not a new concept. However, the techniques we choose are based on the interests of the football club, and these methods have not yet been used together in valuing passes.

2.1 Expected Goals

The first metric that we look into is one that is commonly used, and that is Expected Goals (xG). Multiple publications describe it along the liens of 'one of the first advanced metrics to become widely known among general football fans' (Whitmore 2023). xG measures the probability that a goal will be made from a certain shot-position. This is done by using a logistic regression, based on at least the distance from the goal and the angle from the goal, but depending on the specific xG metric there are also other factors involved. This can be for example the match state (current score), shot types of play, body part, or shot technique.

It is not clear when and by whom the xG metric has been introduced. There are multiple sources online that say that the name 'Expected Goals' has been mentioned for the first time by Vic Barnett and Sarah Hilditch in 1993 (Barnett and Hilditch 1993). In 2004, something that sounds like expected goals was introduced, but not yet with that name, by Richard Pollard, Jake Ensum and Samuel Taylor (Pollard et al. 2004). They use a logistic regression with variables such as the distance from the goal, angle from the goal and three other factors. This is similar to the components that nowadays are incorporated in the definition of xG. A few years later, Sam Green wrote about it in an article (Green 2012), introducing the xG metric, which is the first time the metric and terminology are used together in an official paper. Today, xG is used a lot in football analytics. The drawback of this metric however is that it is specifically meant to measure a good shooting position and thus only can be applied on shots on goals, not on passes in general. Therefore, it will not be used in this thesis, but it has been the start for another metric which we will use.

2.2 Expected Threat

Since we want to value all passes and not only those in scoring positions, we will use a different metric than xG. As an alternative, we can consider Expected Threat (xT). xT assigns a threat value to every location on the field. xT was introduced in 2011, but not yet with that name. Sarah Rudd (Rudd 2011) came up with the idea of using Markov chains to determine how to value players that helped scoring the goal, so that the credit of a goal can be shared instead of going only to the player that scored the goal. Markov chains model the probable outcomes over multiple iterations by considering the transition probabilities between states. To apply this to football, Rudd considers the following states:

- 2 absorbing states: 'Goal' and 'End of Possession'
- 7 set pieces (penalties, corners etc.)
- 30 states defined by zonal location and defensive state

With these states, she determined a transition matrix by calculating the probability of moving from state S_a to S_b for all combinations of the 39 states. Then by multiplying the matrix with itself for n times, this resulting matrix gives the probability of ending in a certain state after n iterations.

In 2018, Karun Singh (Singh 2018) published a blog post about xT, where it is defined as a way to assign a threat value to every location on the pitch. This is done by considering two options: a player can shoot, or move the ball by passing or dribbling. For both options, the expected payoff is considered. This is combined in a formula which is iteratively evaluated, resulting in the probability of scoring a goal after n iterations, similar to the idea of Sarah Rudd.

2.3 Packing

When you are in possession of the ball, the fewer opponents there are between you and the goal, the more likely you are in a scoring position. To measure this, an approach named Packing was introduced, by former German football players Stefan Reinartz and Jens Hegeler (Biermann 2015), with their company IMPECT, that translates football into data. Packing is the number of opponents that are bypassed during a pass (or dribble). The opponents that are outplayed during a pass are then 'packed'. There exist multiple variations of how to determine the packing score. One version considers the number of opponents bypassed on the field length of the field (x-axis). Another variant considers the distance to the goal, which means it first checks how many opponents were closer to the goal than the ball, and after the pass this will be checked again. The difference in this number of opponents is the packing score.

2.4 Pitch Control

The mentioned metrics are all about the potential impact of a pass towards scoring, but we also want to consider the probability that a pass will succeed. Without assessing this risk, we might end up suggesting impractical strategies such as only advocating for passes from our own goal to the opponent's end. To address this, we consider the concept of Pitch Control, introduced by William Spearman (Spearman 2016). Pitch Control is defined at a given location as the probability that a team (or a player) will get possession of the ball if the ball moves directly to that location, considering multiple factors such as the speed and acceleration of the players, speed of the ball, the initial reaction time and the trajectory of the player.

In another paper (Spearman 2018), Spearman introduces the Potential Pitch Control Field (PPCF), which enhances the original Pitch Control model by using the arrival time of the ball in the model. This adjustment means that the PPCF model accounts for the time players have to reach the ball. This model gives us a more realistic understanding of Pitch Control. In Section 4.1.4 this model will be explained in more detail. In Figure 1, we show an example of Pitch Control. The color of the field represents the team that controls that part of the pitch, and the intensity of the color represents the amount of control the team has. For the white areas it means that the amount of control for both teams is roughly the same.

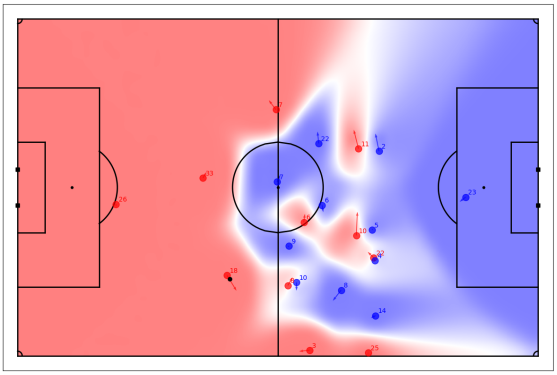


Figure 1: Example of the Pitch Control model: the colored dots are the players of the red and blue teams. The arrow connected to the players is pointed in the direction the players is moving and the size is based on their velocity. The black dot close to the red player 18, left from the centerline, is the ball. The color intensity on the pitch represents the control the teams have on all parts of the field.

2.5 Valuing passes

In this section, we will explore relevant papers that discuss various methods for valuing passes. All of these papers use different approaches and we use these to gain more insight into the subject.

2.5.1 Valuing passes with event data

The main goal of the paper (Bransen et al. 2019) is to measure the contribution of football players to the outcome of the match by valuing their passes. Their approach to value passes follows four steps. To start with, they split the matches into possession sequences. Possession sequences are defined as a series of events such that, for the whole series, the same team maintains possession of the ball.

The next step is to label the possession sequences and passes for its outcome. A pass gets the same label as the corresponding possession sequences gets. If a sequence does not lead to a shot on goal, the possession sequence gets the label 0. If it does lead to a shot on goal, the label of the possession sequence will be the probability of the scoring opportunity resulting in a goal. This probability is determined through the expected goals model.

After this labeling has been done, the similarities between passes is computed. For this, a distance function between two passes is defined that includes the characteristics and the circumstances under which the pass is executed. Then, the similarity between passes is defined. The last step is to determine the value of a pass. This is defined as the expected added reward, which is calculated as the expected end reward minus the expected start reward of the pass.

To determine the expected start and end rewards, the pitch is divided into a grid of cells. Passes are assigned to the cell where the end location of the pass is. For the expected start reward we consider the cell where the pass starts. Then, consider all

passes that start in that particular cell. We determine the expected start reward then as the average label of the passes that start in that cell. The expected end reward for unsuccessful passes is equal to zero. For successful passes, it is determined as a weighted average of the pass labels in that cell, with as weights the similarity function. In this paper, the start and end coordinates of passes are incorporated in the event data. They do not use tracking data, which is a significant difference compared to our research.

2.5.2 Physics-based approach for pass probabilities

In this paper (Spearman et al. 2017), they introduce a model for ball control in football, based on the concepts time-to-intercept and time-to-control. Time-to-intercept refers to the time it takes for a player to reach the ball and time-to-control represents the duration it takes for a player to control the ball, once it is reached. This model is utilized to assess the probability of a pass succeeding. The goal is to make a predictive model that is only based on information from the start of the pass. They determine a physics-approached way for an estimation of the trajectory of the ball as well as the time-to-intercept and the time-to-control. One of the applications of this model in their paper is to give a passing value. Provided that there is a function $f(x)$ which quantifies the value of a specific situation, they compute the value of a pass with the probability whether or not that pass succeeds. The pass value is computed in the following way: they consider the game states 'successful' and 'failed', and compute the probability that the pass succeeds times the function $f(x_{suc})$. From this, the multiplication of the probability that a pass fails times $f(x_{fail})$ is subtracted, i.e. $\mathbb{P}(successful) * f(x_{suc}) - \mathbb{P}(failed) * f(x_{fail})$.

2.5.3 Expected passes

The work of (Anzer and Bauer 2022) computes the probability of any pass to be completed. First, the event and tracking data are synchronized with each other, meaning that the location and exact timing of pass events from event data are matched with tracking data. Then, the problem is considered that for unsuccessful passes, the intended receiver is unknown. This is relevant to determine the difficulty of a pass, so they want to estimate this intended receiver. First, a model is used to determine the potential locations of all players within a specified time frame. Then, they combine this with the physics-based ball trajectory model from (Spearman et al. 2017) to predict the trajectory of the ball. Both steps combined gives a prediction for the intended receiver. After this, a machine learning model is trained to estimate the probability of a pass. This is based on the information in the step before, as well as a very extensive list of 25 features. Lastly, another model is constructed to estimate the likelihood that a pass is blocked.

2.5.4 Pass quality in football

In the following paper (Felices 2023) they also attempt to rate passes. They propose an alternative that takes into account the risk and the reward of a pass. So first they start with defining the pass risk. This is defined as 'the likelihood of successfully completing a pass given a player's possession of the ball and the situation they are in.' This risk is assessed by a standard supervised learning pipeline, where they train a classifier to return a probability based on input features. Three baselines are introduced to compare the risk of a pass: Naive, Ball-Information and Tracking/Feature-crafted. After the pass risk, they continue with the pass reward. This is defined as 'the likelihood that a pass made in a given situation, with a player in possession of the ball and the ability to pass, will result in a shot within the next 10 seconds'. This reward is assessed in a similar way as the risk.

3 Data Description

For this research we were provided with access to football data of the matches of FC Eindhoven in the Dutch First Division. We can access this data via the platform SciSports (SciSports 2013). For every match, we have access to physical reports, tracking data, metadata, event data and the video of the match. For our research we will not use the physical reports. We will mainly focus on the tracking and event data, and the metadata is relevant to structure the tracking data. In this chapter we will start with a description of the datasets. Then we explain the preprocessing we executed. Afterwards, we will elaborate on how we make the connection between the different datasets we use. We will finish this chapter with describing the expert data we have collected for the parameter estimation of the pass value model.

3.1 Tracking data

The first dataset that we will describe is the tracking data. Tracking data consists of the coordinates of every player on the field (for both teams) and the coordinates of the ball. In our data, we have a frame rate of 25, which means we have 25 frames per second, resulting in approximately 175.000 frames per match, accounting for halftime and injury time. The tracking data file is the 'raw data file', where each row in this file represents a frame, containing the coordinates for all players and the ball. For the players, we have the x- and y-coordinates. For the ball, we have x-, y- and z-coordinates and also binary values for 'ball in air' and 'ball alive'. Ball in air is 1 if the ball is in the air, and 0 otherwise. Ball alive is 1 if the match is active, and 0 if not (for example, when the ball crossed the line and has to be thrown-in again). The tracking data does not have column names, which makes it not yet a nicely structured or readable data file. To address this, we utilize metadata to transform the raw data into a well-structured data file. This involves adding column names, splitting the data set for the home and away team and adjusting the coordinates to be able to easily use open source code for the Pitch Control model (Shaw 2020). We will first discuss the metadata, and then elaborate on the preprocessing we did.

3.1.1 Metadata

The metadata of a match consists of some relevant information that is needed for the data analysis. There are many different parameters in this data, but the relevant parameters we will use, are listed below:

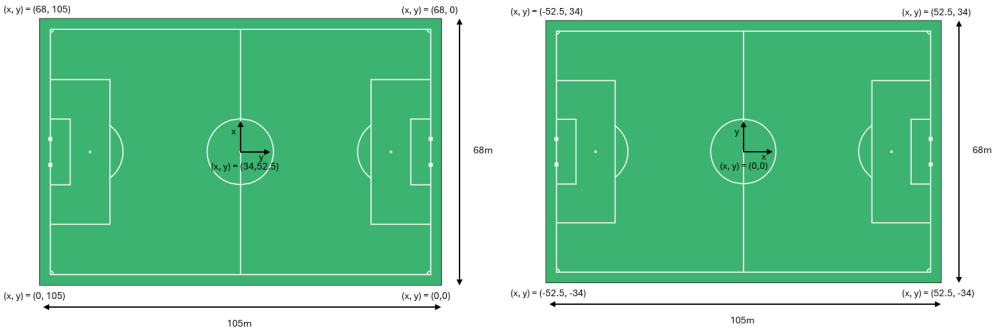
Parameter	**Explanation**
Frame rate	The number of frames per second.
Field size	The length and width of the field, and the coordinates of the centre of the pitch.
Start & End frames	For the whole match, as well as the first and second half, the start and end frame number.
Team	For both teams, a Team-ID and if they are the Home or Away team.
Player	For all players, a Player-ID, shirt number, and the first and last frame the player is part of the line-up.

The parameters mentioned above can be used to make the tracking data file into a better structured data file.

3.1.2 Preprocessing

The first step to improve our tracking data format is to obtain column titles. We discovered that the order of the coordinates in the rows of the raw data file, is the same as the order of the players in the metadata file. We extract the jersey numbers from the metadata, and use these as the column titles. Then we split the tracking data into two files, one for the home team and one for the away team. Additionally, we split the x- and y-coordinates for each player into separate columns to facilitate plotting the positions on the field. We also add a period number to each row, which indicates in which half of the match the game is. We obtain this period numbers from the start and end frames of the metadata.

Furthermore, we adjust the x- and y-coordinates of the match. This is because we will make use of open source code for visualizations and the Pitch Control model (Shaw 2020). The raw tracking data initially has $(34; 52, 5)$ as the center of the pitch, with the x- and y-coordinates reversed compared to the desired format. To ensure compatibility with the existing code, we will transform the coordinates so that the center of the pitch becomes $(0, 0)$, which includes swapping the coordinates and adjusting the values. Making these adjustments is simpler than rewriting the open source code ourselves.
The difference in the coordinate system before and after the transformation are visualized in Figure 2 below.



(a) Coordinates before

(b) Coordinates after

Figure 2: Coordinate system

3.2 Event data

The second dataset we will use is the event data, which consists of certain events that occur during a football match. For these events, data is collected such as the start and end time of the event, what type of event it is (e.g., pass, foul, goal) and which team and player executed the event. The event data of a match consists of the following variables:

Variable	Explanation
Code id	Unique ID for the event.
Period id	Period in which the event took place.
Timestamp	Start timestamp for the event (in seconds).
End timestamp	End timestamp for the event (in seconds).
Code	Team executing the event, type and subtype of the event.
Player	Player executing the event, including shirt number.

During the research, we discovered that in the Dutch First Division football where our club plays in, there are different structures for the event data, that have to be handled differently. Most of the matches have the structure we have discussed above. The matches that have a different structure will be out of the scope of our research, due to the extra time the data processing would cost. Fortunately, we have 26 matches with the structure as discussed above, so we will assume this will not severely impact the quality of this research.

3.2.1 Preprocessing

The column 'code' of the event data contains three types of information: which team it is about, which type of event and which subtype of event. To better organize the data, we split this column into three separate columns. Additionally, since many events appear twice in the dataset (once at the team level and once at the player level), we clean the data so that every event appears in the dataset only once. We also split the player column such that the jersey number is in a separate column; this will be useful later on, since the tracking data has the jersey numbers as column titles.

Since we now have the types and subtypes of events in separate columns, we can consider the different types that are present. Considering the passes, there are two type of events: 'Pass' and 'Pass (Successful)'. For successful passes, both these events exist at the same timestamps, while for unsuccessful passes, we only have the 'Pass' event.

A detail to point out is that especially for passes, but also for some other events, the timestamps are not exactly the start and end of the event. According to SciSports (SciSports 2024), the start time for passes is 3 seconds before the actual start and the end is 3 seconds after the actual end. However, in practice, this approximation may not always be accurate. While examining the data, we already noticed some passes for which the time it took was smaller than 6 seconds, which contradicts the statement by SciSports. This indicates that the actual start and end times of events may vary and require further investigation. In Section 4.1.1 we will address this issue by adjusting these timestamps to make them reflective of the actual start and end times of the events.

3.3 Linking tracking and event data

A crucial step for the data analysis is the connection between the tracking- and event datasets. The event data does not contain the locations of players, and the tracking data does not contain the events happening. To execute our research, we want to be able to have this data combined. In the tracking data, there are frame numbers included. The event data consist of timestamps in seconds. In the metadata, we have the start frame number of the match. Starting from there, we take time steps of size $\frac{1}{\text{frame rate}}$ and add them as timestamp. In our case with 25 frames per second, we take time steps of 0.04 seconds. With adding these timestamps to the tracking data, we are able to link the two datasets.

3.4 Expert data

To determine the parameters and select the best pass value model to get a representative model for FC Eindhoven, we want to use 'real' pass values. However, valuing passes is not yet common practice, so we do not have access to real values. To address this, we have asked experts to rate passes. We have given them schematic overviews of the passes, showing the start situation for all players, and an arrow from the start player to the end coordinates of the pass, along with a description of who was the receiver of the pass. An example of this is shown in Figure 3. In total, 34 passes selected from 2 different matches have been rated on a scale of 1 to 5.

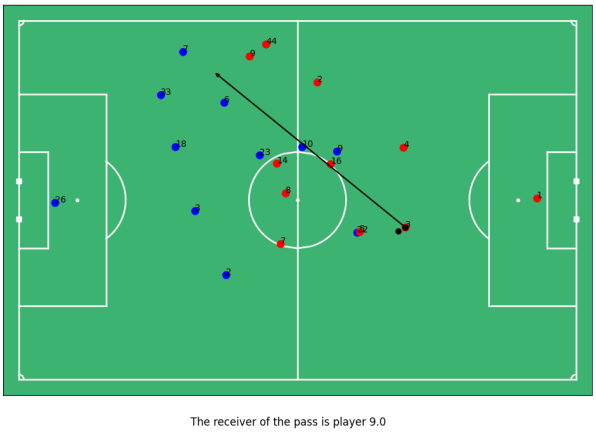


Figure 3: Example of schematic overview of a pass

The initial plan was for these five experts to be from the staff of FC Eindhoven, since that would be the best representation of how they value passes, which would be best if the model would be used by them later on. However, because of circumstances beyond our control, this could not be executed. To solve this problem, we have asked five football-enthusiast employees from Pipple to be our experts.

Given this data, we consider the pairwise correlation between the scores of the different experts. These coefficients are Pearson's correlation coefficients (Chattamvelli 2024), which is determined with the following formula:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \tag{1}$$

In this formula, $x \in X$ are the values of variable X (for example, Expert 1), and $y \in Y$ are the values of variable Y (for example, Expert 2). n is the number of observations, in our case 34.

Our pairwise correlations are given in Table 1. Furthermore, in Appendix A.1, all expert ratings are shown. The majority of the correlation coefficients are positive, which indicates that there is moderate agreement in ratings of the experts. For the interpretation of the number of the correlation, to name them as weak. moderate or strong, there are several approaches introduced but there is not one clear definition of how to interpret them (Schober et al. 2018). In the example in this article, they state that when the

coefficient is above 0.7, it is a strong correlation. In other articles, that is sometimes 0.75 or 0.8 or even higher. Most of our coefficients are below 0.7, and all are below 0.75. There could be a discussion about if the coefficient of 0.73 is strong, but we choose the value of 0.7 from the article as the boundary and state that none of our correlation coefficients are strong. Also, a coefficient that stands out is the correlation between Expert 3 and 5, which is almost 0, which indicates there is no relationship between their scores, according to the Pearson correlation metric.

	Expert 1	**Expert 2**	**Expert 3**	**Expert 4**	**Expert 5**
Expert 1	1.0	0.39	0.28	0.73	0.67
Expert 2		1.0	0.19	0.4	0.36
Expert 3			1.0	0.49	-1,04 $\times 10^{-17}$
Expert 4				1.0	0.58
Expert 5					1.0

Table 1: Expert ratings correlation

From all of this we can draw the conclusion that valuing passes is very subjective. Because of the difference in scores and noticing in some cases some specific outliers of one expert, we will take the median of the expert values to obtain the value per pass that we will use as the 'real' value. We prefer to use the median instead of the mean, as with the mean the outliers have a considerable influence on the score (Khorana et al. 2022).

4 Methodology

In this chapter we will begin by explaining the different components that we will use in the construction of the pass value model, followed by the model itself. We will then elaborate on our approach on how to estimate the parameters and selecting the appropriate pass value model. Furthermore we will explain the approach we will use to explore any potential relationship between the pass value from our model and certain statistics. We will conclude this chapter with our ideas on how to apply the pass value in practice. The results that arise from this will be discussed in Chapter 5.

4.1 Valuing passes

In this section, we introduce a new metric to value passes in a football match. When considering passes, there are numerous factors that could be taken into account. Since we want this metric to be useful for FC Eindhoven, we have included the suggestions of the experts of the football club to come to the components we will use. This will be Expected Threat, Packing and Pitch Control. We will elaborate on these components in more detail in this section. Before we can determine the actual values for these components, we first encountered another problem, being that in the event data, start and end coordinates of passes are not available.

4.1.1 Start and end coordinates

For our passes, we want to know the start and end coordinates of the pass, because we will need this later on for the metrics we will use. For this, we need to use a combination of our tracking and event data. There are two difficulties we have to tackle:

- Extract who the receiver of the pass is.
- Extract the correct x- and y-coordinates of the start and end of the pass.

We will start with the first bullet: how do we know who the receiver of the pass is? In our research we will only consider successful passes. This is because for unsuccessful passes, it is very difficult to determine the potential receiver of the ball. This is definitely interesting to look into, but for now out of the scope of this research.

Considering the successful passes, it is a bit easier but not trivial. The event data is after preprocessing better structured, but still not perfect. For example, some events occur twice because they are labeled in two different categories (there can be two exactly the same 'Pass (Successful)' events, but one has the subtype 'Open Play', which all passes have, and one 'Assist'). This is not a big problem, but is something we have to tackle while looking for the receiver.

To find the receiver, we iterate through the rows of the event data. We aim to locate the next row where the jersey number is different from the sender's (and not an empty value), the team is the same as the sender and the Type is not 'Physical'. This is because 'Physical' is about run types and has event types such as 'Centre To Flank Run' or 'Deep Run', but this does not necessarily involve someone who is in possession of the ball but can also be about a player that is unrelated to the events happening around the same time. We continue searching until we find such a row within the time frame of the event. This can also be considered in pseudo code:

Algorithm 1 Finding the receiver of the pass

```
start_row ← index of the starting event
next_row ← start_row + 1
while next_row is within the dataset do
    if Jersey number at next_row = Jersey number of start_row
     or Jersey number at next_row is empty
     or TeamType at next_row is different from TeamType of start_row
     or Type at next_row is 'Physical' then
        next_row += 1
    else
        result_number ← Number at next_row
        Assign result_number to the 'Receiver' column of the event
    end if
end while
if next_row exceeds the number of rows in the dataset then
    Set 'Receiver' at start_row to None
end if
```

After determining the receiver, the second part is to extract the coordinates of the pass from the tracking data. Because the timestamps of the event data are not precise, as described in the data section, we want to consider another strategy. If you want to execute a pass, you have to be close to the ball. Because of this, we have defined the function `find_minimum` which needs as input a dataset and a distance column of the player you want to consider. Then as output it returns the minimum distance, and the index of the row in the dataset where the distance between the player and the ball is the smallest.

Algorithm 2 `find_minimum`(data, distance_column)

```
min_distance ← minimum distance in distance_column
min_distance_rows ← rows in data with distance equal to min_distance
if min_distance_rows is not empty then
    min_distance_row ← the first row in min_distance_rows
    return min_distance, min_distance_row
else
    return None, None
end if
```

To determine the actual start and end coordinates, we will use this function `find_minimum`. For the start of the pass, we search for the time between the start and end time of the pass where the sender of the pass is closest to the ball. This means that as input for `find_minimum`, we use a subset of the tracking data, with only the data that is between the start and end time of the event, and the distance column of the player that executes the pass. With this, we find the start coordinate and time of the pass.
For the end coordinates we do this in a similar way. For this, use as input data again a subset of the tracking data with the data between the start and end time of the pass, but we use the start time we got as a result of the start coordinates. For the distance column, we use the distance column of the receiver of the pass which we have found in Algorithm 1. We also describe this process in pseudo code:

Algorithm 3 Extracting coordinates

1: *distance_column_send* ← column indicating distance of the sender of the pass to the ball
2: *distance_column_rec* ← column indicating distance of the receiver of the pass to the ball
3: *data* ← tracking data from event start timestamp to event end timestamp
4: *min_distance*, *min_distance_row* ← find_minimum(*data*, *distance_column_send*)
5:
6: *start_coordinates_x* ← x-coordinate of the player at *min_distance_row*
7: *start_coordinates_y* ← y-coordinate of the player at *min_distance_row*
8: *result_time_start* ← timestamp at *min_distance_row*
9:
10: Update the event with the found start coordinates and timestamp
11:
12: *data* ← tracking data from result_time_start to event end timestamp
13: *min_distance*, *min_distance_row* ← find_minimum(*data*, *distance_column_rec*)
14:
15: *end_coordinates_x* ← x-coordinate of the player at *min_distance_row*
16: *end_coordinates_y* ← y-coordinate of the player at *min_distance_row*
17: *result_time_end* ← timestamp at *min_distance_row*
18:
19: Update the event with the found end coordinates and timestamp

4.1.2 Expected Threat

As introduced in Section 2.2, Karun Singh (Singh 2018) introduced Expected Threat, defined as a way to assign a threat value to every location on the field. For this, two options are considered: a player can shoot, or move the ball by passing or dribbling. For both options, we have an expected payoff. These will be combined into one formula later in this section. To compute this Expected Threat, we divide a football field into different zones, in an $a \times b$ grid. For every zone (x, y) in this grid, there are a few aspects that we will take into consideration:

- Move probability $m_{x,y}$: when a player is in possession of the ball in zone (x, y), what is the probability that a player chooses to move the ball in their next action, i.e. the player performs a pass or dribble.
- Shoot probability $s_{x,y}$: when a player is in possession of the ball in zone (x, y), what is the probability that a player chooses to shoot the ball in their next action.
- Move transition matrix $T_{x,y}$: when a player moves from zone (x, y), what is for each of the other zones the probability that a player will move to this specific zone.
- Goal probability $g_{x,y}$: what is the probability that when a player shoots from zone (x, y), the shot results in a goal.

In this model we consider that players will always move or shoot, so $m_{x,y} + s_{x,y} = 1$. For Expected Threat, the following equation is defined (Singh 2018):

$$\mathrm{xT}_{x,y} = (s_{x,y} \cdot g_{x,y}) + (m_{x,y} \cdot \sum_{z=1}^{a} \sum_{w=1}^{b} T_{(x,y)\to(z,w)} \cdot \mathrm{xT}_{z,w}) \tag{2}$$

We will explain the formula, breaking it down step by step. The formula consists of two parts:

- $s_{x,y} \cdot g_{x,y}$
- $m_{x,y} \cdot \sum_{z=1}^{a} \sum_{w=1}^{b} T_{(x,y)\to(z,w)} \cdot \mathrm{xT}_{z,w}$

We start with the first bullet. This part of the equation values how good it is to shoot from position (x, y). It consists of the probability that a player chooses to shoot the ball ($s_{x,y}$) and the probability that the shot results in a goal ($g_{x,y}$), which essentially is the expected payoff of a shot. Then we have the second part of the equation. ($m_{x,y}$) is the probability that a player chooses to move the ball. But if you choose to move the ball, you also have to make the decision to which zone on the field you will move it. Say that the reward for a zone (z, w) is $\mathrm{xT}_{z,w}$. To determine the expected payoff for this zone, we have to consider the reward and the probability of moving to that zone. For that, we use the transition matrix $T_{x,y}$. So for a specific zone (z, w) the probability of moving to that zone is $T_{(x,y)\to(z,w)}$. If we combine these two components we get the expected payoff for zone (z, w). If we want to consider the total expected payoff we have to consider all different zones. Then we come to $\sum_{z=1}^{a} \sum_{w=1}^{b} T_{(x,y)\to(z,w)} \cdot \mathrm{xT}_{z,w}$. To determine this value, observe that you already need to know the xT value for all zones. To work with this we will iteratively evaluate the formula until we reach convergence. We start with $\mathrm{xT} = 0$ for all zones. If you look at the formula in that case, we are only left with $s_{x,y} \cdot g_{x,y}$, which we can extract from our data. This means, that in this first iteration the only option is to shoot the ball (you can compare this to an expected goals model). In the following iteration, we also consider the option to move before shooting. Essentially this means that we consider the probability of scoring within the next two actions. If we extend this, we can say that after n iterations, xT is the probability of scoring in the next n actions.

In practice using the above, we have chosen for a 12×8 grid, and $n = 5$. Karun Singh uses in his blog a 16x12 grid (Singh 2018), but mentions that you can choose a different resolution based on the amount of data you have. He looked at a whole season of matches, while we only consider matches of one club within a season. Because of this, we have chosen a smaller grid. The choice of $n = 5$ is based on the blog of Karun Singh as well. It can be interesting for future research to look into the influence of the choice of n

In this research, we will use xT in the following way. For a pass, we consider in which cell of the grid the start coordinates lie and the same for the end coordinates. We then extract from the xT grid the xT value for the start and end, i.e., $\mathrm{xT}_{\mathrm{start}}$ and $\mathrm{xT}_{\mathrm{end}}$. The xT value that we will use in the pass value is defined as the difference between the end and the start of the pass, i.e., $\mathrm{xT}_{\mathrm{pass}} = \mathrm{xT}_{\mathrm{end}} - \mathrm{xT}_{\mathrm{start}}$, since we want to increase the xT as much as possible with a pass. An easily understandable interpretation of this value is that a pass's worth is measured as the change of percentage points it brings to the chance of the team scoring within the next 5 actions. For example, if for a pass, $\mathrm{xT}_{\mathrm{start}} = 0.05$ and $\mathrm{xT}_{\mathrm{end}} = 0.15$, then the worth of this pass is 0.1, which indicates that the chance of scoring within the next 5 actions is increased by 10 percentage points by this pass.

To use the xT value in valuing passes, we normalize the scores to a range between 0 and 1 to ensure comparability with other variables. Normalization prevents xT from disproportionately influencing the model due to its original scale, allowing for balanced and fair contributions from all factors.

4.1.3 Packing

As discussed before in Section 2.3, there are several definitions of Packing. In our method we will use the version that considers the distance to the goal of the ball and the opponents. This is because we believe this provides a more direct measure of the threat posed by the opposing team in comparison with considering the x-coordinates. To determine the Packing score, for both the start- and end situation, the number of opponents that are closer to the goal compared to the ball have to be determined, and the difference between the end and start situation is the Packing score.

Algorithm 4 Determine Packing score

```
for each pass event in events do
    Determine the team executing the pass
    distance_ball ← the distance from the ball to the goal
    Packing_score_start = 0
    for each opponent in opposing team do
        distance_opponent ← the distance between the opponent and the goal
        if distance_opponent < distance_ball then
            Packing_start += 1
        end if
    end for
    Packing_score_end = 0
    for each opponent in opposing team do
        distance_opponent ← the distance between the opponent and the goal
        if distance_opponent < distance_ball then
            Packing_end += 1
        end if
    end for
    Packing = Packing_start − Packing_end
end for
```

To use this Packing score in valuing passes, we will normalize the values to scores between 0 and 1. This normalization ensures comparability with other variables, as explained previously. We will use the notation P for Packing later on in the pass value.

4.1.4 Pitch Control

Lastly, we consider the Pitch Control model. For this we consider the Potential Pitch Control Field (PPCF) model of William Spearman (Spearman 2018).
In the paper Beyond Expected Goals a differential equation is introduced to determine the control probability of the ball for each player j at location r at time t. This probability is derived as the probability that a player will be able to intercept the ball at location r, multiplied by the probability that none of the other players will control the ball at the same location.

First, we consider the probability that player j reaches location r at time t within time frame T, with an uncertainty in player arrival time of s, which will be stated as $f_j(t, \overrightarrow{r}, T|s)$. To determine this, first the expected intercept time $\tau_{exp}(t, \overrightarrow{r})$ will be computed. For this, we calculate the time it takes for player j to reach location $\overrightarrow{r}$ from their starting position $\overrightarrow{r_j}(t)$. This calculation assumes the player begins with an initial velocity $\overrightarrow{v_j}(t)$ and accelerates at a constant rate until reaching a maximum speed. The

probability is then given by:

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

Then we consider a parameter λ_i which we call the control rate, which means it is the time it takes for a player to control a ball when it has arrived to a certain location. This time is given by $1/\lambda_i$, so a higher value of λ_i means there is less time needed for a player to control the ball. In the paper, they make the difference between the control rate for attacking players (set A) and for defensive players (set B), since an attacking player would want to make a more accurate touch while a defending player is already content with just kicking away the ball. Because of this an extra parameter κ is introduced.

$$\lambda_i = \begin{cases} \lambda & \text{if } i \in A \\ \kappa\lambda & \text{if } i \in B \end{cases}$$

With the probability and the control rate, we can state the differential equation for the control probability in a formula in the following way:

$$\frac{d\text{PPCF}_j}{dT}(t, \overrightarrow{r}, T|s, \lambda_j) = (1 - \sum_k \text{PPCF}_k(t, \overrightarrow{r}, T|s, \lambda_j)) f_j(t, \overrightarrow{r}, T|s)\lambda_j \tag{4}$$

To apply the PPCF-model in practice, we use an open source implementation by Dr. Laurie Shaw (Shaw 2020). In this implementation they make some assumptions, which are the following:

- The ball has a constant speed of 15 m/s.
- A player has a maximum speed of 5 m/s .
- A player has a maximum acceleration of 7 m/s^2.
- A player takes the fastest possible path to the considered location at the pitch.
- A players has a initial reaction time of 0.7 seconds wherein the players continues their current trajectory and after this, they run directly to the target location at maximum speed.
- There is no difference in the control rate for attacking players and defensive players, so $\kappa = 1$

The open source implementation that we use determines the PPCF-value for the field divided in a 50x32 grid.

To use Pitch Control in a pass value, we have to consider for which part of the pass you want to calculate the PPCF-value: only the beginning and/or end coordinates, an area around the pass or the whole line of pass? Since we want the pass to succeed, and there is also a chance that the other team takes control of the ball halfway the pass, we are going to consider the average PPCF-value along the line of pass.

To determine the cells we need to consider in the line of pass, we will use Bresenham's line generation algorithm (Bresenham 1965). This algorithm is an efficient method which is used for drawing a straight line between two given points on a discrete grid. It is commonly used in computer graphics and image processing to provide a solution to the problem of determining which pixels to illuminate to create a close approximation

of a straight line between two points.

The algorithm works by incrementally plotting the pixels closest to the true path. At each step, it evaluates whether the next pixel to be plotted should be horizontally or vertically adjacent to the current pixel, based on the line's slope. By intelligently choosing the nearest pixel to the ideal line path, the algorithm guarantees that the line will traverse through the most appropriate grid cells. It is a highly efficient and accurate algorithm: unlike methods that evaluate every pixel, this algorithm minimizes the number of calculations required while ensuring it is a well-defined line. With the below pseudo code we show how we implemented the algorithm. The starting and end position are not the actual coordinates, but the grid coordinates that are used in the Pitch Control model.

Algorithm 5 Calculate passing line with Bresenham's line generation algorithm

function CALCULATE_PASSING_LINE(start_pos, end_pos)
  Get the starting position (x_0, y_0) and the end position (x_1, y_1)
  Calculate the change in x-coordinate: $dx = |x_1 - x_0|$
  Calculate the change in y-coordinate: $dy = |y_1 - y_0|$
  Determine the direction of movement in x-coordinate: $sx = 1$ if $x_0 < x_1$ else -1
  Determine the direction of movement in y-coordinate: $sy = 1$ if $y_0 < y_1$ else -1
  Initialize error: $err = dx - dy$
  Initialize an empty list for passing line: $passing\_line$
  while True **do**
    Add the current position (x_0, y_0) to the passing line
    if Reached the end position **then**
      break
    end if
    Calculate $e2 = 2 \times err$
    if $e2 > -dy$ **then**
      Adjust the error and move in x-direction: $err = err - dy$, $x_0 = x_0 + sx$
    end if
    if $e2 < dx$ **then**
      Adjust the error and move in y-direction: $err = err + dx$, $y_0 = y_0 + sy$
    end if
  end while
  return passing_line
end function

From this algorithm we get all the cells in which the line occurs. Then, to get for a pass a Pitch Control value to use in the pass value, we will determine the average Pitch Control over all cells of the passing line, which we will note as PC. Choosing the average Pitch Control reflects the dynamic nature of real match scenarios and ensures a realistic evaluation of pass quality while maintaining interpretability.

4.1.5 Pass value

We will elaborate on the potential pass value models we will consider. To do so, we first have to take into consideration that Expected Threat and Packing metrics often become negative when the ball is played towards your own goal. If we want to handle such cases, we should look into another way to construct a model compared to the positive situations, because otherwise passes that are not necessarily bad could receive negative scores. Therefore, we we will only consider passes where the distance to the opponent's goal is smaller at the end of the pass, compared to the beginning. This approach will likely address most cases of negative Expected Threat and Packing. After the pass value has been performed, we will also normalize the pass value to a score between 0 and 1 again, and translate this into values between 1 and 5, to get a more intuitive feeling.

We will look into two different models. Both models are a linear combination of the different metrics. The choice of a linear combination offers simplicity and interpretability, and it allows for clear weighting of each metric, reflecting their relative importance in determining pass values.

Model 1
The three considered metrics are metrics that the experts have mentioned as important metrics. We construct a pass value PV by combining the three metrics in a linear model.

$$\text{PV} = \alpha \cdot P + \beta \cdot \text{xT} + \gamma \cdot \text{PC} \tag{5}$$

where P is the Packing score, xT the Expected Threat score and PC the Pitch Control value. We want to normalize our value PV so we determine the maximum PV value and then calculate the normalized PV.

Model 2
While talking to experts at FC Eindhoven, we also noticed that when valuing a pass they mention different factors being important when the pass is in the last part of the field (i.e. closest to the opponents goal) compared to the first part of the field. In the beginning of the field it is most important to keep the ball in control and try to make as much distance as possible, while closer to the goal the distance is not as important, but the Packing and xT become more important. Also, the xT values vary more in the last part of field compared to the first part. For the first part of the field, we will consider a new component, namely the distance. We also normalize the distance of the pass by dividing it by longest pass of the match, and the result is variable D.
For the first $\frac{2}{3}$ of the field, we state the pass value as follows:

$$\text{PV} = \gamma_1 \cdot \text{PC} + \delta \cdot D \tag{6}$$

For the last $\frac{1}{3}$ of the field, as in model 1:

$$\text{PV} = \alpha \cdot P + \beta \cdot \text{xT} + \gamma_2 * \text{PC} \tag{7}$$

4.1.6 Parameter estimation and model selection

In the previous subsection we developed two potential pass value models. However, these models still contain parameters that require estimation and afterwards we want to select which model works best. To estimate the parameters, we need real pass values, but such data is not readily available. To tackle this problem, we have asked five experts to rate 34 passes. We have gathered this data and already did some analysis in Section 3.4.

To estimate the parameters of our pass value models, we use a regression approach. The regression will help us determine the weights for the different factors that are incorporated in the pass value.

There are several regressions techniques available, including linear regression, ridge regression and lasso regression (James et al. 2013). The independent variables are the different factors we use in the pass values. Considering the small number of independent variables (at most three), lasso regression might not be practical. This is because Lasso tends to produce overly sparse models by zeroing out coefficients, with can remove important variables. If there exists multicollinearity among the independent variables, which indicates that there is a high correlation between them, we will choose ridge regression, otherwise we will use a linear regression technique, and specifically the Ordinary Least Squares (OLS) method.

With the OLS method, we can get the coefficient vector with the following formulas. For linear regression:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{8}$$

In the case of ridge regression, it is similar to the OLS estimator but we get the simple ridge estimator:

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y \tag{9}$$

For both equations, X is the matrix of independent variables (P, xT and PC), which means that X has three columns. y is the vector of dependent variables PV. For the ridge regression, we have λ which is the regularization parameter, and I is the identity matrix.

The multicollinearity will be determined with the Variance Inflation Factor (VIF). The VIF assesses the correlation of one independent variable with a group of other variables. The VIF is determined in the following way for every independent variable X_i. We start by running an OLS regression, where X_i is regressed on all other independent variables. For instance, when $i = 1$ and there are three independent variables in total, this equation is:

$$X_1 = \alpha_0 + \alpha_2 X_2 + \alpha_3 X_3 + \epsilon \tag{10}$$

In this equation, α_0 is a constant, and we have the error term ϵ. Then, the VIF factor for X_1 is determined with the following formula:

$$VIF_1 = \frac{1}{1 - {R_1}^2}, \tag{11}$$

where ${R_1}^2$ is the coefficient of determination. We have real values of the independent variable $y_1, \ldots, y_n$ and modeled values from the OLS regression $f_1, \ldots, f_n$. Then for the coefficient of determination we have the following formula:

$${R_1}^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{12}$$

where $SS_{res} = \sum_i (y_i - f_i)^2$ and $SS_{tot} = \sum_i (y_i - \overline{y})^2$, with $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ the mean of the observed data.

For model 1 and the last third of the field in model 2, we will use the VIF. For the first $\frac{2}{3}$ of the field in model 2, since there are only two independent variables, we will check the correlation between these variables for the multicollinearity.

Once we have conducted the regression analysis and obtained the regression coefficients, we can compare the performance of our models against the evaluations of the experts. To do this, we will use the median value of the experts' ratings for each pass. We will assess the effectiveness of the models using (Root) Mean Squared Error ((R)MSE) and the Mean Absolute Error (MAE) as measures of predictive accuracy. We also look into the R^2, that has been introduced in the VIF definiton.
The MSE is determined in the following way:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{13}$$

In this equation, n is the sample size (in our case 34), y_i is the i^{th} expert value and $\hat{y}_i$ the corresponding predicted value from the model. For the RMSE, you take the square root of the MSE.
The MAE is calculated as follows:

$$\text{MAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \tag{14}$$

where the variables are the same as in the MSE.

The (R)MSE is valuable because it penalizes large errors more heavily, providing insight into the average size of errors in predicted pass values. Meanwhile, MAE is more robust to outliers and gives a direct measure of the average absolute difference between predicted and actual values (Hossain 2023). By considering both metrics, we aim to evaluate how accurately our models predict pass values compared to expert ratings, ensuring the reliability and practicality of our pass value models in real-world scenarios.

4.2 Relationship analysis

Now we have constructed and chosen a pass value model, we will explore if there are certain relationships between the pass values and some other statistics of a football match. First, we are going to consider some independent variables and execute correlation and regression analysis on these. Then, we are going to elaborate on some comparative analysis.

4.2.1 Correlation and regression analysis

We have two categories of independent variables we are going to consider:

1. Variables related to specific passes
2. Variables related to an entire match

We will begin by examining specific passes, considering two variables: pass distance and player position. These variables have been selected based on our hypothesis that there is a relationship between them and pass value, as well as their accessibility in our data sources. The player position data will be extracted from the FC Eindhoven website.

- Pass distance
  For each pass, we will categorize the length as short, medium, or long.
- Player position
  We consider the assigned player position of the sender: is it a keeper, a defender, a midfielder or an attacker?

In addition to individual passes, we will consider the relationship between the average pass value of a team in a certain match and three variables: pass accuracy, ball possession and the number of goals. These variables will be extracted from the website FotMob ("FotMob" 2024), since this is not contained in our metadata.

- Pass accuracy
  The percentage of accurate passes per team in each match.
- Ball possession
  The percentage of time each team spends in possession of the ball.
- Number of goals
  The total number of goals scored by each team in the match.

There are always more ideas for independent variables to explore, but we have chosen the available, relevant data that we could find. Our selection of variables is based on their accessibility in our data sources and their potential to provide valuable insights into pass value and team performance.

For both types of independent variables, we will conduct two types of analysis: correlation analysis and regression analysis. These analyses will provide insight into the relationship between the independent variables and the pass value.

First, for the correlation analysis, we will calculate the correlation coefficient between each independent variable and the pass value. This coefficient measures the strength and the direction of the linear relationship between the variables.

Then, in the regression analysis, we will use linear regressions to model pass value as a function of the independent variables. We will conduct separate linear regressions for each independent variable to examine the relationship between each variable and pass value individually.

4.2.2 Comparative analysis

In addition to to the correlation and regression analysis, we also consider a comparative analysis. For this we are also going to consider the subtype labels that successful passes have in the event data. We have the following list of subtypes, with definitions of these types from SciSports, the data supplier (SciSports 2024).

- Open Play
  "A pass that successfully reaches a teammate during open play."
- Open Play — Switch
  "A successful pass played from one side of the field to the other during open play, usually to change the point of attack or exploit open space."
- Open Play — Final 3rd
  "A successful pass originates from the final third and reaches the intended recipient, potentially leading to a scoring opportunity."
- Between Lines
  "A successful between-the-lines pass is an advanced tactic where a pass is initiated behind the opponent's midfield, penetrating through their defensive layers and landing between their midfield and defensive blocks."
- Behind last Line
  "A successful pass played behind the last line of defense of the opposing team, often providing an opportunity for the receiving player to create a scoring chance."
- Pre-Key Pass
  "A successful pass that sets up the key pass, contributing to the creation of a goal-scoring opportunity."
- Key Pass
  "A successful pass that directly leads to a goal-scoring opportunity, typically a shot."
- Assist
  "A successful pass that directly leads to a goal scored by a teammate."

An assumption could be that passes with such a subtype (different from Open Play, which can be stated as the 'standard' pass) should be rated higher in general than the 'normal' passes. We will consider the average score for the available matches for each type of pass.

4.3 Applications

Besides investigating the relationships of the pass value with the variables mentioned before, it is useful to see how we can apply this pass value in practice, especially for the football club. We will consider a few different things, first we look at alternative pass options and then we consider rating players.

4.3.1 Considering alternative passes

With ratings of passes, we can simply judge if a pass was good or not, but it is also interesting to see if it was the best available option. To consider this question, we will, for all rated passes, determine the pass value for the alternative passes, for which the assumption holds that at the end of the pass the ball is closer to the goal then before.

To determine the pass value for alternative passes, we have to make some assumptions. First, for the coordinates of the hypothetical receiver, we take the coordinates of the player at the start of the pass, because we do not know exactly where the player would be in the hypothetical case, and this is the information we do have from a player. Also, the Pitch Control is determined for the start of the pass, so to be able to use that, this is the most logical choice to make. Furthermore, for the Packing we need an end timestamp. For this, we take the end timestamp of the original pass. For future research it can be interesting to take an average speed of the ball, and determine the distance between the start and end coordinates of the hypothetical pass and determine the time it takes to travel this distance on average, but since that takes more computational time, we will not consider this approach right now.

Then, we will consider the number of passes that are rated higher than the original pass and analyze the difference between the highest possible rating and the original. For every match, a summary can be generated, allowing for insightful analysis. For instance, trainers can identify if specific players, or players in specific areas of the field, tend to make non-optimal choices more frequently. Such analysis provides trainers with valuable insights into where to focus on.

4.3.2 Player rating

In the article of (Bransen et al. 2019), a metric is introduced to rate players on their pass value called the ECOM rating. This metric is defined as follows:

$$ECOM = \frac{\text{sum of the values of the player's passes}}{\text{number of minutes played}} \cdot 90 \tag{15}$$

In this metric, we can use our constructed pass value in the numerator. With this, we can rate the individual players of the football club. For the football club it can be interesting to see who scores well and who can improve more on their passes. They can also distinguish between the categories of players: the keeper, defender, midfielders and attackers, and see within these categories who performs best and who can improve. They can take this into consideration when constructing their formation.

4.3.3 Relationship analysis

For both of these applications, it is also interesting to do some relationship analysis. To start with the alternative passes, we get a number of how many passes were the best possible option in the match. We are going to consider a regression between this number of best passes and the three independent variables as seen in 5.2.1: pass accuracy, ball possession and the number of goals in a match.

For the player rating we will do some comparative analysis. On the FotMob website mentioned before, there is a rating of the players of FC Eindhoven available. For the players of FC Eindhoven, we will compare the FotMob rating to the ECOM rating that we get from using our constructed pass value. To compare the rankings with each other, we will use Spearman's rank correlation coefficient (Chattamvelli 2024). If all ranks are distinct integers, the Spearman rank correlation coefficient can be determined with the following formula.

$$r_s = 1 - \frac{6 \sum {d_i}^2}{n(n^2 - 1)} \tag{16}$$

In this formula, $d_i = R(X_i) - R(Y_i)$, where $R(X_i)$ and $R(Y_i)$ are the ranks that come from the 'raw' values X_i and Y_i. n is the number of observations.

5 Results

In this chapter we will elaborate on the results we have obtained based on the methodology in the previous chapter. We will start with considering the pass value. For the pass value we will study the parameter estimation and the selection of the best performing model. Afterwards we will discuss the relationship analysis and see if there are relations between certain other football statistics, based on correlations and regression, and also with comparative analysis. We will conclude this chapter with the possible application of this model in practice with considering alternative passes and rating players based on their pass value.

5.1 Pass value

As stated before the main goal of this thesis is to construct a representative pass value model. The construction of the potential models has been discussed in Section 4.1.5. In this part we will elaborate on the results of the parameter estimation and the model selection. With the model that we select based on the results, we execute the relationship and comparative analysis and look at its applications.

5.1.1 Parameter estimation

We start off with the parameter estimation. As stated in Section 4.1.5, we will begin with determining the Variance Inflation Factor (VIF) for the independent variables to see if there is multicollinearity. Then we will perform the regression, depending on the results of the VIF scores, this will be a linear or ridge regression.

Model 1
The first model is, as introduced in Section 4.1.5, defined as $\text{PV} = \alpha \cdot P + \beta \cdot \text{xT} + \gamma \cdot \text{PC}$. We will first determine the VIF scores between the three variables that are contained in this pass value: Packing (P), Expected Threat (xT) and Pitch Control (PC). The results of this are shown in Table 2. For the interpretation of VIF, it is a rule of thumb that when the value exceeds ten, it indicates an amount of collinearity which can be noted as problematic (Neter et al. 1983). All our VIF scores are smaller than ten which means there is no multicollinearity present.

Independent variables	**VIF Score**
P	6.556
xT	2.049
PC	4.813

Table 2: VIF scores model 1

Because there is no case of multicollinearity we will as discussed in Section 4.1.5 perform a linear regression. As 'real' values we use the expert values seen before. With that we obtain the parameters shown in Table 3 below.

Variable	**Parameter**	**Estimation**
P	α	0.7877
xT	β	0.7820
PC	γ	0.0215

Table 3: Parameter estimation model 1

From this parameter estimation, we can conclude that mostly Packing and Expected Threat have a high influence on the pass value, especially compared to the Pitch Control.

Model 2
In model 2 we distinguish two cases: the first $\frac{2}{3}$ of the field and the last third. For the first $\frac{2}{3}$, the model is defined as $\text{PV} = \gamma_1 \cdot \text{PC} + \delta \cdot D$ and for the last, it is the same as model 1, $\text{PV} = \alpha \cdot P + \beta \cdot \text{xT} + \gamma_2 \cdot \text{PC}$. We start off with determining the multicollinearity for the first $\frac{2}{3}$: for this, there are only two independent variables, so as stated before we will consider the correlation between these to determine whether there is multicollinearity. For the two independent variables PC and D, we get a correlation of -0.1561, which indicates that there is no multicollinearity since it is close to 0. Because of this, a linear regression will be performed.

For the last third of the field we consider the same kind of model as model 1, so based on those VIF scores we again do a linear regression. Performing the linear regression separately on the first and last part of the field, we obtain the parameters that are shown in Table 4.

Variable	**Parameter**	**Estimation**
PC	γ_1	0.3383
D	δ	1.0584
P	α	0.1982
xT	β	2.8797
PC	γ_2	0.0272

Table 4: Parameter estimation model 2

In the parameter estimation of model 2, we notice that for the first part of the field, the distance has a high influence on the pass value. For the last part of the field, we notice that the xT has a very high influence, while Packing and Pitch Control do not. Compared to model 1, we can conclude that xT has the most influence on the last part of the field.

5.1.2 Model selection

With the parameters we have retrieved in Section 5.1.1 we can determine the pass values for both models. We will compare these retrieved pass value to the original expert values and consider three different metrics that measure the predictive accuracy. Below is Table 5 with the MSE, RMSE and MAE for the normalized values. From this table we can clearly see that model 1 performs better than model 2, since for all metrics the values are lower for model 1 than for model 2.

Metric	**Model 1**	**Model 2**
MSE	0.0468	0.0952
RMSE	0.2164	0.3085
MAE	0.1687	0.2417
R^2	0.28	-0.46

Table 5: Model selection (normalized)

For intuition, we also added Table 6, where we consider the same metrics but instead of the normalized values from 0 to 1, we consider the originally given scale from 1 to 5,

which means that for model 2, the MAE around 1 means that on average, the error is one whole point from the 'real' value.

Metric	Model 1	Model 2
MSE	0.7491	1.5231
RMSE	0.8655	1.2341
MAE	0.6702	0.9714
R^2	0.28	-0.46

Table 6: Model selection (scale 1 to 5)

For all performance metrics, the differences between the models are minimal, but in all cases model 1 performs better. However, we also calculated the R^2 values for both models. Model 1 has an R^2 of 0.28, while model 2 has an R^2 of -0.46. The R^2 value of -0.46 for model 2 indicates that it performs worse than a naive prediction using the mean of the observed values. In other words, the model's predictions are less accurate than simply predicting the average of the observed data. On the other hand, the R^2 value of 0.28 for model 1 indicates that 28% of the variance in the observed data is explained by the metrics used in this model. This suggests that model 1, while not perfect, provides some predictive power and performs better than model 2. From now on, we will continue with using model 1 in our analysis.

To visualize the performance of model 1, we compare the median of the expert values to the values we have obtained in our model. The results are shown in the following Figure 4.

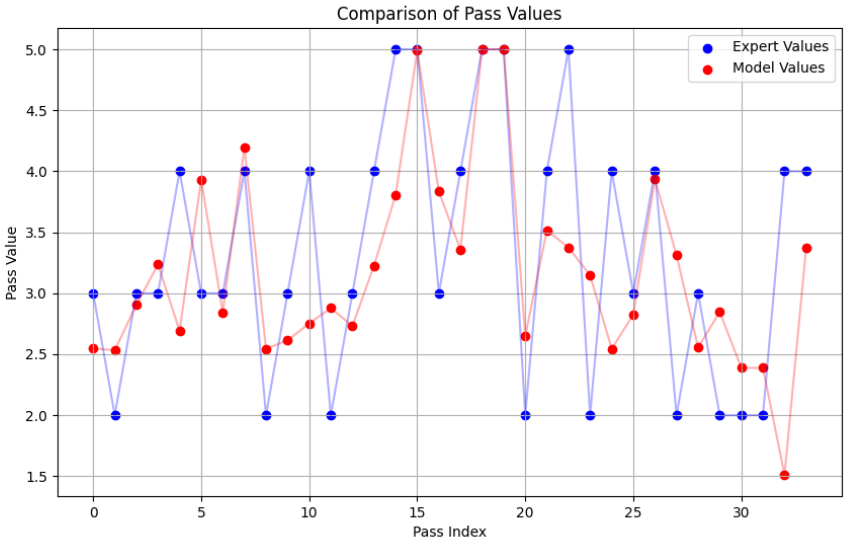


Figure 4: Expert values vs. model values

5.2 Relationship Analysis

Now that we have selected our pass value model, we will explore certain relationships between this pass value and certain statistics about passes and matches.

5.2.1 Correlation and regression analysis

We have two types of independent variables to consider and two types of analysis. Starting with the first type of independent variables: variables related to specific passes.

Pass distance
To execute the regression for pass distance, we have categorized the passes into three categories: Short, Medium and Long. Short are passes that are between 0 and 13.75 meters, Medium passes between 13.75 and 27.5 and Long passes are longer then 27.5 meter (Cordón-Carmona et al. 2023). We have used for this all the forward-passes performed by FC Eindhoven, which is a total of 4879. The OLS regression results for the pass distance (Short, Medium, Long) are as follows:

Variable	Coefficient
Intercept (Short)	2.1259
Medium	0.1144
Long	0.4259

Table 7: OLS regression pass distance

We can interpret the results as follows. The intercept term represents the baseline pass value grade for the category Short. The pass value grade is expected to increase or decrease compared to the Short passes by the coefficient for each pass type when those types of passes occur. In our case, that states that for Medium passes it increases with 0.1144 compared to short passes. For Long passes it increases with 0.4259. It seems that long passes are valued higher on average. If we consider the mean values for all three types of passes performed by FC Eindhoven over all available matches in Table 8, we see that is the case.

Pass distance	Mean pass value	# passes
Short	2.13	1723
Medium	2.24	2116
Long	2.55	1040

Table 8: Mean pass values for pass distance

Player position
The OLS regression results for the player position (Keeper, Defender, Midfielder and Attacker) are as follows:

Variable	Coefficient
Intercept (Keeper)	2.3997
Defender	-0.1276
Midfielder	-0.1579
Attacker	-0.2041

Table 9: OLS regression player position

The intercept states that for the baseline of the position Keeper, the expected value of the pass value is 2.3997. For the other positions, the coefficient determines the difference compared to the baseline value. We see that all player positions are rated lower than the keeper, but the differences between the players is not that big. If we consider the mean values for the different player position in Table 10, we see that the differences between these are indeed very small.

Player position	Mean pass value	# passes
Keeper	2.40	484
Defender	2.27	2354
Midfielder	2.24	1403
Attacker	2.20	631

Table 10: Mean pass values for player positions

For both of these regressions, the R^2 is 0.008, which indicates that the independent variables only explain a small proportion of the variance in the pass value and the relationship between our variables is weak.

Next, we have the variables that are related to an entire match. For this, we consider each team in each available match of FC Eindhoven with the correct type of event data (26 matches in total, 52 teams). For all those matches, we have extracted the values of three variables: pass accuracy, ball possession and number of goals from the website FotMob ("FotMob" 2024). We will examine the correlation and regression analysis for these three variables separately.

Pass accuracy
The correlation coefficient for pass accuracy and the pass value is -0.216. The pass accuracy is the percentage of successful passes in the match. The negative correlation indicates that if the pass value increases, the pass accuracy decreases. So, if there are relatively fewer successful passes in the match, they are of higher value. Since the correlation coefficient is close to zero, it indicates a weak correlation.

With the linear regression we got the following values:

Intercept	Slope
3.045	-0.00974

The slope is -0.00974, which means that for every 1 percentage point of increase in pass accuracy, the pass value decreases with 0.00974, so this is a very small change per

percentage point. The intercept states that if the pass accuracy would be 0, the pass value would be 3.045, but this has no meaning as in practice this never occurs.

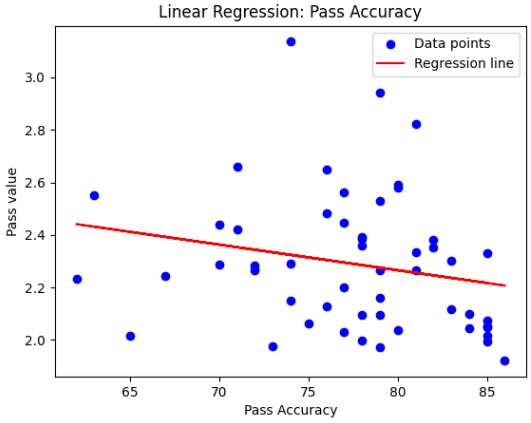


Figure 5: Linear regression with pass accuracy

Ball possession

We continue with ball possession. The correlation coefficient is -0.104, which again means there is a small negative relation. If the pass value increases, the ball possession decreases, so if the team has less ball possession, the pass values become higher. But the score is very close to zero, so again a weak correlation. Then we determine the linear regression and we get the following values:

Intercept	**Slope**
2.437	-0.00293

The slope is -0.00293, which means that for every 1 percentage point of increase in pass accuracy, the pass value decreases with 0.00293. This is again a very small change per percentage point. The intercept for the ball possession states that if the ball possession would be 0, the pass value would be 2.437. Again, this is not interpretable as this will never happen.

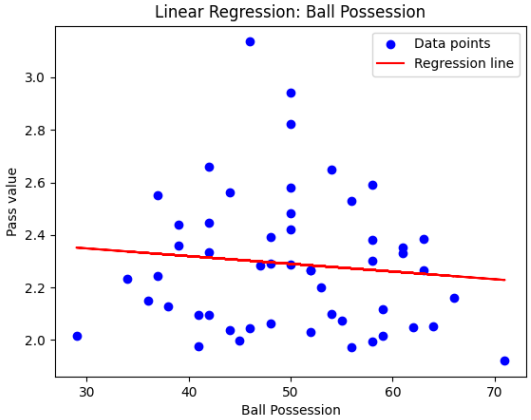


Figure 6: Linear regression with ball possession

Interesting to notice is that the results for the pass accuracy and ball possession are quite similar. We have looked into the correlation between these two variables, and this is 0.715, which is a strong positive relationship, which explains the similar results.

Number of goals

Lastly we have the number of goals. The correlation coefficient is 0.196, which means there is a small positive relation. If the pass value increases, the number of goals also increases. The score is again close to zero, so it is a weak correlation. Then we determine the linear regression and we get the following values:

Intercept	Slope
2.229	0.050

The slope is 0.050, which means that for every goal, the pass value increases with 0.05, and the intercept states that when the number of goals is zero, the pass value is 2.229.

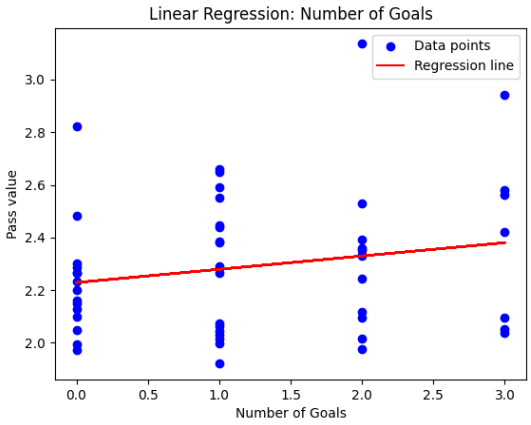


Figure 7: Linear regression with number of goals

For the three variables we also computed the (R)MSE, MAE and R^2, to observe the performance of the regression, shown in Table 11.

Metric	Pass accuracy	Ball possession	Number of goals
MSE	0.0640	0.0664	0.0645
RMSE	0.2529	0.2576	0.2540
MAE	0.2016	0.2030	0.1945
R^2	0.0466	0.0109	0.0385

Table 11: Performance regressions

The performance metrics (MSE, RMSE, and MAE) are relatively low for all three variables, indicating that the regression models predict these values quite well. However, the R^2 values, which measure the proportion of variance explained by the model, are also low. These low R^2 values suggest that while the models predict the variables with reasonable accuracy as indicated by the error metrics, they do not explain a large portion of the variance in the data. In other words, there seems to be a relation between pass accuracy, ball possession, and number of goals with the pass value. However, these relations are not very strong. While these variables explain some of the variations in

pass value, there may be additional factors influencing pass value that were not included in the analysis.

For the pass accuracy and ball possession, we can possibly explain the negative relationship because we only consider successful passes. First, we consider the pass accuracy. Because the pass value model only considers passes that are successful, we only measure the effectiveness of the pass and do not consider the completion rate. In offensive play, teams often take risks with daring passes, which can lead to a lower overall pass accuracy. However, successful execution of these riskier passes frequently results in higher valued passes. Thus, although pass accuracy may decrease, the average value of the successful passes rises. This can explain the observed negative correlation between pass accuracy and pass values. For ball possession, a similar logic applies. When a team has lower ball possession, they are more likely in a defensive position which offers more opportunity for counterattacks with daring chances when regaining possession. The successful passes in daring chances are often more valuable and since we only consider successful passes, this can explain the negative correlation between ball possession and pass values.

For the number of goals, a positive relationship seems plausible: as the number of goals increases, we expect pass value to increase as well. However, it's also worth noting that pass value might not have a significant impact on the actual number of goals scored, which aligns with realistic expectations.

5.2.2 Comparative analysis

For the comparative analysis we consider the different subtypes of the successful passes. From the available data and correct event data, with the restriction of only forward passes, we get 10708 passes considering both FC Eindhoven and the opponent. In Table 12, we see the results of the average pass value per subtype and the amount of passes per subtype.

Event subtype	**Average pass value**	**# passes**
Open Play	2.20	9135
Open Play — Switch	2.28	79
Open Play — Final 3rd	2.43	615
Between Lines	2.83	431
Behind last Line	2.78	175
Pre-Key Pass	2.70	73
Key Pass	3.04	178
Assist	3.53	22

Table 12: Average pass value for subtypes

In this table we observe that important passes, such as key passes and assists, are indeed rated higher than other passes, as we assumed. Notably, the 'standard' successful pass, without any special label in the subtypes, receives the lowest rating, which aligns with the expectations given that the subtypes typically denote events with greater impact. These findings strongly suggest that the pass value model is representative of real-game dynamics, which is a very promising result.

5.3 Applications

In this part we are going to look at the practical applications that follow from the pass value model. We first start with alternative passes and then elaborate on the ECOM rating to rate players. We finish with performing a relationship analysis on these two applications.

5.3.1 Considering alternative passes

For considering alternative passes we used all accessible matches that were in the correct data format. For all forward passes of FC Eindhoven, which was a total 4879 passes, we have computed the number of alternative passes with a higher pass value, determined the best alternative and computed the difference between the best alternative and the modeled pass value. Because the alternative pass cannot go to the sender or receiver of the original pass, the maximum of alternative passes is 9. Because of the assumption that we only rate passes with our model that go towards the goal, it can be that this maximum number is lower in practice. We have constructed an overview in Appendix A.2 that lists the frequency of events where a specific number of alternative passes with a higher value occurred.

As stated before, it can be interesting for trainers to analyse a match afterwards based on the alternative passes. The data for one match is stated in Table 13. There were a total of 157 passes considered.

# better alternatives	# events
0	17
1	8
2	25
3	15
4	17
5	23
6	24
7	11
8	11
9	6

Table 13: Better alternatives

It is important to take into account that football is a very dynamic, in the moment game and it is very likely that passes where our model states that there were several better options, in practice the original pass was the best feasible option.

The practical application is that a football trainer can get this overview and look at the events that had a lot of better options after a match. But, if an event had a lot of alternatives but the difference in pass value is not that high, the number of better options is not very important. That is why we have looked into the size of the difference from the best alternative with the original pass value. With this, we can also look at certain interesting values in Table 14.

Range of the difference d	**# events**
$d < 0.5$	50
$0.5 \leq d < 1$	58
$1 \leq d < 1.5$	47
$d \geq 1.5$	2

Table 14: Range of differences

With the results from this table, a trainer can look into the passes with the largest difference, and look for example at all passes that have a difference of more than one. This means that there is a pass that would have received a rating one point higher on a scale of 1 to 5, which can be a significant increase. As stated before, everything depends on a lot of factors in the match, so these results should be taken into consideration as a guideline to look into certain passes but not as a 'holy grail'. We look into an example of an event that had a difference bigger than 1.5 in Figure 8 to illustrate this.

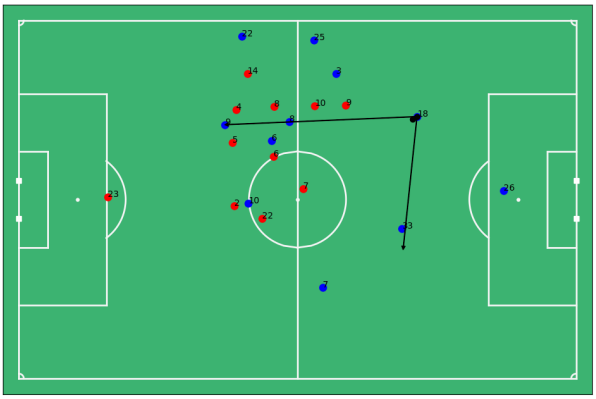


Figure 8: Example of pass with $d \geq 1.5$

We asked one of our experts for his opinion on this best alternative. The original pass to player 33 got a value of 1.83, while the best alternative to player 9 received a value of 3.38. The first thing that he mentions is that player 9 is offside, so this pass could not have been executed in play. Furthermore, one of their own players is in the line of pass, making it difficult to execute. Finally, we see that there are many opponents around the line of pass, giving a high chance of an intercept of the pass by the opponent. The fact that this alternative pass received a much higher rating in our model can be explained, since the Packing of the alternative is ten while for the original pass, it is zero. Even though the Pitch Control for the alternative is much lower, this metric is accounted for the least in the pass value model. These kind of observations might indicate the need for more investigation in measuring the feasibility or difficulty of a pass before using the pass value in considering alternative passes.

The last part of this application is to look at specific players, their mean difference and mean number of better events among the passes in a match. From this, you can obtain the information on which players maybe have to improve on their passing choices.

Player	**Mean number of better events**	**Mean difference**	**# passes**
33	5.2	0.91	30
18	5	0.84	34
3	4.06	0.80	16
25	4	0.75	6
26	4.79	0.68	14
7	3.76	0.64	21
6	3.9	0.64	10
11	3	0.56	4
8	2.33	0.43	9
10	1.71	0.34	7
22	0.75	0.30	4
9	1	0.19	1
2	0	0	1

Table 15: Player evaluation

It can be interesting to take into account the amount of passes a player performed. In this case we see for example that if you would look at player 2, not knowing the amount of passes, it would appear that this is an amazing player. However, now we know that only 1 pass of this player is considered in this data, so it is less representative for the skills of the player. Statistics like this (for one match or a whole season) can give some insight into the pass quality of a player.

5.3.2 Player rating

With the formula from Section 4.3.2 we can determine the ECOM rating per player for FC Eindhoven. We get the results in Table 16.

Jersey number	Player position	ECOM rating
26	Keeper	0.74
12	Keeper	0.69
3	Defender	0.95
33	Defender	0.73
2	Defender	0.62
15	Defender	0.98
18	Defender	1.04
32	Defender	0.7
25	Defender	0.44
7	Midfielder	0.38
8	Midfielder	0.71
27	Midfielder	0.99
99	Midfielder	0.97
23	Midfielder	0.51
6	Midfielder	0.70
5	Midfielder	0.37
11	Attacker	0.24
14	Attacker	0.78
22	Attacker	0.25
9	Attacker	0.14
19	Attacker	0.13
10	Attacker	0.46

Table 16: ECOM rating

The results of the ECOM rate can be intuitively interpreted as the amount of goals that is expected to be made following the passes a player have executed during 90 minutes of play (Bransen et al. 2019).

Player position	Average ECOM
Keeper	1.43
Defender	0.78
Midfielder	0.66
Attacker	0.33

Table 17: Average ECOM ratings

It is interesting to observe that the attackers have on average a lower ECOM rating than other players. We think this is because of the components that are used in the pass value: during the attacker's pass, more opponents are closer together, which likely results in lower Pitch Control compared to defenders and midfielders. Also, the distance covered is most of the times lower and the opponents move along with the ball more easily than with longer passes, so the Packing score is probably also lower.

5.3.3 Relationship analysis

In this section, we will look into the relationship of the introduced applications. For alternative passes, we will look into the relation of the amount of 'perfect' passes per match, while for the player rating we will compare the ECOM rating to the FotMob rating.

Considering alternative passes
For the relationship analysis of the alternative passes, we execute the same kind of research as we did in Section 5.2.1, with as independent variables again the pass accuracy, the ball possession and the number of goals. As dependent variable, we look at the number of 'perfect' passes, so passes that had 0 better alternatives, which would mean that in our model it would be the perfect pass for that moment.

Starting with the pass accuracy, we get a correlation coefficient of 0.367, indicating a moderate positive correlation. Then we perform the linear regression and we get the following values:

Intercept	Slope
-31.86	0.8203

The slope is 0.8203, which means that for every 1 percentage point of increase in pass accuracy, the number of perfect passes increases with 0.8203, and the intercept states that if the pass accuracy would be zero, the amount of perfect passes is -31.86. While this negative number might seem unrealistic, it can be explained by the fact that the pass accuracy only has values between 60% and 85%. The regression line extends beyond this range, leading to a negative intercept.

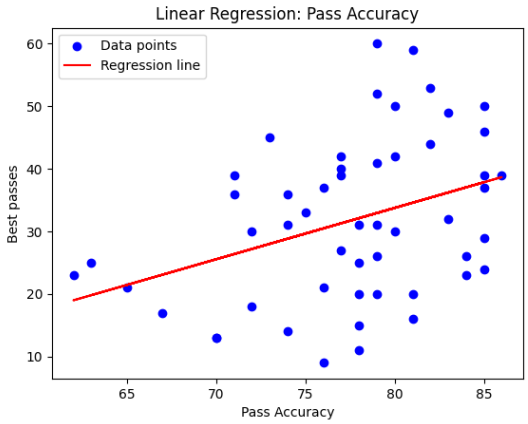


Figure 9: Linear regression with pass accuracy

For the ball possession, the correlation coefficient is 0.478, which means that there is a moderate positive correlation between the percentage of time the team has possession of the ball, and the amount of 'perfect' passes that have been made. Next we perform the linear regression, and we get the following values:

Intercept	Slope
-1.61	0.6664

The slope is 0.6664, which means that for every 1 percentage point of increase in ball possession, the number of perfect passes increase with 0.6664, and the intercept states that if the ball possession would be zero, the amount of perfect passes is -1.61.

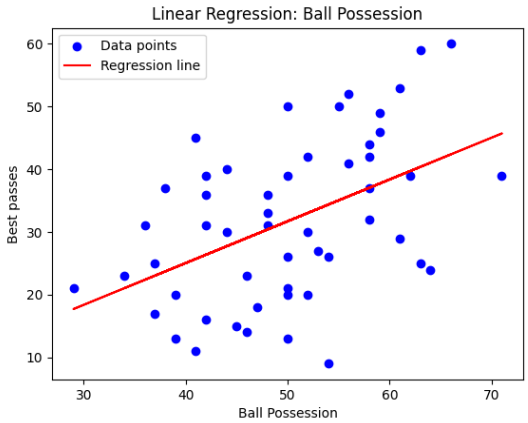


Figure 10: Linear regression with ball possession

Last, for the number of goals, the correlation is 0.027, very close to zero, which indicates there is no correlation. When we perform the linear regression, we get the following values:

Intercept	**Slope**
31.29	0.345

The slope is 0.345, which means that for every goal, the number of perfect passes increases with 0.345, and the intercept states that if the number of goals would be zero, the amount of perfect passes is 31.29.

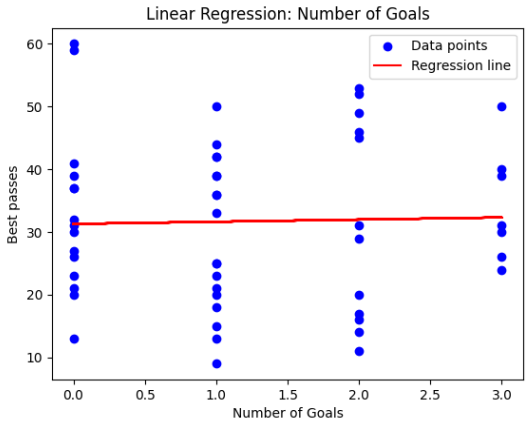


Figure 11: Linear regression with number of goals

For the three variables we also computed the (R)MSE, MAE and R^2, to observe the performance of the regression, shown in Table 18.

Metric	Pass accuracy	Ball possession	Number of goals
MSE	141.6425	126.4659	163.6616
RMSE	11.9014	11.2457	12.7930
MAE	10.0478	9.6710	10.7421
R^2	0.1352	0.2278	0.00074

Table 18: Performance regressions

From the table, we observe relatively high scores for (R)MSE and MAE. The MAE being the most interpretable metric, indicates the average discrepancy between the predicted and actual values. For instance, in the case of Pass Accuracy, the regression model predicts, on average, 10 fewer successful passes than the actual values. These discrepancies suggest that the relationship may not be captured by the regression models.

The R^2 values further support this observation. They are relatively higher compared to previous relations with the pass value in Section 5.2.1, but they still indicate that only a small portion of the variance in the best passes is explained by the independent variables. Interestingly, the R^2 value for the number of goals is extremely low, suggesting that the number of goals explain almost none of the variance in the amount of best passes.

Player rating

We can compare the ECOM rating we have for the players of FC Eindhoven, to the FotMob rating ("FotMob" 2024) that is available for these players. In Appendix A.3 two tables are included. First, Table 26 states the jersey number and type of player, together with their FotMob and ECOM rating. The next table is Table 27, which compares the rating from highest to lowest for both FotMob rating and ECOM rating.

Within these tables we cannot directly see if there is a relationship. To compare these ratings in a better way, we are going to separately look at the different positions. Since there are only two keepers, we will leave these out. We consider defenders, midfielders and attackers separately.

Starting off with the defenders, if we sort the players on highest to lowest score for both the FotMob and ECOM rating we get the following table.

#	Sorted on FotMob	Sorted on ECOM
1	3	18
2	18	15
3	15	3
4	25	32
5	2	33
6	33	2
7	32	25

Table 19: Player rating defenders

In this table, we see that the top 3 is the same in both ratings, only in a different order, so the same players are rated as the best defenders in both ratings. If we look more at the bottom we see that numbers 2 and 33 switched places in the different orders so they are close to each other. What is striking, is that number 25 and 32 are switched in the different orders.

We have constructed the same table for the midfielders.

#	**Sorted on FotMob**	**Sorted on ECOM**
1	27	27
2	7	99
3	6	8
4	8	6
5	23	23
6	99	7
7	5	5

Table 20: Player rating midfielders

We see that for both ratings, the best and lowest rated midfielder are the same. For the midfielders, the most striking is number 7 and 99 that are in one of the ratings on the second place and on the other in the second to last place. For the FotMob ratings, the midfielders are all very close to each other, but for the ECOM there is a big difference between the players. A likely explanation is that it is possible that a player like player 7 is rated higher on FotMob because of other factors than the pass value takes into account.

Last, we look at the attackers.

#	**Sorted on FotMob**	**Sorted on ECOM**
1	10	14
2	14	10
3	22	22
4	11	11
5	19	9
6	9	19

Table 21: Player rating attackers

These ratings look pretty similar. The first two, the middle two and the last two are in both ratings on the same place or switched with each other, which means the best and lowest rated players are in both ratings about equally good.

To compare the ranks based on something else than just our view, we consider Spearman's Rank Correlation (SRC), shown in Table 22. The interpretation of this correlation is similar to Pearson's correlation coefficient, which means a correlation is strong starting from 0.7 (Schober et al. 2018).

Player position	**SRC**
All	-0.1648
Defenders	0.5714
Midfielders	0.4144
Attackers	0.8857

Table 22: Spearman's Rank Correlation

It is interesting to note that while considering all player positions together, the correlation coefficient is negative, suggesting a weak negative association. However, when considering the positions separately, we observe moderate and even strong correlations. This suggests that the ECOM rating can be effective in assessing passing skills among players, specifically compared to players within the same position.

6 Conclusion

In this chapter, we will start with giving a summary of the research. After that, we continue with the conclusions that can be drawn from this thesis. We will finish this chapter with recommendations for future research on the subject of valuing passes in football.

6.1 Summary

In this thesis, we have constructed a representative pass value model for FC Eindhoven and analyzed the relationship between passes and certain other statistics. To start with, we constructed two different pass value models for which the parameters were estimated based on expert data we gathered. These parameters are estimated via a linear regression, since there was no multicollinearity among the independent variables. Then we selected the best performing model based on (R)MSE and MAE as metrics to measure the predictive accuracy of the models.

After selecting the best performing model we looked into the potential relationship of the pass value and other statistics. First for individual passes we looked at the pass distance and player position, and secondly we looked at statistics for a complete match such as pass accuracy, ball possession and the number of goals. For these we performed an OLS regression. Furthermore, we compared the average values of event subtypes with each other to see if passes that were highlighted by the event data as important passes (such as key passes and assists) were also valued as better passes in our model. Lastly, we looked into possible practical applications of the pass value model. For this we considered two options: alternative passes and player ratings.

6.2 Conclusions

To state our conclusions, we will consider the three research questions that were formulated at the beginning of the thesis.

How can the expert-identified factors be incorporated into a representative pass value model?
Based on the suggestions of the experts, we have constructed two different types of pass value models. While the first model combines the three components in a linear way, the other model makes a distinction between passes in the first $\frac{2}{3}$ of the field and the last third, where in the beginning of the field, the distance travelled and Packing are found more important, and for the last part the first model is considered.

Based on the expert data we estimated the parameters of the two different models. For model 1, especially Packing and Expected Threat seem to be good explanatory variables for the pass value. In model 2, we see that for the first part of the field, the distance has a high influence on the pass value, while for the last part Expected Threat explains the pass value most. With metrics as the (R)MSE and MAE we measure the predictive accuracy of the models. For all these metrics, we found in Section 5.1.2 that model 1 came out as the best model of the two.

What is the relationship between our pass value model and various statistics regarding passes and matches?
We have looked at different statistics, and we will shortly recap for each of these what we have found. For the pass distance, we labeled each pass as 'Short', 'Medium' or 'Long'. It seems that on average, the longer the pass, the better pass value it received. This can possibly be explained since both Expected Threat and Packing are on average higher when a pass travels a longer distance. Then, for the player position we considered the position on the field a player has, and we see that the average pass values for the different player positions are very close to each other. After the pass-specific statistics, we looked at match statistics. We started with the pass accuracy, which is the percentage of successful passes in a match, and the ball possession, which is the percentage of times the team had possession of the ball. For both of these, it seems that there is a negative relationship between these properties and the pass value. This can be explained based on the fact that we only consider successful passes, which means that more risk is taken, which means there may be fewer successful passes and less ball possession, but the successful passes that are made are rated higher. Lastly, for the number of goals a team made during a match, it seems there is a small positive relationship.

Furthermore, we compared the average pass values of different event subtypes and noticed that all subtypes got on average higher scores than the 'normal' pass. Passes such as key passes and assists are rated the highest, which is a nice result that is in line with what we would have expected.

How can this model be applied in practice to optimize team performance?
We introduce two applications that can be used by football trainers. First we look at alternative passes. Based on the starting position of an executed successful pass, we consider the potential alternative receivers and corresponding passes. For these passes we compute the pass value with our model. With this, you can get valuable information about how often, compared to the alternatives, a very good or very bad pass has been made and the difference between the real pass value and the best alternative can be considered. This application can be used to analyze specific matches, players or situations that a trainer wants to look into.

The second application we looked into is rating players with our pass values. For this, our pass values are used as input for the ECOM rating. The performance of players can be compared based on this rating. Rating players on their pass value can be useful to see how well the player performs, and compare them with other players. This can also be used when making the line-up for a match.

6.3 Future research

There are a lot of additions that can be made in the future based on our research. We filtered our data and considered only forward passes, since Packing and Expected Threat are negative for most passes directed away from the goal you are targeting, so that would have given negative pass values. For these backwards passes, a suggestion could be to look into the positions at the start of the pass and to investigate the coordinates of the opponents. If there are a lot of opponents in the line of pass when the player would have played forward, playing backward can be a good idea to create space. But, for these passes, we recommend to first again talk to experts to find out what they think is important in these situations.

Furthermore, we only considered successful passes, because for those, we can relatively easily determine the receiver of the pass. Future research could consider unsuccessful passes as well. The first step then would be to determine a way to predict the intended receiver. Then, the pass could be rated similarly to successful passes, but an additional factor could be incorporated to account for the fact that it is an unsuccessful pass. This would ensure that successful and unsuccessful passes from the same start and (predicted) end coordinates receive different ratings, reflecting the importance of the pass's success in its overall value.

Additionally, looking into the alternative passes can be a promising area to improve on. Since our approach indicates that the feasibility and difficulty of the pass are not accounted for in a proper way, it is interesting to do more research on that subject. Afterwards, exploring the possibility of constructing a player rating that combines the ECOM rating and the alternative passes analysis could be an interesting approach. This method would account for the pass value while also comparing it to the available alternatives, offering a more comprehensive evaluation of player decision-making.

Finally, it can be good to explore other models, take into account more components, or get a bigger 'real' expert data set. There are several possible ways to look for improvements. We leave this for future research.

References

Anzer, Gabriel, and Pascal Bauer. 2022. "Expected Passes." *Data Min Knowl Disc* 36 (January): 295–317. https://doi.org/10.1007/s10618-021-00810-3.

Barnett, V., and S. Hilditch. 1993. "The Effect of an Artificial Pitch Surface on Home Team Performance in Football (Soccer)." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 156 (January): 39. https://doi.org/10.2307/2982859.

Biermann, Ch. 2015. "Reinartz & Hegeler." *11 Freunde* 168:62.

Bransen, Lotte, Jan Van Haaren, and Michel van de Velden. 2019. "Measuring soccer players' contributions to chance creation by valuing their passes." *Journal of Quantitative Analysis in Sports* 15 (February). https://doi.org/10.1515/jqas-2018-0020.

Bresenham, J. E. 1965. "Algorithm for computer control of a digital plotter." *IBM Systems Journal* 4 (1): 25–30. https://doi.org/10.1147/sj.41.0025.

Chattamvelli, Rajan. 2024. *Correlation in Engineering and the Applied Sciences: Applications in R.* Springer Nature.

Cordón-Carmona, Antonio, Víctor Álvarez, Santiago Morales, Daniel Mon, Abraham Garcia, and Ignacio Román. 2023. "The Influence of Pass Length and Height in Europe's Top 5 Leagues in Men's Football." *The Open Sports Sciences Journal* 16 (December). https://doi.org/10.2174/011875399X263057231127051556.

Felices, Alex Marin. 2023. "Beyond Succes and Failure: Measuring Pass Quality in Football with Tracking Data Analysis and Use Cases," June. https://medium.com/@marin11amf11/pass-quality-assessment-in-soccer-a-multi-dimensional-approach-with-continuous-evaluation-febee2c3b900.

"FotMob." 2024. Accessed on 01-05-2024. https://www.fotmob.com/nl/teams/6416/squad/fc-eindhoven.

Green, Sam. 2012. https://www.statsperform.com/resource/assessing-the-performance-of-premier-league-goalscorers/.

Hossain, E. 2023. *Machine Learning Crash Course for Engineers.* Springer International Publishing. ISBN: 9783031469909. https://books.google.nl/books?id=Q1frEAAAQBAJ.

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. 2013. *An introduction to statistical learning.* Vol. 112. Springer.

Khorana, Arjun, Ayoosh Pareek, Matthieu Ollivier, Sophia Madjarova, Kyle Kunze, Benedict Nwachukwu, Jón Karlsson, Erick Marigi, and Riley Williams III. 2022. "Choosing the appropriate measure of central tendency: mean, median, or mode?" *Knee Surgery, Sports Traumatology, Arthroscopy* 31 (November): 1–4. https://doi.org/10.1007/s00167-022-07204-y.

Neter, John, William Wasserman, and Michael Kutner. 1983. "Applied Linear Regression Models." https://api.semanticscholar.org/CorpusID:126196130.

Pollard, Richard, Jake Ensum, and Samuel Taylor. 2004. "Estimating the probability of a shot resulting in a goal: The effects of distance, angle and space." *Int. J. Soccer Sci.* 2 (January).

Rudd, Sarah. 2011. "A Framework for Tactical Analysis and Individual Offensive Production Assessment in Soccer Using Markov Chains." https://www.nessis.org/nessis11/rudd.pdf.

Schober, Patrick, Christa Boer, and Lothar Schwarte. 2018. "Correlation Coefficients: Appropriate Use and Interpretation." *Anesthesia & Analgesia* 126 (February): 1. https://doi.org/10.1213/ANE.0000000000002864.

SciSports. 2013. "SciSports Website." https://www.scisports.com/.

———. 2024. "SciSports Events." https://support.scisports.com/en/articles/7937107-scisports-events.

Shaw, Laurie. 2020. https://github.com/Friends-of-Tracking-Data-FoTD/LaurieOnTracking/blob/master/Metrica_PitchControl.py/.

Singh, Karun. 2018. "Introducting Expected Threat (xT)." https://karun.in/blog/expected-threat.html.

Spearman, William. 2016. "Quantifying Pitch Control." February. https://doi.org/10.13140/RG.2.2.22551.93603.

———. 2018. "Beyond Expected Goals." March.

Spearman, William, Austin Basye, Greg Dick, Ryan Hotovy, and Paul Pop. 2017. "Physics-Based Modeling of Pass Probabilities in Soccer." March.

Whitmore, Johnny. 2023. https://theanalyst.com/eu/2023/08/what-is-expected-goals-xg/.

A Appendix

A.1 Expert data

#	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Mean	Median
1	3	4	5	4	2	3.6	4
2	5	5	1	5	5	4.2	5
3	2	2	3	2	2	2.2	2
4	1	3	2	2	1	1.8	2
5	3	3	2	1	3	2.4	2
6	2	4	3	2	2	2.6	3
7	2	4	3	2	3	2.8	2
8	2	4	3	3	1	2.6	3
9	3	4	3	3	2	3	3
10	3	5	4	3	3	3.6	3
11	4	5	4	5	4	4.4	4
12	2	3	2	4	3	2.8	3
13	2	2	3	3	2	2.4	2
14	3	4	4	4	3	3.6	4
15	2	1	3	4	2	2.4	2
16	3	1	2	2	2	2	2
17	2	4	1	2	4	2.6	2
18	4	4	4	4	3	3.8	4
19	5	5	2	5	5	4.4	5
20	4	2	4	5	4	3.8	4
21	3	4	2	4	5	3.6	4
22	4	5	5	5	4	4.6	5
23	2	3	1	2	3	2.2	2
24	2	2	2	2	3	2.2	2
25	3	2	4	3	4	3.2	3
26	4	5	4	5	4	4.4	4
27	4	1	3	4	5	3.4	4
28	4	5	4	5	5	4.6	5
29	3	4	4	4	3	3.6	4
30	3	3	3	4	4	3.4	3
31	3	5	2	3	4	3.4	3
32	2	4	2	3	4	3	3
33	4	5	4	5	5	4.6	5
34	2	4	4	4	4	3.6	4

Table 23: Expert ratings

A.2 Alternative passes

# Better alternatives	# Events
0	556
1	321
2	380
3	387
4	433
5	512
6	650
7	622
8	571
9	446

Table 24: Better alternatives

Player	Mean number of better events	Mean difference	# Passes
27	3.36	0.88	314
99	3.15	0.85	102
18	5.23	1.25	614
3	5.03	1.24	648
22	1.91	0.55	154
26	6.17	1.41	467
6	3.85	0.97	408
2	3.29	0.84	289
15	4.63	1.11	477
10	2.05	0.62	307
9	1.24	0.40	51
33	4.57	1.11	219
14	2.48	0.74	33
32	4.92	1.20	52
23	3.35	0.91	223
7	3.32	0.80	157
8	3.32	0.84	186
11	1.89	0.55	70
25	3.86	0.93	55
19	1.75	0.51	16
5	2.46	0.68	13
12	4.82	1.05	17

Table 25: Player evaluation whole season

A.3 Player rating

Jersey number	Type of player	FotMob rating	ECOM rating
26	Keeper	6.73	0.74
12	Keeper	6.88	0.69
3	Defender	6.91	0.95
33	Defender	6.61	0.73
2	Defender	6.77	0.62
15	Defender	6.81	0.98
18	Defender	6.87	1.04
32	Defender	6.02	0.7
25	Defender	6.80	0.44
7	Midfielder	6.98	0.38
8	Midfielder	6.89	0.71
27	Midfielder	7.00	0.99
99	Midfielder	6.77	0.97
23	Midfielder	6.86	0.51
6	Midfielder	6.89	0.70
5	Midfielder	6.1	0.37
11	Attacker	6.67	0.24
14	Attacker	6.96	0.78
22	Attacker	6.86	0.25
9	Attacker	5.80	0.14
19	Attacker	6.38	0.13
10	Attacker	7.37	0.46

Table 26: Player rating

#	**Sorted on FotMob**	**Type of player**	**Sorted on ECOM**	**Type of player**
1	10	Attacker	18	Defender
2	27	Midfielder	27	Midfielder
3	7	Midfielder	15	Defender
4	14	Attacker	99	Midfielder
5	3	Defender	3	Defender
6	6	Midfielder	32	Defender
7	8	Midfielder	14	Attacker
8	12	Keeper	26	Keeper
9	18	Defender	33	Defender
10	22	Attacker	8	Midfielder
11	23	Midfielder	6	Midfielder
12	15	Defender	12	Keeper
13	25	Defender	2	Defender
14	2	Defender	23	Midfielder
15	99	Midfielder	10	Attacker
16	26	Keeper	25	Defender
17	11	Attacker	7	Midfielder
18	33	Defender	5	Midfielder
19	19	Attacker	22	Attacker
20	5	Midfielder	11	Attacker
21	32	Defender	9	Attacker
22	9	Attacker	19	Attacker

Table 27: Player ranking