UTRECHT UNIVERSITY

Department of Information and Computing Science

Applied Data Science master thesis

Clustering passes from PEC Zwolle's goalkeeper

Applying Gaussian Mixture Models to football Event Data

First examiner:
Judith Verstegen

Candidate:
Matthijs de Koning

Second examiner:
Oliver Schmitz

In cooperation with:
PEC Zwolle

July 2, 2024

Abstract

Goalkeepers in the game of football have an unique role, as they are the only player on the field that are allowed to handle the ball with their hands. However, in recent years the game has shifted to where their ability to handle the ball with their feet, e.g. passing, has become just as important. But research regarding goalkeeper performance is scarce, especially in the context of passing the ball. This thesis attempts to fill this knowledge gap by investigating what type of passes a goalkeeper makes during a match. This is done by clustering passing *event data* by means of Gaussian Mixture Models. Here we show that the models could find a distinction between ground passes and passes made through the air. Furthermore, there is also an indication of the presence of more complex type of passes, however this could not be concluded for certain and needs further investigation. To improve the classification, it is suggested that the spatial-temporal data of all the players on the field during a match should be used to provide more relevant information that is missing in *event data*.

Contents

1. Introduction

The goalkeeper has a different involvement in the game of football compared to the other outfield players [1]. Most time of a match the goalkeeper is actively trying to defend the goal, and as a consequence, their performance have a direct impact on the result of the game [2]. However, it is found that not shot stopping qualities distinguish an elite level goalkeeper from an average goalkeeper, but their ability to pass the ball [3]. The importance of the goalkeeper being able to handle the ball with their feet is also reflected in their salary. Research found that passing accuracy mattered for goalkeeper pay [4].

By analysing the passes of a goalkeeper, trainers will be able to prepare more relevant training sessions and be more specific in their coaching. And hopefully, as a result of this, improve the performance of the goalkeeper during the game.

Currently, most passing metrics are designed to analyse outfield players (by quantifying the value of pass based on how much it improves scoring chances) and are not always relevant for goalkeepers [5], [6]. Furthermore, goalkeepers are often excluded from analysis and, in general, dedicated research on performance analysis of goalkeepers is scarce [7].

The overall aim of this thesis is therefore to contribute to performance analysis for goalkeepers. Specifically, this thesis aims to make a start in quantifying the way we look at goalkeeper passes.

We will be trying to do this by answering the following research question: What type of passes (i.e. long ball to winger, switching play or short goal kick, etc.) does a goalkeeper make during a game? And what factors characterise each type of pass? Factors such as the goalkeepers first touch, the amount of pressure on the ball and the effect of the event that happened before the pass will be quantified and investigated. The focus herein lies on classifying a pass and not on the value of a pass. To answer the first part of the research question clustering algorithms will be applied on event data. Event data quantifies what happens on the pitch in relationship with individual player actions e.g. tackles, shots, throw-ins, passes, etc. For the second part, the contextual factors above will be quantified and their values compared between clusters. Since this thesis is in collaboration with PEC Zwolle, their goalkeeper playing in the Dutch Eredivisie

throughout the 2023-2024 season will be analysed.

Similar research has been done on clustering pass sequences [8], [9] or classifying team playing styles [10], [11]. However, only a few studies on classifying individual passes were found. The research that was available considered classes for a pass based on the value it provided ("good", "ok", "bad") [12], or considered just 3 types of passes ("supportive", "penetrative", "backwards") [13]. This study will build upon the latter example by adding more types and shifting the context to the role of the goalkeeper. Identifying the clusters will be done together with expert opinion, e.g. the coaching staff.

Additionally, a study about Rugby used principal component analysis to look at the effect of match factors and performance indicators on the expression of playing styles [14]. This research will follow the same idea for the second part, but for types of passes instead of team playing styles.

2. Data

2.1 Description

This research used event data from PEC Zwolle's football matches in the Dutch Eredivisie during the 2023/2024 season. This data was provided by Opta [1]. Since the focus is on the passes made by PEC's goalkeeper, only events and information relevant to those passes were extracted and analysed. Furthermore, this study will *not* take into account the spatial-temporal data of all the other players on the pitch.

2.2 Event data

Event data consists of detailed information about all the actions that define a football match. I.e. passes, shots and fouls. But also when and where the ball goes out of play, how many stoppage time is added and information regarding player substitutions. In general, for each such event the location on the pitch, which player was involved, the outcome, and the time is recorded. Additionally, depending on the type of event, some contextual information is also available in so-called "qualifiers". Data in these qualifiers range from with what body-part a pass was made to the reason why a match is paused. The definitions of relevant events for this research can be found in table A.1 The main focus, however, is on the pass event itself.

2.2.1 Qualifiers

Each pass event also had qualifiers attached to it with contextual information. The definition of these qualifiers can be seen in table A.2. The data regarding the distance, angle and coordinates of the end location of the pass are contained in qualifiers and are always available for a pass event. Furthermore, it is possible that one event has multiple qualifiers. Also some qualifiers are a subcategory of another qualifier, both the subcategory and the overarching category are than available as qualifier for that specific event. For example, this occurs with the "over-arm" and "throw" qualifier, since the over-arm throw is a subcategory of a throw.

[1] https://www.statsperform.com/opta/

Feature	Description
x_{end}	X-coordinate of the end location of a pass
y_{end}	Y-coordinate of the end location of a pass
x_{origin}	X-coordinate of the start location of a pass
y_{origin}	Y-coordinate of the start location of a pass
$distance$	The distance the ball has travelled during a pass in meters
$angle$	Defined as the angle of the pass relative to the direction of play in radians

Table 2.1: Definitions of the pass features

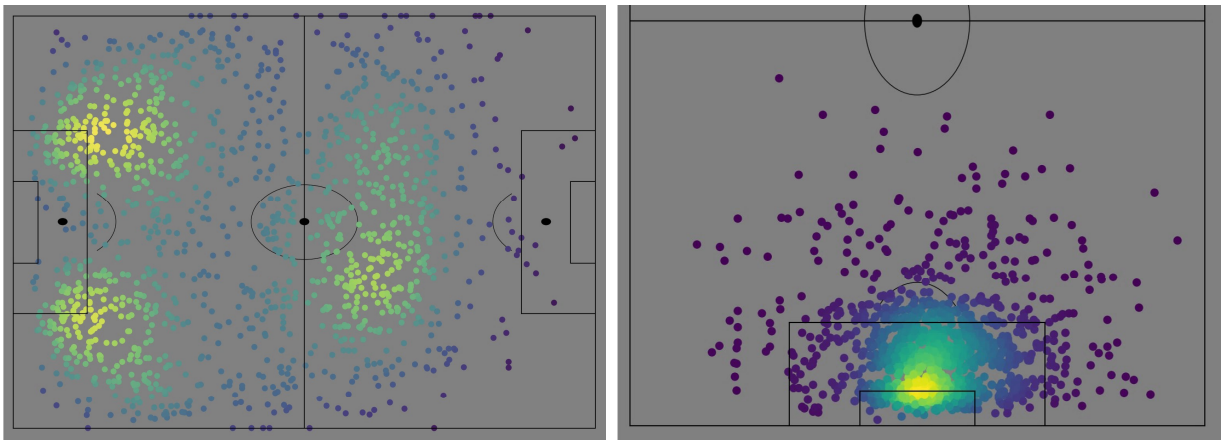
2.2.2 Collecting and quality

Event data is automatically collected from video streams of a match. Humans assist these algorithms by checking for their correctness and they add extra (contextual) information where necessary. Especially when labelling the passes with qualifiers. From Opta's website we read: "Event data is generated in real-time though a combination of human annotation, computer vision and AI modelling." The data is stored in .xml files and can be downloaded per match. Quantitative information regarding the accuracy of the data is not provided.

2.3 Pass features

A pass is defined as an attempt by one player to play the ball to a teammate. This includes set pieces like a free-kick or a throw-in. A pass can either be successful when it reaches a teammate or unsuccessful when it does not. The features that define a pass are shown in table 2.1. The cluster algorithm will try to find clusters based on combinations or edited variants of these features. This will be further elaborated in section 3. The range of values the coordinates of a pass can take is [0,100] for both the x- and y-axis.

The goalkeeper made 1355 passes spread over 32 games with a success percentage of 68.1%. Figure 2.1 shows the end and starting points of each pass. From 2.1a, it can be seen that there are roughly three areas where the goalkeeper passed the most to. Two relatively circular shaped clouds of points on either side of the penalty box, and one oblong shaped cloud of passes further up the field that is skewed to the right. Most of the starting points of the passes (see figure 2.1b) are inside the penalty area. The further the goalkeeper goes from goal, the fewer amount of passes he has given.



(a) End locations of a pass **(b)** Start locations of a pass

Figure 2.1: Every pass form the goalkeeper in the 2023/2024 season. The direction of play is from left to right. Brighter colours indicate a higher density of points.

2.4 Preparation of the data

From these qualifiers, nine pass categories were defined based on what they represent and how often they occurred. Their names, definitions and fraction in the dataset can be seen in table 2.2. These categories can than be used in evaluating the characteristics of each cluster.

Furthermore, two additional pass features have been created by subtracting the origin x- and y- coordinates of a pass from the end x- and y- coordinates of a pass, respectively. As a result, every pass is than in the same reference frame (all passes originate from (0,0)), which makes it easier to compare the different type of passes based on shape (distance and angle) instead of on the location on the pitch.

Next, to be able to compare the passes some more (mainly contextual) features were extracted from the match event data. Those features are as follows: the event before and the event after a pass, the minute of the game a pass was given, the goal difference at the time of the pass and the outcome. Also the pass features of the previous pass (if the previous event was a pass), the time on the ball (defined as the the difference between the previous event and the time the goalkeeper passed the ball), speed of play (the average time of events from the 5 events before and 5 events after the pass), the location of the goalkeeper's pass in a pass chain and the result of the pass chain in which the goalkeeper's pass was a part of are all extracted. A pass chain is a collection of successive passes.

Finally, the pass features were scaled to a range of [0,1] based on their minimal and maximum values before they are used in the clustering algorithms.

Name	Definition	Fraction
launched long ball	Pass played from a player's own half up towards front players. Aimed to hit a zone rather than a specific player. Defined as the passes having the "long ball" and "launch" qualifier.	22.0%
goal kick	Goal kick through the air. Contains passes with the "goal kick" qualifier, but not the "low" qualifier.	9.5%
chipped long ball	Pass which was chipped into the air (short pass through the air). Defined as the passes with the "long ball" qualifier and not having the "launch" qualifier.	14.2%
throw	Goalkeeper throws the ball. Defined as all the passes with the "throw" qualifier, excluding those with the additional "over arm" qualifier.	11.6%
low goal kick	Indicates a ground pass goal kick. All passes with the "low" qualifier.	8.2%
volley	Goalkeeper kicks the ball forward straight out of his hands. Contains all passes with the "Gk kick from hands" qualifier.	6.1%
over-arm	Indicates over-arm throw by the goalkeeper. All passes with the "over arm" qualifier.	1.9%
hoof	Goalkeeper drops the ball on the ground and kicks it long towards a position rather than a specific player. All passes with the "gk hoof" qualifier.	1.0%
in play	All other passes that were not included in the above categories. The bulk consists of ground passes.	25.5%

Table 2.2: Names and definitions of the nine pass categories used in this thesis

3. Method

3.1 Overview

The method of this thesis can be summarised as follows. First, .xml files containing the event data of PEC Zwolle's matches in the 2023/2024 season were downloaded. From this, the pass events of the goalkeeper were extracted. In particular data about the pass features, the relevant qualifiers and information necessary for context. This raw data was than preprocessed by constructing contextual features, creating pass labels and scaling the pass features. Next, Gaussian Mixture Models were trained on combinations of the pass features to find the optimal number of clusters. The optimal number of clusters was decided based on i) metric scores and ii) domain knowledge. After the clusters were chosen, the occurrences of pass labels and the values of the contextual features were compared between these clusters.

3.2 Gaussian Mixture Model

3.2.1 Motivation

In this thesis the Guassian Mixture Model (GMM) [15] was used to cluster the goalkeeper's passes. A GMM can be seen as a generalisation of the K-means algorithm. With both algorithms the user sets the number of clusters. However, where the K-means algorithm assumes spherical clusters and favours similar number of observations in each cluster [1], the GMM is able to find clusters of different shapes and cluster sizes. This is its main advantage, because it is assumed that the type of passes during a match also have different shapes and sizes, i.e. a pass to the centre-back occurs much more often than a pass to a central midfielder.

Another benefit of using a GMM instead of the K-means algorithm is its ability for soft classification. Whereas, K-means assigns an observation to one specific cluster, leading to hard cut-offs of cluster boundaries with outliers having a large impact on the final result, the GMM assumes that the observations are generated from a mix of different

[1] https://www.geeksforgeeks.org/demonstration-of-k-means-assumptions/

Gaussian distributions. Consequently, it assigns a probability for each observation to belong to a certain cluster. This is also useful, because distinguishing passes in a game is not always that clear-cut, e.g. due to a bad touch or variation inherent to a pass activity.

Finally, the drawback of being able to distinguish clusters of unequal sizes and shapes is that the GMM has more parameters to estimate and as a result tends to overfit with small datasets.

3.2.2 Model specification

Formally, a GMM is a weighted sum with mixture weights w_i of M Gaussian distributions defined by mean vectors μ_i and covariance matrices Σ_i. Where $i = 1, \ldots, M$. Each Gaussian distribution is a specific cluster, or sometimes also called a mixture. To estimate the optimal values of the parameters the GMM is trained by means of the Expectation-Maximisation (EM) algorithm [16]. It is possible to create variants of this model by demanding a certain shape of covariance matrix. For example, one can have the constraint that each mixture has the same covariance matrix, as a result clusters would have the same shape. The mixture weights indicate the probability of a random observation to belong to cluster i, they have the constraint that $\sum_{i=1}^{M} w_i = 1$. Intuitively, the weights give a rough estimation of the size of each cluster in the model. The cluster centres are given by the mean vectors. w_i, μ_i and Σ_i fully define a GMM and these parameters will be optimised.

3.3 Clustering

In this thesis a GMM is used where each cluster has its own covariance matrix. Because it is unknown what model will work best, it is decided to use the most general version. To speed up and improve performances of the EM algorithm the initial guess for the weights, the means and the covariance matrices is provided by a K-means algorithm.

3.3.1 Performance metrics

To find the number of clusters that best splits up the data based on performance metrics a GMM is trained for $M = 2, \ldots, 15$. Both from expert opinion and by testing a handful of GMMs with larger values for M it is concluded that it is unlikely more than 15 types of passes exist in the dataset. To speed up computations, $M = 15$ is therefore chosen as cutoff. The GMM is initialised by a K-means algorithm that uses random starting points,

therefore variation exists in the final convergence locations of the GMMs. To reduce this uncertainty, ten GMMs were trained for each M and their performances were averaged.

These models were trained on combinations of the features shown in table 2.1.

To evaluate the performance of each model the Silhouette score, the Davies-Bouldin index and the Calinski-Harabasz index were calculated. The Silhouette score and Calinski-Harabasz index are similar as both measure the distinctness of clusters by quantifying within-cluster distances and between-cluster separation. The former can take on values between -1 and 1, and the latter is unbounded. For both methods higher scores indicate more compact and well-separated clusters. Alternatively, the Davies-Bouldin index measures the similarity of clusters. Lower values indicate better clustering results. As there is no standard metric for this specific problem, multiple metrics have been used to create a more complete picture of the performance of the models.

3.3.2 Domain knowledge

The other method of choosing the number of clusters is based on domain knowledge. This has been done in consultation with the coaching staff. Here it has been chosen to analyse six clusters clustered from a GMM trained on the modified end coordinates (end coordinate minus the coordinates of the start of the pass) of a pass, the distance and the angle of the pass. These four features have been chosen, because of a trade off between including the most information and, at the same time, minimising the dimensionality (number of features) of the dataset for better performances. These six clusters are chosen in such a way that they roughly represent actual zones on the pitch a goalkeeper would pass to. It is known that there exist more than six type of passes, but by considering more, the analysis of comparing the clusters would become too hard as not enough samples would be present in each cluster.

3.4 Comparing clusters

After clustering the data on the pass features, the resulting clusters will be evaluated on how the contextual features and the pass categories (constructed from the qualifiers) are distributed within each cluster. This gives an idea about what type of pass each cluster represents. For each cluster, summary statistics are calculated (like the mean, standard deviation, min/max values and size of the cluster) and the distribution of the values of each contextual feature are visualised. This is done for the following: *time on the ball*, *speed of play*, *minute of the game*, *the event before the pass*, *the event after the pass*, *the*

event in which the pass chain ended, the start and end coordinates of the pass to the goalkeeper if the previous event was a pass, the direction of the goalkeeper's first touch and the distance of the goalkeeper's first touch. The continuous variables are compared with boxplots and histograms. The distribution of categorical variables are visualised with stacked bar charts. Lastly, it is calculated how the nine pass categories form table 2.2 are spread over the clusters.

3.5 Implementation

Analysis and visualisation has been done in Python 3.11 [17]. The Scikit-learn [18] package has been used for: building and applying the GMMs, for scaling the pass features in the preprocessing step and for applying performance metrics. The Matplotlib library [19] has been used to create visualisations. The NumPy [20] and Pandas [21] packages were used for data manipulation.

4. Results

First the clustering results are shown, then the results from analysing the characteristics of each cluster. For both the clustering results and the part about the characteristics of these clusters the method of determining the best number of clusters based on metric scores is treated first. Furthermore, the names of colours on the y-axis and the colours of the boxes in the boxplots refer to the clusters in figure 4.1 and figure 4.2.

4.1 Clustering

4.1.1 Based on metric scores

The clustering results on the combinations considered in this research are shown in table A.3, A.4 and A.5.

The best partitioning of the data was from GMMs trained on the "distance" and "angle" features, the most optimal number of clusters is three according to all three metrics. The silhouette score had a value of 0.5686 +/- 0.0004, the Calinski-Harabasz index was equal to 2216 + 0.03 and the Davies-Bouldin score was equal to 0.47 + 0.004. The models all converged to the same maximum, hence the extremely small errors. For larger number of clusters the model found it harder to find a maximum and as a result the standard deviations were larger as well.

Therefore 3 clusters has been chosen based on the "distance" and "angle" pass feature combination. These clusters are shown in figure 4.1.

The success percentage of the passes in the blue (n=691), green (n=333) and red (n=331) clusters are 39.8%, 97.9% and 97.3% respectively.

4.1.2 Based on domain knowledge

The passes based on the method of choosing the number of clusters with domain knowledge are shown in figure 4.2.

The success percentage of the passes in the blue (n=322), green (n=67), black (n=280), red (n=255), yellow (n=208) and purple (n=223) clusters are 31.0%, 94.0%, 98.2%, 98.4%, 53.8% and 54.7% respectively.

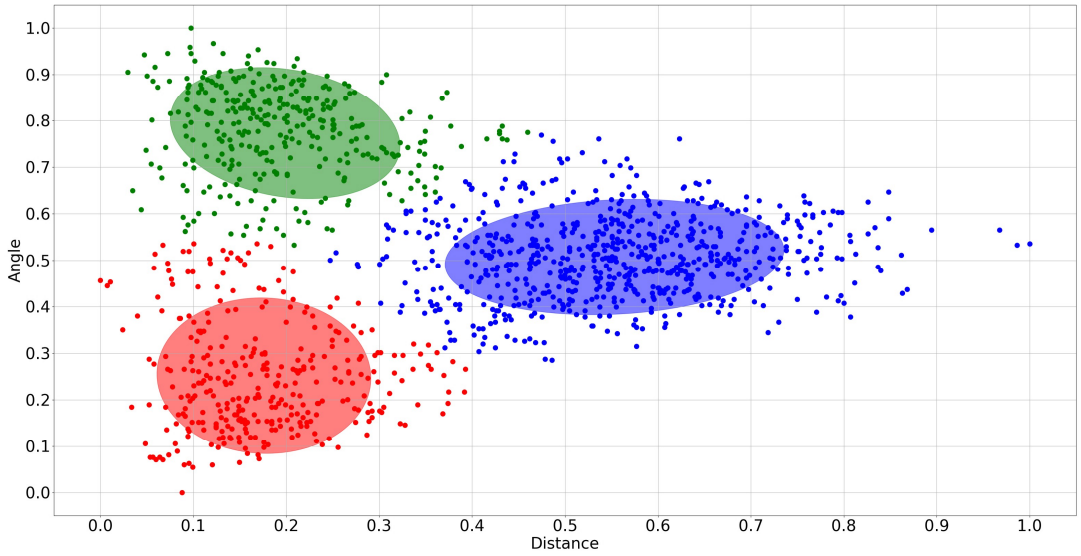


Figure 4.1: Three clusters found by a GMM trained on the "distance" and "angle" pass features. The ellipses indicate the shape of the underlying Gaussian distributions. The axis are scaled to a range of [0,1] based on the minimum and maximum values of that axis respectively. A high value of the angle means that the goalkeeper passed to his left, a low value of the angle means a pass to his right.

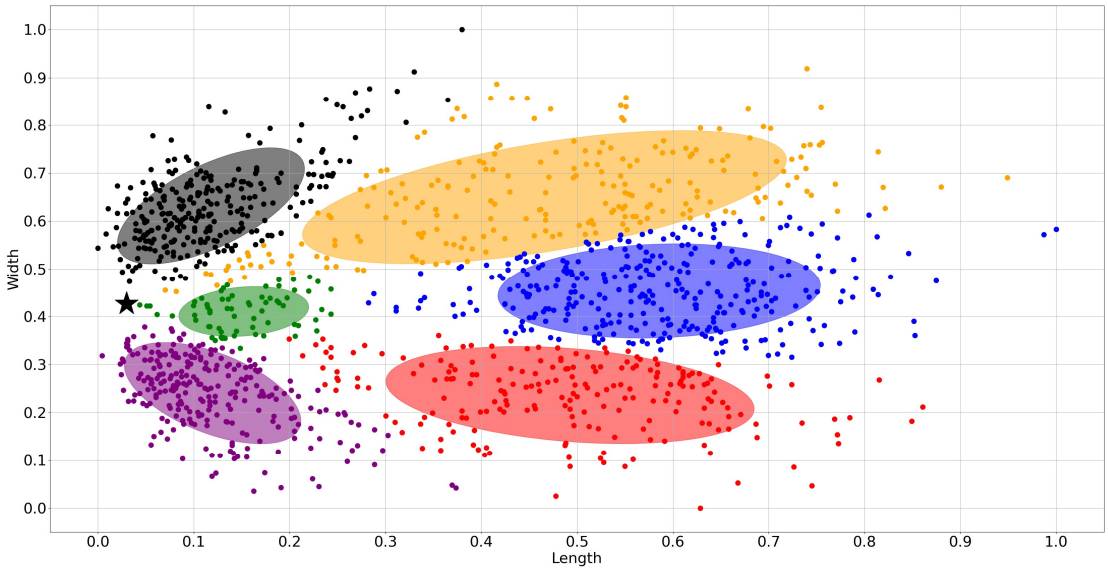


Figure 4.2: Six clusters from a GMM model trained on the end coordinates, distance and angle pass features. The ellipses indicate the shape of the underlying Gaussian distributions. The star indicates the goalkeeper. The axis are scaled to a range of [0,1] based on the minimum and maximum values of that axis, respectively.

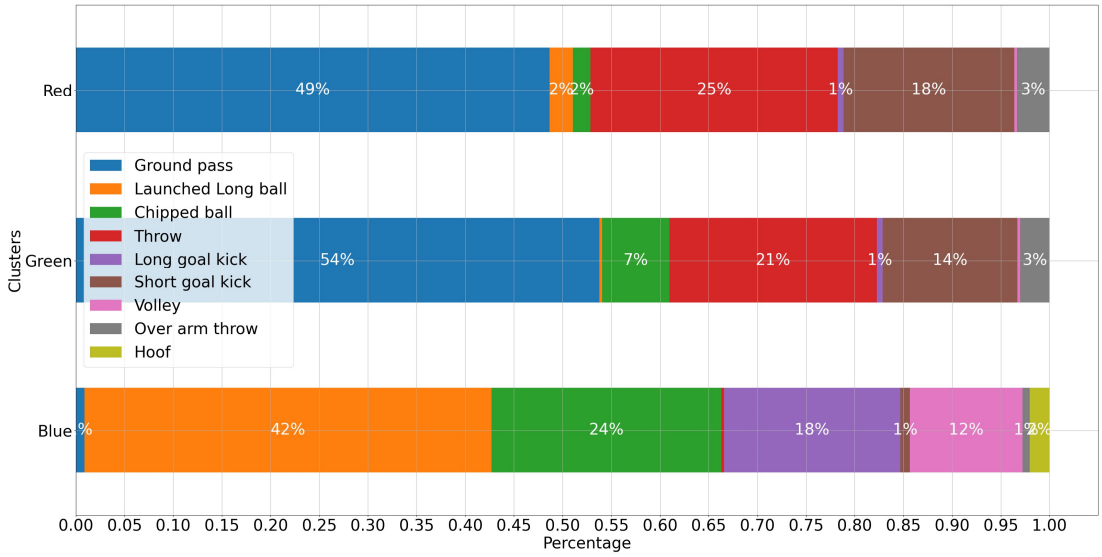


Figure 4.3: The proportion of pass categories for each cluster.

4.2 Analysis of contextual features of clusters

4.2.1 Optimal metric score clusters

4.2.1.1 Pass categories

In figure 4.3 it can be seen that the red and green clusters have a similar proportion of pass categories. The most striking result is the difference in percentages of pass categories launched long ball, chipped ball, long goal kick and volley between the blue and red/green clusters.

4.2.1.2 Characteristics of long balls and ground passes

In figure 4.4 it can be seen that the blue cluster has a large portion of pass chains compared to the other two clusters that end in an aerial, e.g. a header duel. The take-on event is not present in the blue cluster, but is present in the red and green ones.

Note that in figure 4.5 it can be seen that passes in the blue cluster again have a larger proportion of aerials compared to the red and green clusters. When passing to the latter clusters almost all (roughly 92%) events immediately after are also a pass.

4.2.1.3 Switch of play characteristic

Figures 4.6 and 4.7 are concerned about the location of the ball before the goalkeeper receives the ball and how the goalkeeper handles the ball. Specifically, in figure 4.7 it

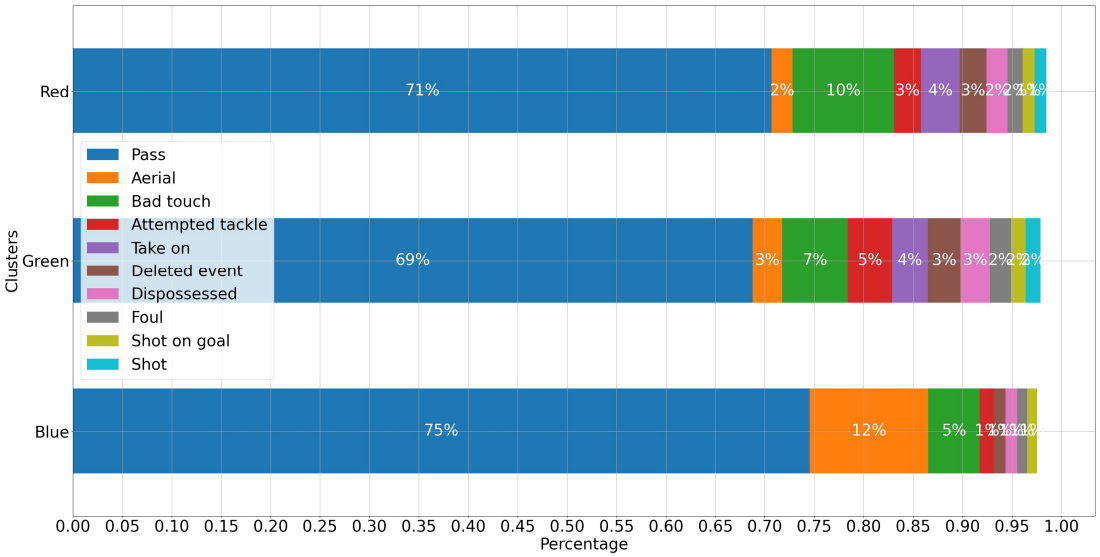


Figure 4.4: The distribution of events in which a pass chain ended that the goalkeeper was a part of.

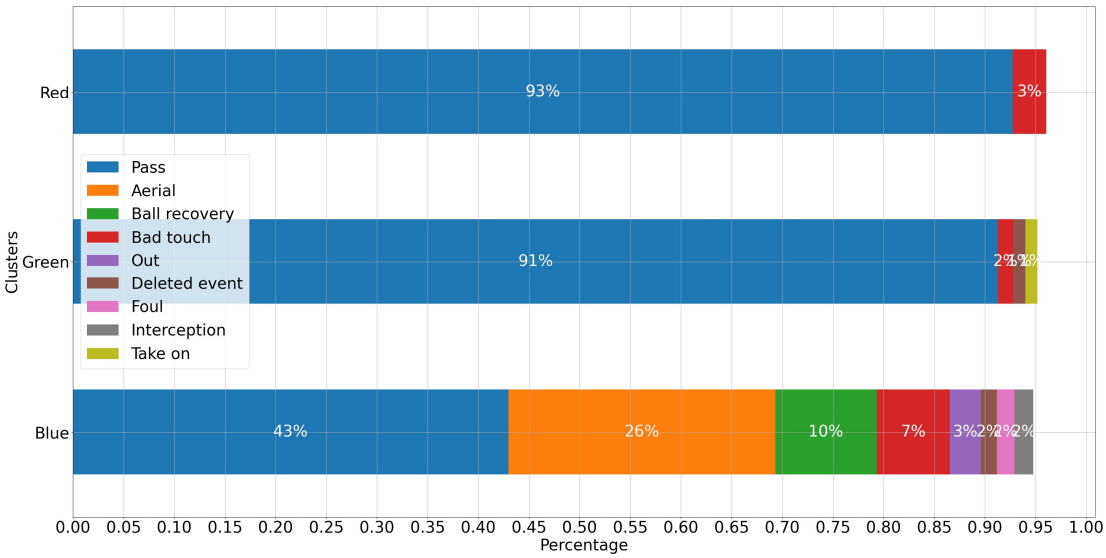


Figure 4.5: The distribution of events right after the goalkeeper made a pass for each cluster.

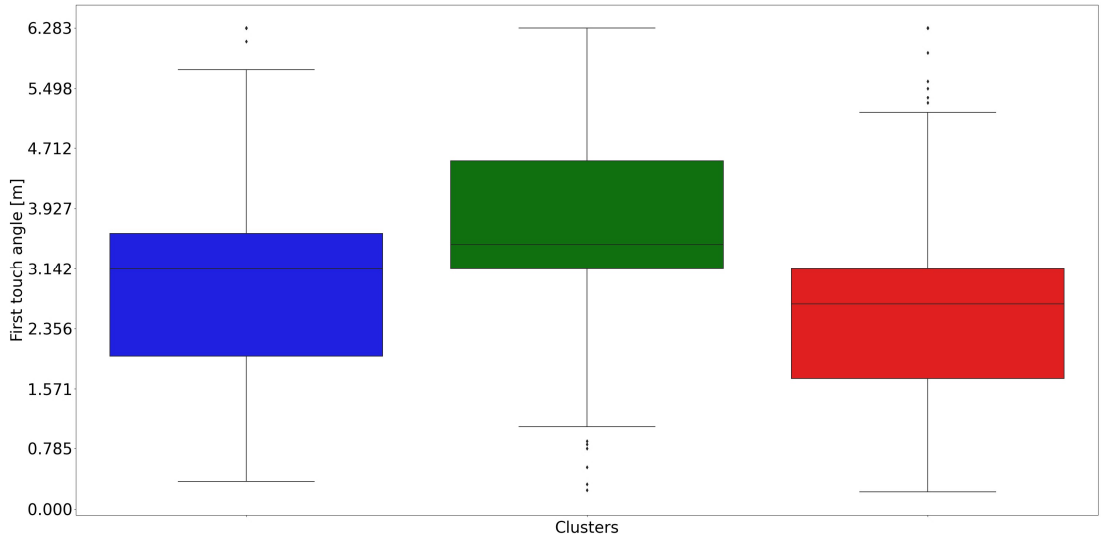


Figure 4.6: The direction in which the goalkeeper makes his first touch with the ball. Values larger than π indicate moving the ball to his left, values smaller than π indicate moving the ball to his right.

can be seen that the goalkeeper received more passes from locations with y-coordinates smaller than fifty when passing to the green cluster. The inverse can be seen for passes to the red cluster. Also note how in 4.6 it can be seen that the goalkeeper's first touch angle is larger than π for the green cluster and smaller than π for the red cluster. The same characteristics can be noted for the black and purple clusters in the six cluster classification. This can be seen in figures A.1 and A.2 in the appendix.

4.2.2 Six clusters based on domain knowledge

4.2.2.1 Green cluster is different

Figure 4.8 shows the distribution of the pass categories for the domain knowledge clusters. Note the similar distribution between the purple, black and green clusters and the yellow, red and blue clusters. Also not how the these two groups look similar to the red/green group and blue group in figure 4.3, respectively. Furthermore, it can be seen that the green cluster does not have chipped ball passes and the largest percentage of throws.

All clusters in figure 4.9 look similar at first eye. However, the green cluster has a smaller percentage of the out event before a pass and a larger percentage of ball recovery compared to the other clusters. Together with the purple cluster the green cluster has the largest percentage of keeper pick-up events before a pass.

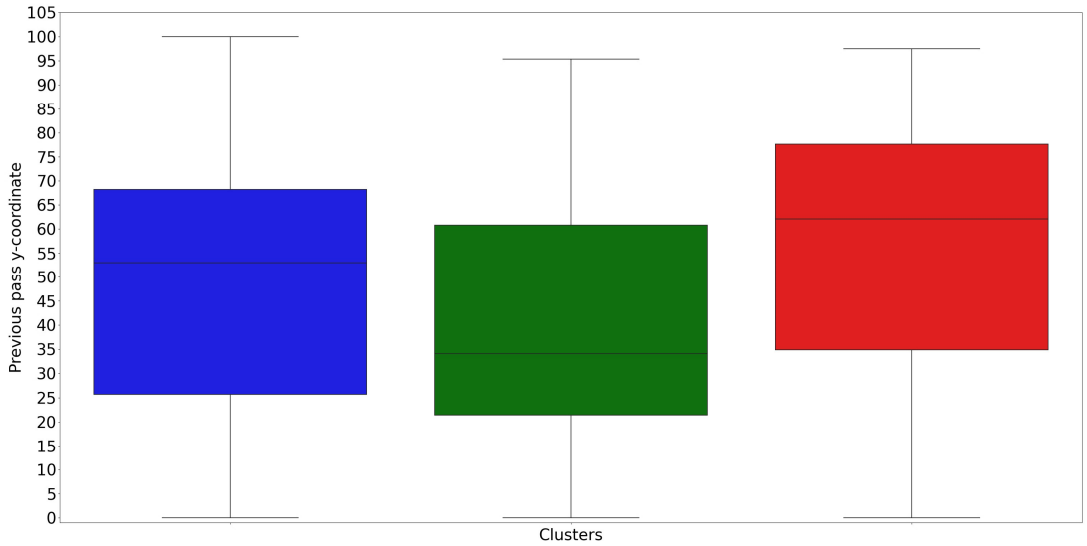


Figure 4.7: Distribution of the values of the y-coordinates of the pass to the goalkeeper. The y-coordinate indicates the location along the width of the field. Values range between [0,100].

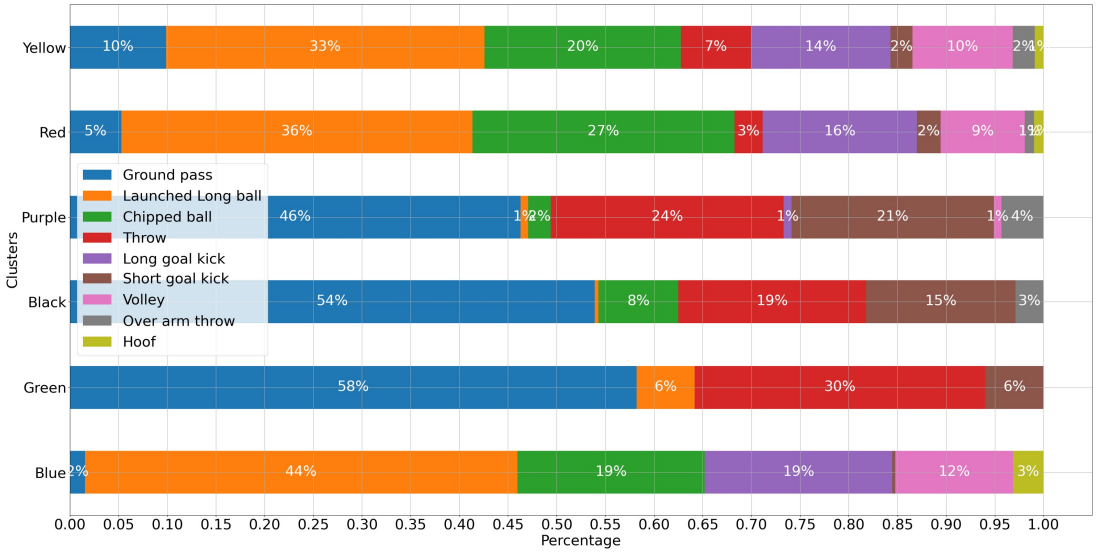


Figure 4.8: The distribution of pass categories for each cluster. The colors on the y-axis refer to the clusters in Figure 4.2.

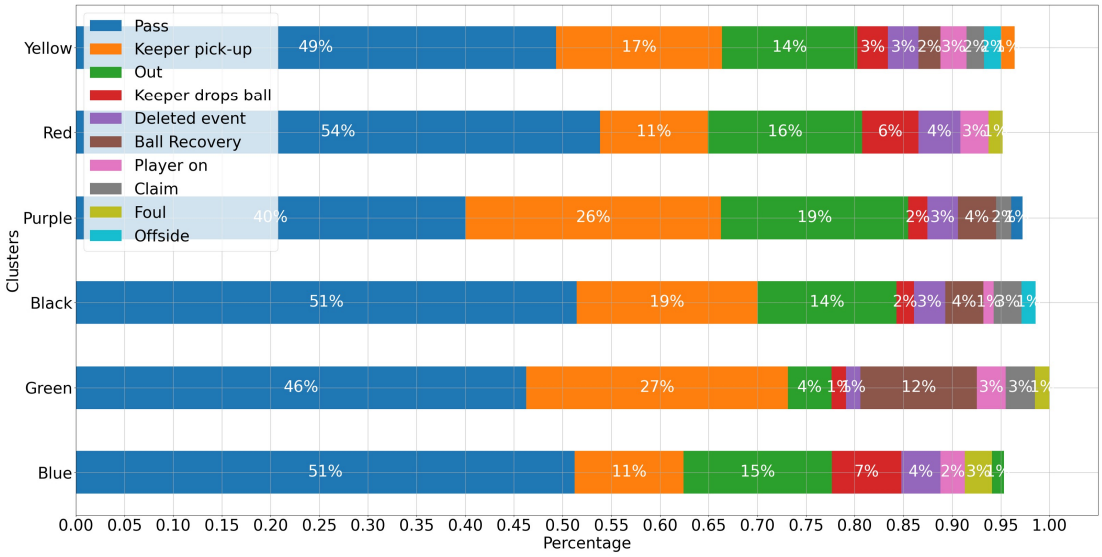


Figure 4.9: The proportion of event types before the goalkeeper made a pass for each cluster.

4.2.2.2 Time related characteristics

Note, in figure 4.10 the shallow distribution of how long the goalkeeper had the ball when he passed to the green cluster.

From figure 4.11 it can be seen that the red, blue and yellow clusters are slightly more skewed to later in the game compared to the green, black and purple clusters.

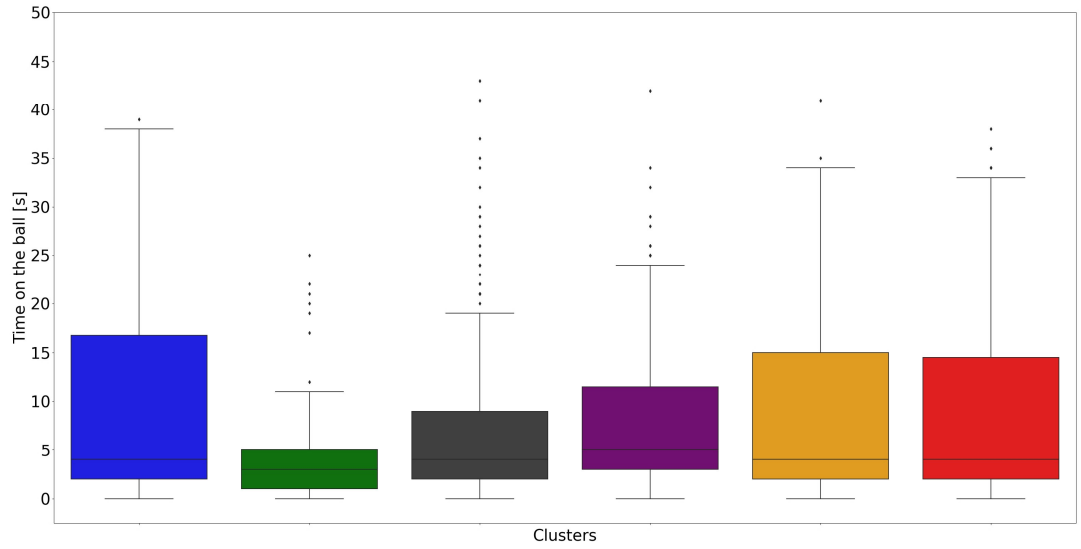


Figure 4.10: Distribution of how long the goalkeeper had the ball before he made a pass.

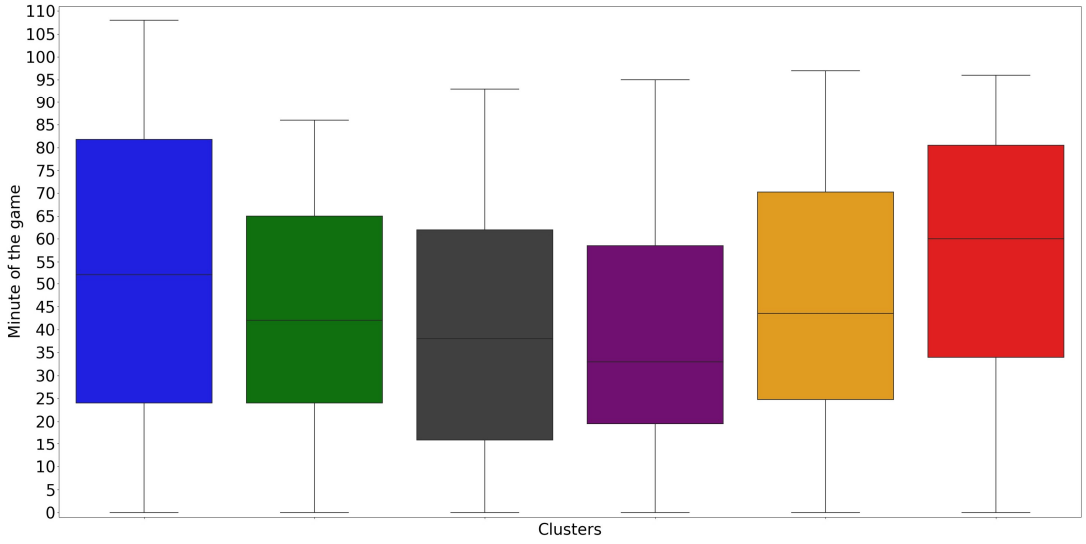


Figure 4.11: Distribution of the minute of the game a goalkeeper made a pass.

5. Conclusion

The results suggest that human judgement is better able to determine pass types than a GMM trained on pass features and tested with the Silhouette score, the Davies-Bouldin index, and the Calinski-Harabas. This is concluded based on the detail of the pass categories that were constructed, but the lack of this detail in the number of clusters the GMM found.

Furthermore, it can be concluded that it is, however, possible for a GMM to distinguish passes made through the air from ground passes. This could be inferred from the event types surrounding a pass and the nine pass categories.

Evidently, the contextual features considered in this research were not able to significantly distinguish between clusters. But there are hints that more complex pass types are present, like a switch of play pass or the passes in the green cluster in the six cluster case.

5.1 Discussion

Figure 4.1 implies quite clear that the red and green clusters consist mostly of short distance passes, most probable directed to the goalkeeper's centre backs who are on either side of the goalkeeper. The blue cluster consists of longer distance passes. The types of passes in these clusters also align with this, as seen in figure 4.3. Here the red and green clusters consist of the ground pass type, throws and short goal kicks. In contrast, the blue cluster mostly consist of passes through the air, i.e. launched long ball, long goal kicks, hoofs and volleys. The argument that the blue cluster passes go through the air can be further supported with the results in figure 4.5, where 26% of the events after the pass in the blue cluster consists of a header duel. In contrast with the red and green clusters where the aerial event is not present at all. Finally, from experience it is already known that long balls have a lower success percentage. This is confirmed with the 39.8% success percentage in the blue cluster (compared to the 97% success percentage in the other two clusters). Figure 4.4 hints at a reason for this. The aerial type is more often the end event for a pass chain when passing to the blue cluster, indicating that the pass becomes contested with a header duel, the duel however results

in loss of possession, ending the pass chain.

Another interesting result comes from figures 4.6 and 4.7. The goalkeeper tends to move the ball with his first touch to the area where he will pass to. On top of this, the direction of his first touch is more often opposite to the location where the ball came from. The combination of these results indicate that the goalkeeper switches the play, e.g. moves the ball from one side to the other side of the pitch to relieve pressure inflicted by the opposing team or to find an opening for attack. The same characteristic can be seen for the black and purple clusters in figures A.1 and A.2.

From figure 4.8 it can be inferred that the six clusters are a more detailed version of the three pass clusters in 4.3. The yellow, red and blue clusters in figure 4.2 roughly correspond to the blue cluster in figure 4.1.

However, the main difference here being that the green cluster for the six cluster case is different. The green cluster passes are "in-play", meaning that they are not from a goal-kick or free-kick. As can be seen from figure 4.8 by not having the goal-kick pass category, from the smallest percentage of *out* and largest percentage of *ball recovery* event types before the pass in figure 4.9, and the shortest time on the ball in figure 4.10. A goalkeeper would either give that type of pass when they are dominant on the field (without any pressure from the opposition) or when the goalkeeper tries to play out from the back when there is a certain type of pressure from the enemy team. Furthermore, there being no chipped balls (medium distance precision pass through the air) reinforces this, as those are risky passes that one normally wouldn't make in a central, crowded and vulnerable part of the field.

Finally, there is a phenomena in football where long distance passes are used when a team needs to score a goal to equalise or win, but the match is coming to an end. The idea is that by just launching the ball in the opposition box and trusting on luck that the ball ends up in front of a teammate who can then shoot the ball in the goal. Figure 4.11, shows this by the extent in which the blue boxplot reaches over into additional time. This could signifying that the goalkeeper kicked long balls late in the game, presumably to try for a late game equaliser or win.

In this research the main shortcoming is the absence of spatial-temporal information of the players on the field. When clustering only on raw pass features (beginning and end coordinates of a pass), it is not considered whether there is a teammate there or not. The same goes for whether there is space for the teammate to control the ball, or if the pass lane is blocked by an opposition player, and if it is in line with agreements made in the game plan. Also the type of teammate is important; you wouldn't pass a long ball to

a smaller teammate to force him into a header duel. Arguably, these are more important aspects of determining what type of pass to give than knowing for example what event came before a pass. Event data will therefore have an incomplete picture and will only give a rough indication of what type of pass there could have been.

One way this spatial-temporal information of players could be used is by quantifying the value of a pass and then cluster based on the value of said pass (or by taking the value of a pass into account). In this research, every pass had the same weight when clustering, but a "wrong" pass should have a lower weight when trying to classify it (or perhaps put it in its own "wrong pass" cluster).

Furthermore, in the six cluster scenario, together with the nine pass categories, there where on average 1355/6/9 = 25 passes per pass category per cluster. Since the ground pass and long ball pass categories had larger sample sizes compared to the other pass categories. The percentages in the stacked bar chart plots of these small sample events or pass categories would therefore fluctuate heavily if it contained one pass more or one pass less. This makes it hard to say anything definitive about it.

Ultimately, in this thesis only one goalkeeper for one season has been considered. A reason for why the results are very basic could be because the data is too sparse for an algorithm too detect rarer pass types. We already saw the great disproportions between the nine pass categories (ground passes occur, for example, much more often compared to hoofs). Increasing the size of the data by adding passes from other goalkeepers and more matches might help in detecting these rare pass types.

A. Appendix A

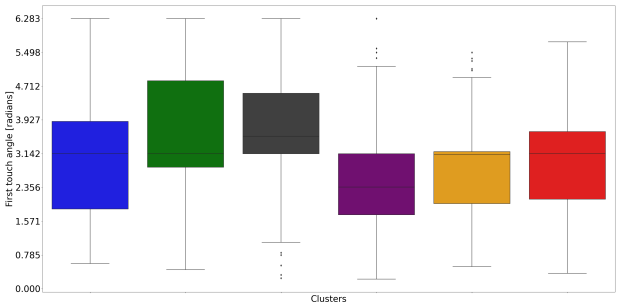


Figure A.1: The direction in which the goalkeeper makes his first touch with the ball. Values larger than π indicate moving the ball to his left, values smaller than π indicate moving the ball to his right.

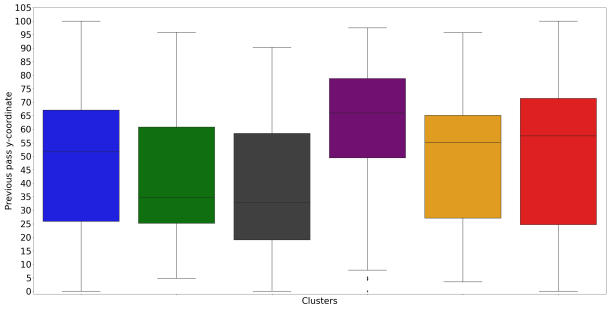


Figure A.2: Distribution of the values of the y-coordinates of the pass to the goalkeeper. The y-coordinate indicates the location along the width of the field. Values range between [0,100].

Name	Definition
Pass	The attempted delivery of the ball from one player to another player on the same team. A player can use any part of their body (permitted in the laws of the game) to execute a pass.
Aerial	Two players from opposing teams contest an aerial ball.
Bad touch	Used when a player makes a bad touch on the ball and loses possession. Could be either because the ball hit the player unintentionally or because the player unsuccessfully controlled the ball.
Attempted tackle	A player has attempted to tackle an opponent but has been unsuccessful.
Take on	Attempted dribble past an opponent.
Deleted event	The event has been deleted.
Dispossessed	Player is successfully tackled and loses possession of the ball.
Foul	Indicates a foul has been committed.
Shot on goal	Shot on goal saved by the goalkeeper.
Shot	Any shot on goal which goes wide or over the goal.
Ball recovery	A player gathers a loose ball and gains control of possession for their team.
Out	Indicates the ball has gone out of play for a throw-in or goal kick.
Interception	When a player intercepts any pass event between opposition players and prevents the ball reaching its target.
Keeper pick-up	Goalkeeper picks up the ball.
Keeper drops-ball	A goalkeeper drops the ball close to himself with the intention to play it with his feet.
Player on	Player comes on as a substitute.
Claim	Goalkeeper catches a crossed ball.
Offside	An offside decision is given against an attacker.

Table A.1: Event type definitions

Name	Definition
Long ball	Intended long balls
Launch	Pass played from a player's own half up towards front players. Aimed to hit a zone rather than a specific player.
Goal kick	Pass event - Goal kick.
Low	Indicates a ground pass goal kick.
Throw	Pass event - Goalkeeper throws the ball.
Over-arm	Indicates over-arm throw out by the goalkeeper.
Gk kick from hands	Goalkeeper kicks the ball forward straight out of his hands.
Gk hoof	Goalkeeper drops the ball on the ground and kicks it long towards a position rather than a specific player.

Table A.2: Qualifier definitions.

Feature combination	Silhouette score	Standard deviation	Clusters
(x_{end}, y_{end})	0.4519	0.0003	4
(x_{start}, y_{start})	0.4645	0.0002	2
(distance, angle)	0.5686	0.0004	3
$(x_{end}, y_{end}, x_{start}, y_{start})$	0.27	0.01	2
$(x_{end}, y_{end}$, distance, angle)	0.4660	0.0002	3
$(x_{start}, y_{start}$, distance, angle)	0.383919	0.000003	3
$(x_{end}, y_{end}, x_{start}, y_{start}$, distance, angle)	0.3563	0.0002	3

Table A.3: Best clusters based on silhouette scores.

Feature combination	Index value	Clusters
(x_{end}, y_{end})	1521	4
(x_{start}, y_{start})	722	14
(distance, angle)	2216	3
$(x_{end}, y_{end}, x_{start}, y_{start})$	536	2
$(x_{end}, y_{end}$, distance, angle)	1312	3
$(x_{start}, y_{start}$, distance, angle)	765	3
$(x_{end}, y_{end}, x_{start}, y_{start}$, distance, angle)	766	3

Table A.4: Best clusters based on Calinski-Harabasz index.

Feature combination	DB Score	Standard deviation	Clusters
(x_{end}, y_{end})	0.7246	0.0003	4
(x_{start}, y_{start})	0.838	0.008	14
(distance, angle)	0.564301	0.0004	3
$(x_{end}, y_{end}, x_{start}, y_{start})$	1.0967	0.0005	3
$(x_{end}, y_{end}$, distance, angle)	0.7433	0.0002	3
$(x_{start}, y_{start}$, distance, angle)	0.994283	0.000003	3
$(x_{end}, y_{end}, x_{start}, y_{start}$, distance, angle)	1.0373	0.0002	3

Table A.5: Best clusters based on Davies-Bouldin score.

Bibliography

[1] F. Otte, T. Dittmer, and J. West, "Goalkeeping in modern football: Current positional demands and research insights," *International Sport Coaching Journal*, vol. 10, pp. 112–120, 1 Jul. 2022, ISSN: 2328-9198. DOI: 10.1123/ISCJ.2022-0012. [Online]. Available: https://journals.humankinetics.com/view/journals/iscj/10/1/article-p112.xml.

[2] J. West, *A review of the key demands for a football goalkeeper*, Dec. 2018. DOI: 10.1177/1747954118787493.

[3] M. Jamil, A. Phatak, S. Mehta, M. Beato, D. Memmert, and M. Connor, "Using multiple machine learning algorithms to classify elite and sub-elite goalkeepers in professional men's football," *Scientific Reports*, vol. 11, 1 Dec. 2021, ISSN: 20452322. DOI: 10.1038/s41598-021-01187-5.

[4] D. Berri, D. Butler, G. Rossi, R. Simmons, and C. Tordoff, "European sport management quarterly issn: (print) (salary determination in professional football: Empirical evidence from goalkeepers," 2023. DOI: 10.1080/16184742.2023.2169319. [Online]. Available: https://www.tandfonline.com/action/journalInformation?journalCode=resm20.

[5] F. Goes, E. Schwarz, M. Elferink-Gemser, K. Lemmink, and M. Brink, "A risk-reward assessment of passing decisions: Comparison between positional roles using tracking data from professional men's soccer," *Science and Medicine in Football*, vol. 6, pp. 372–380, 3 2022, ISSN: 24734446. DOI: 10.1080/24733938.2021.1944660.

[6] T. Decroos, L. Bransen, J. V. Haaren, and J. Davis, "Actions speak louder than goals: Valuing player actions in soccer," vol. 11, 2019. DOI: 10.1145/3292500.3330758. [Online]. Available: https://doi.org/10.1145/3292500.3330758.

[7] A. Tienza-Valverde, V. H. andez-Beltr an, M. C. Espada, A. B.-S. anchez, F. J. Santos, and J. M. Gamonales, "Analysis of individual performance indicators of football goalkeeper," 2023. DOI: 10.1016/j.apunsm.2023.100420. [Online]. Available: https://doi.org/10.1016/j.apunsm.2023.100420.

[8] D. Coutinho, B. Gonçalves, T. Laakso, and B. Travassos, "Clustering ball possession duration according to players' role in football small-sided games," *PLoS ONE*, vol. 17, 8 August Aug. 2022, ISSN: 19326203. DOI: 10.1371/journal.pone.0273460.

[9] M. Merlin, S. A. Cunha, F. A. Moura, R. da Silva Torres, B. Goncalves, and J. Sampaio, "Exploring the determinants of succes in different clusters of ball possession sequences in soccer," *Research in Sports Medicine*, 2020.

[10] S. J. Mofatt, R. Gupta, N. F. COllier, and B. S. Keller, "Classifying and quantifying team playing styles in the australian football league," *International Jounral of Performance Analysis in Sport*, 2024.

[11] J. Fernandex-Navarro, L. Fradua, A. Zubillaga, P. R. Ford, and A. P. Mcrobert, "Attacking and defensive styles of play in soccer: Analysis of spanish and english elite teams," *Journal of Sports Sciences*, 2016.

[12] S. Chawla, H. J. Estephan, J. Gudmundsson, and M. Horton, "Classification of passes in football matches using spatiotemporal data," *ACM Transactions on Spatial Algorithms and Systems*, 2017.

[13] L. I. Gómez-Jordana, J. Milho, Á. Ric, R. Silva, and P. Passos, *Barça sports analytics summit landscapes of passing opportunities in football-where they are and for how long are available?* 2019.

[14] C. Wedding, C. T. Woods, W. H. Sinclair, M. A. Gomez, and A. S. Leicht, "Exploring the effect of various match factors on team playing styles in the national rugby league," *International Journal of Sports Science and Coaching*, 2021.

[15] R. Chellappa, A. Veeraraghavan, and N. Ramanathan, "Gaussian mixture models," *Encyclopedia of Biometrics*, pp. 659–663, 2009. DOI: 10.1007/978-0-387-73003-5_196. [Online]. Available: https://link.springer.com/referenceworkentry/10.1007/978-0-387-73003-5_196.

[16] A. Dempster, N. Laird, and D. Rubin, "Expectation-maximization algorithm," *Encyclopedia of Machine Learning*, pp. 387–387, 2011. DOI: 10.1007/978-0-387-30164-8_291. [Online]. Available: https://link.springer.com/referenceworkentry/10.1007/978-0-387-30164-8_291.

[17] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009, ISBN: 1441412697.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[19] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. DOI: 10.1109/MCSE.2007.55.

[20] C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. DOI: 10.1038/s41586-020-2649-2. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2.

[21] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, S. van der Walt and J. Millman, Eds., 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.