

École polytechnique de Louvain

Statistical analysis of build-up phases in football in order to determine the success factors

Author: **Adrien SIMON**

Supervisors: **Jean-Charles DELVENNE, Martin TAMINIAUX**

Reader: **François GLINEUR**

Academic year 2023–2024

Master [120] in Mathematical Engineering

Acknowledgements

Before I start, I would like to thank a number of people.

First of all, I would like to thank Jean-Charles Delvenne, my supervisor at UCLouvain, for accompanying, helping and advising me throughout the year. I would also like to thank him for giving me the freedom to choose the subject of this master thesis.

Secondly, I would like to thank Martin Taminiaux, my external supervisor, for making himself available to help me throughout the year. His knowledge of football was a crucial help to me.

Finally, I would like to thank François Glineur, reader and member of the jury for this master thesis.

Louvain-La-Neuve, June 2024
Adrien Simon

Abstract

In today's football, data analysis plays a very important role. It allows clubs and teams to analyse and optimise multiple aspects of the game in order to continually progress. One of the essential aspects of football is the build-up phase. If a team wants to progress, it must optimise its build-up phases and be as effective as possible in this aspect of the game. The subject of this master thesis is the statistical analysis of build-up phases in order to determine the success factors. We use two statistical methods, correspondence analysis and the chi-square test of independence, to determine which features have an influence on the degree of success of a build-up phase. The features can be collective characteristics but also individual characteristics specific to the players participating in the build-up phases. At the end of this thesis, we illustrate the different features found using two build-up phases of the 2022 World Cup.

Contents

List of Figures	x
List of Tables	xi
1 Introduction	1
2 Definition of a successful build-up phase	9
2.1 Definition of a build-up phase	9
2.1.1 End of a build-up phase	9
2.1.2 Start of a build-up phase	11
2.2 Definition of the success of a build-up phase	13
2.2.1 Packing	14
3 Statistical methods used	19
3.1 What will we apply our statistical methods to	19
3.1.1 Definition of contingency table	19
3.1.2 Contingency table for a build-up phase feature	20
3.2 Correspondence analysis	22
3.2.1 How correspondence analysis works	22
3.2.2 Interpretation of the results of a correspondence analysis	29
3.3 Chi-square test of independence	30
4 Analysis of the features of a build-up phase	35
4.1 Type of pass	35
4.2 Passing of players	37
4.3 Weak foot of players	38
4.4 Preferred foot of players	40
4.5 Skills of players	41
4.6 Positioning of players	42
4.7 Vision of players	44
4.8 Pace of players	45
4.9 Ball control of players	47
4.10 Physic of players	48
4.11 Agility of players	49
4.12 Work rate of players	51

4.13	Tactical system	52
4.14	League	54
4.15	Preferred foot of the left center back	55
4.16	Preferred foot of the left back	56
4.17	Summary of success features for a build-up phase	58
5	Illustration with real build-up phases	59
5.1	Example of a Very successful build-up phase	59
5.2	Example of a Failed build-up phase	61
6	Conclusion	63
A	Appendix : Datasets	67
A.1	StatsBomb	67
A.1.1	Events dataset	67
A.1.2	Lineups dataset	68
A.1.3	360 frame dataset	68
A.2	Players dataset	69
B	Appendix : Correspondence analysis plots	71
B.1	Interpreting the symmetric plot and the asymmetric row plot	71
B.2	Type of pass	73
B.3	Passing of players	73
B.4	Weak foot of players	74
B.5	Preferred foot of players	74
B.6	Skills of players	75
B.7	Positioning of players	75
B.8	Vision of players	76
B.9	Pace of players	76
B.10	Ball control of players	77
B.11	Physic of players	77
B.12	Agility of players	78
B.13	Work rate of players	78
B.14	Tactical system	79
B.15	League	79
B.16	Preferred foot of the left center back	80
B.17	Preferred foot of the left back	80
C	Appendix : Chi-square distribution table	81
	Bibliography	83

List of Figures

1.1	Ratio of scoring shots in significant zones of the pitch from [1]	3
1.2	Image of the Manchester City - Bayern Munich match on 11 April 2023 in the Champions League 2022/23 quarter-finals from [2]	5
1.3	Comparison of the styles of the 2023/2024 Premier League teams from [3] . . .	5
1.4	Images showing the progress made by StatsBomb 360 technology from [4] . .	7
2.1	Football pitch dimensions by StatsBomb from [5]	10
2.2	Images showing a pass that ends a build-up phase, a pass that is part of a build-up phase and a pass that is not part of a build-up phase	10
2.3	Division of a football pitch into thirds from [6]	11
2.4	Images showing a ball recovery initiating a build-up phase and a ball recovery not initiating a build-up phase	12
2.5	Two build-up phases for Brighton in the 2023/2024 Premier League from [3] .	13
2.6	Image illustrating the packing metric concept	15
2.7	Build-up phase of Brighton during Brighton - Newcastle in Premier League 2023/2024 from [3]	16
2.8	Build-up phase of Portugal during Portugal - Spain at World Cup 2018 from [7]	16
3.1	Example of a contingency table from [8]	20
3.2	Example of scree plot from [9]	27
3.3	Example of an asymmetric column plot from [10]	30
4.1	Asymmetric column plot for the correspondence analysis of type of pass . . .	36
4.2	Asymmetric column plot for the correspondence analysis of passing of players	37
4.3	Asymmetric column plot for the correspondence analysis of weak foot of players	39
4.4	Asymmetric column plot for the correspondence analysis of preferred foot of players	40
4.5	Asymmetric column plot for the correspondence analysis of skills of players .	41
4.6	Asymmetric column plot for the correspondence analysis of positioning of players	43
4.7	Asymmetric column plot for the correspondence analysis of vision of players .	45
4.8	Asymmetric column plot for the correspondence analysis of pace of players .	46
4.9	Asymmetric column plot for the correspondence analysis of ball control of players	47
4.10	Asymmetric column plot for the correspondence analysis of physic of players	49

4.11	Asymmetric column plot for the correspondence analysis of agility of players	50
4.12	Asymmetric column plot for the correspondence analysis of work rate of players	52
4.13	Asymmetric column plots for the correspondence analysis of tactical system .	53
4.14	Asymmetric column plot for the correspondence analysis of league	54
4.15	Asymmetric column plot for the correspondence analysis of preferred foot of the left center back	56
4.16	Asymmetric column plot for the correspondence analysis of preferred foot of the left back	57
5.1	Spain's build-up phase during the Spain-Germany match at the 2022 World Cup from [11]	60
5.2	Canada's build-up phase during the Belgium-Canada match at the 2022 World Cup from [11]	62
B.1	Symmetric plot in standard coordinates of our example illustrating how a correspondence analysis works from [12]	72
B.2	Example of an asymmetric row plot from [10]	72
B.3	Plots resulting from correspondence analysis of the type of pass	73
B.4	Plots for Passing of players	73
B.5	Plots resulting from correspondence analysis of weak foot of players	74
B.6	Plots resulting from correspondence analysis of preferred foot of players	74
B.7	Plots resulting from correspondence analysis of skills of players	75
B.8	Plots resulting from correspondence analysis of positioning of players	75
B.9	Plots resulting from correspondence analysis of vision of players	76
B.10	Plots resulting from correspondence analysis of pace of players	76
B.11	Plots resulting from correspondence analysis of ball control of players	77
B.12	Plots resulting from correspondence analysis of physic of players	77
B.13	Plots resulting from correspondence analysis of agility of players	78
B.14	Plots resulting from correspondence analysis of work rate of players	78
B.15	Plots resulting from correspondence analysis of tactical system	79
B.16	Plots resulting from correspondence analysis of league	79
B.17	Plots from correspondence analysis of preferred foot of the left center back	80
B.18	Plots resulting from correspondence analysis of preferred foot of the left back	80
C.1	Chi-square distribution table from [13]	81

List of Tables

2.1	Different levels of success of a build-up phase	17
3.1	Example of a contingency table for the type of pass feature	21
3.2	Example of a contingency table for the physic of players feature	21
3.3	Example of a contingency table to illustrate how a correspondence analysis works from [12]	22
3.4	Table of rows standard coordinates for our example illustrating how a correspondence analysis works	28
3.5	Table of columns standard coordinates for our example illustrating how a correspondence analysis works	29
3.6	Example of a contingency table to illustrate how a chi-square test of independence works	30
3.7	Example of a contingency table with rows and columns totals to illustrate how a chi-square test of independence works	31
3.8	Table showing the expected counts for our example contingency table illustrating how a chi-square test of independence works	31
3.9	Preparation to calculate our test statistic of our example illustrating how a chi-square test of independence works	32
4.1	Table of success features for a build-up phase	58

CHAPTER 1

Introduction

Today, data analysis is used in almost every sport. It continues to play an increasingly important role, and teams are investing more and more resources in it. But that was not always the case! Let's go back to the beginning of the history of data analysis in sport.

Data analysis in sport originated in the United States. In 1964, Earnshaw Cook, an American engineer, published *Percentage baseball*. This book is known as one of the first to use statistics in baseball, and more particularly in sport in general. After that, other articles were written, but there is one person in particular who is credited with popularising the use of data analysis in sport. That person is Billy Beane. In 2003, when he was General Manager of the Oakland Athletics (an American baseball team), he used sabermetrics¹ to evaluate players and make the right decisions, despite a very limited budget. [14]

Little by little, the use of data analysis has also spread to other professional sports leagues in the United States: National Football League (NFL), National Hockey League (NHL), National Basketball League (NBA), etc. The NBA is a perfect example. NBA teams now have access to dozens and dozens of advanced statistics that facilitate quantitative analysis of player and team performance. One man who has greatly increased the use of data analysis in NBA is Daryl Morey. In 2007, he became general manager of the Houston Rockets (an NBA franchise). A statistician rather than a basketball player, Morey has made extensive use of advanced statistics to define his team's style of play. Despite the criticism and scepticism of many fans and journalists, it has worked, as the Houston Rockets have become one of the best NBA franchises under Daryl Morey. [15]

Data has taken a little longer to make its mark in football. In 2019, the deputy head of recruitment at Stade de Reims (a professional football club in France) said: "Football, for its part, is only at the beginning of advanced stats, with a level of use of data that would be similar

¹A concept of statistical analysis in baseball proposed by Bill James in 1981.

to what the NBA experienced in the mid-2000s". This slight delay can be explained, in part, by the fact that football is a more fluid and less sequenced sport than most professional sports in the USA, whose phases of play are less diverse and therefore easier to analyse. One of the first clubs to adopt data analysis was Liverpool, which for some years now has had its own research department with a dedicated director, Ian Graham. He has helped the club return to the top echelons of European football. Today, the biggest clubs are developing their own algorithms in-house and adapting these algorithms to fit in perfectly with their project and their playing philosophy. Other clubs are putting data-recruitment (the use of data analysis and advanced statistics to identify, evaluate and recruit potential players based on their past performance and their suitability for a team's tactical and strategic needs) at the heart of their sporting policy. Examples include Brentford, Midtjylland and Hoffenheim. [16]

Football is a simple sport in which two teams of eleven players play against each other and try to score more goals than the opposing team. However, to score a goal, you first have to create chances and have the opportunity to take a shot at the opponent's goal.

Definition 1.0.1 (Shot). A shot is when a player tries to score by kicking the ball towards the opponent's goal.

To have the opportunity to take a shot at goal, you need to be able to put yourself in a favourable situation on the pitch. But what is a favourable situation? It is a situation in which the position on the pitch of the player about to take the shot and the positioning of the opponents mean that the shooter has a fairly high probability of scoring a goal.

Definition 1.0.2 (To score a goal). Action where a player manages to get the ball into the opposing team's goal, resulting in a point for his team. This can be achieved by striking the ball with the foot, head or any other legal part of the body, and sending it over the opposing team's goal line.

Obviously, a player has the right to shoot at goal regardless of his position or that of his opponents. But the closer the shooter is to the goal, and the fewer opponents there are to counter the shot, the greater the likelihood that the shot will be converted into a goal. Figure 1.1, from [1], illustrates this perfectly. It shows a football pitch with typical dimensions of 120 metres long and 80 metres wide. The two goals are located at coordinates (0,40) and (120,40). The figure shows that the probability of scoring from the two areas closest to the opponent's goal is 0.7, while for the areas furthest from the opponent's goal, the probability is close to 0.

Definition 1.0.3 (Scoring shot). Shot that turns into a goal.

A player's position on the pitch is not the only factor influencing whether or not a shot is successful. Another important factor is the positioning of the opposing players. The fewer defenders there are between the shooter and the opposing goalkeeper, the more time and space the shooter has to get his shot away. There is also less chance of the shot being blocked or deflected by an opposing defender. It is thus in a football team's best interest to be as effective as possible during possession phases, aiming to get as close as possible to the opposing goal

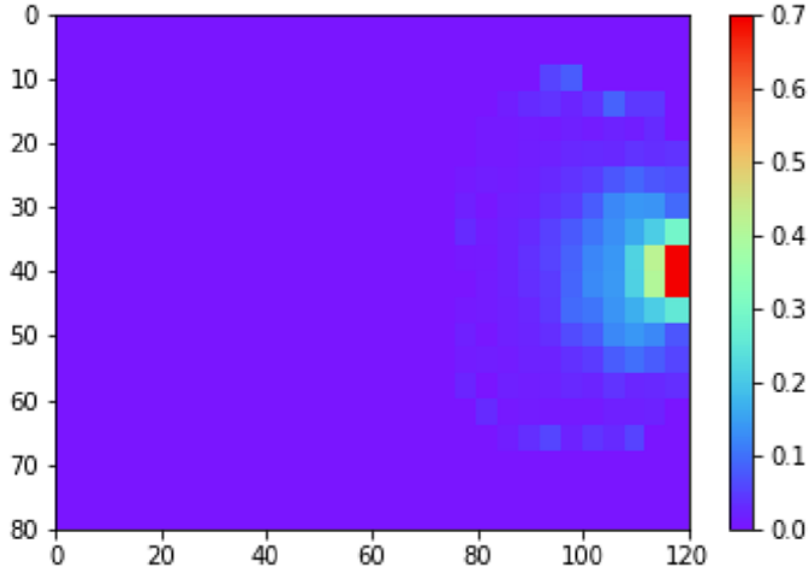


Figure 1.1: Ratio of scoring shots in significant zones of the pitch from [1]

for a shooting position and to bypass as many opponents as possible to increase the likelihood of a successful shot.

Definition 1.0.4 (Possession phase). Possessions are the combination of one or more sequences in a row belonging to the same team. That possession continues until the opposition wins the ball back by making a controlled action on the ball themselves — thus starting their own sequence of play. [17]

Definition 1.0.5 (Sequence). A sequence starts with a player making a controlled action on the ball — i.e. making a pass is controlled but making a tackle is not controlled. Opta² defines a sequence as a “passage of play which belongs to one team and are ended by defensive actions, stoppages in play or a shot”. For example, if a team makes 15 passes and works it forward before the opposition defender clears it out for a throw-in, that would be one sequence. When the player throws it back in to a team-mate, that starts a second sequence but importantly, the team still has possession of the ball. [17]

Definition 1.0.6 (Action). An action refers to a sequence of play involving passes, dribbles, shots, or other interactions among players aimed at scoring a goal. It is the essence of the game where teams try to advance towards the opponent’s goal to score.

Definition 1.0.7 (Pass). A pass is when a player kicks or directs the ball to a teammate to maintain possession, advance play, or create scoring opportunities.

Definition 1.0.8 (Dribble). A dribble is when a player manoeuvres with the ball to avoid opposing defenders and advance towards goal.

Definition 1.0.9 (Tackle). A tackle is a defensive attempt to regain possession of the ball by

²Opta specialises in the collection and analysis of sports data.

intercepting it or throwing the opponent off balance. It often involves a sliding or stretching movement of the leg to disrupt the opponent's play.

Definition 1.0.10 (Defensive action). A defensive action consists of preventing the opposing team from scoring by protecting your own goal, whether by marking, tackling, intercepting or blocking shots.

Definition 1.0.11 (Throw-in). A throw-in occurs when the ball goes out of play across the sidelines. To restart play, one team throws the ball from the point of exit, using their hands to throw overhead, while keeping one foot on the ground.

As mentioned above, the aim of a football team in possession of the ball is to get as close as possible to the opposing goal while eliminating as many opponents as possible. To achieve this, you need to be in the right conditions from the start of the possession phase. For the purpose of this master thesis, we will be looking at the start of a possession phase: the build-up phase.

Definition 1.0.12 (Build-up phase). A build-up phase refers to the process by which a team regains possession of the ball from their own half of the field and seeks to initiate an attack from the back. This often involves defenders or the goalkeeper distributing the ball to midfielders or forwards to start an offensive move. During this phase, emphasis is placed on accurate passing, vision of the game, and the ability to create space to advance towards the opposing goal. An effective build-up phase can enable a team to take control of the game and create scoring opportunities. [18]

In this master thesis, we will therefore focus solely on the build-up phase of a team in its own half of the pitch. We are therefore not interested in the course of a build-up phase after crossing the halfway line.

The reason I decided to work on this subject is that build-up phases are an essential aspect of football. Indeed, a successful build-up phase will allow a team to attack the opponent's half of the pitch with few opposing defenders and plenty of free space. Nevertheless, it is not easy to eliminate opponents during build-up phases. This aspect of the game is something that teams need to work hard on in training. Some teams are better known than others for having particularly effective build-up phases. For example, Pep Guardiola's Manchester City are formidable in build-up phases [2], thanks in part to the use of their goalkeeper Ederson. The use of goalkeepers in build-up phases has evolved considerably in recent years [19]. In the past, goalkeepers were used purely for their stopping ability, whereas nowadays they really play a part in the game. Today, technical skills and passing abilities are increasingly sought after in goalkeepers. Figure 1.2 is a good illustration of Pep Guardiola's use of goalkeeper Ederson. It can be seen that Ederson (the player in black in the image) takes part in his team's build-up phase.

In a slightly different style and with more limited financial resources than Manchester City, Brighton & Hove Albion Football Club are also known for their build-up phases. Like Manchester City, this team also relies heavily on its goalkeeper. However, as can be seen



Figure 1.2: Image of the Manchester City - Bayern Munich match on 11 April 2023 in the Champions League 2022/23 quarter-finals from [2]

in Figure 1.3, Brighton's style of play is quicker and more direct than Manchester City's [3]. Manchester City is the team whose build-up phase is the slowest and most complex in the league. This can be seen in the Direct Speed and Passes Per Sequences metrics.

Definition 1.0.13 (Direct Speed). How quickly teams progress up the pitch in open play sequences, measured in metres per second.

Definition 1.0.14 (Passes Per Sequences). Average number of passes made in a sequence.

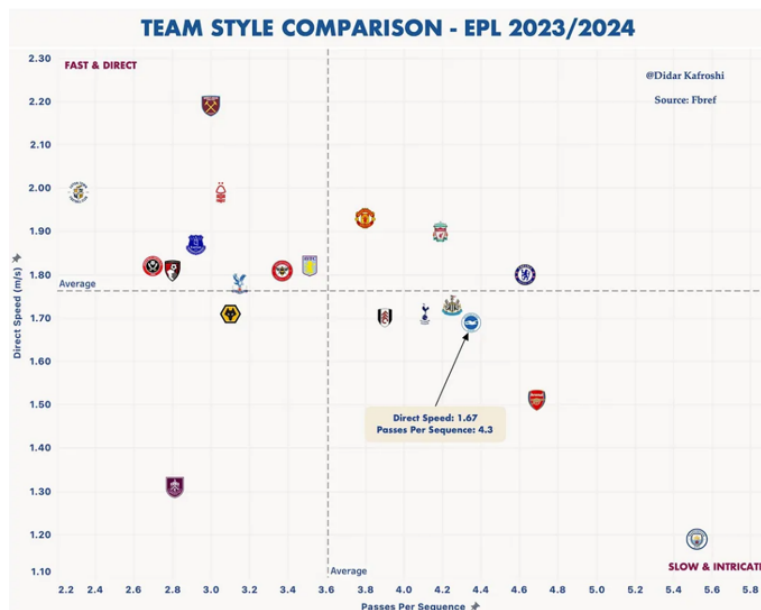


Figure 1.3: Comparison of the styles of the 2023/2024 Premier League teams from [3]

In both cases, whether it is Manchester City or Brighton, the effectiveness of their build-up phase has been evident as both teams had a very successful 2022/2023 season: Manchester City won the English league, the FA Cup, and the European Cup, while Brighton finished in a very respectable 6th place in the English league.

The objective of this thesis will be to determine the success factors of a build-up phase in a team's own half of the field. Before identifying the success factors, it will be necessary to first define what constitutes a successful build-up phase, which is not as straightforward as it may seem. It will also be essential to determine how a build-up phase begins, as well as the various possible outcomes for a build-up phase. Regarding the success factors, we will analyze all types of characteristics of build-up phases. We will focus on the profiles of players participating in build-up phases, the actions to prioritize (type of pass, dribble, etc.), the passing patterns to favor, and even the most effective tactical systems.

To determine the success factors of a build-up phase, we will use two methods of statistical analysis: correspondence analysis and chi-square test of independence. To conduct statistical analyses, we need data to work with. Therefore, the first step will be to collect this data. For this purpose, we have utilized datasets provided free of charge by the company StatsBomb [20]. StatsBomb is a company specialized in data analysis and statistics in the field of football. StatsBomb provides data on over 70 competitions. A large portion of the data is on the Spanish league, but the company also offers data on the other four major leagues (Premier League, Bundesliga, Serie A, and Ligue 1), as well as on minor leagues. The company also possesses datasets on the Champions League. The data is available for different seasons (a significant portion of the data pertains to seasons of the 21st century). There are even data on the 2018 and 2022 World Cups, Euro 2020, the 2019 and 2023 Women's World Cups, and the 2020 Women's Euro.

The company provides five types of data sets. The *Competitions* dataset lists all available competitions, the relevant year, the country, etc. The *Matches* dataset lists all matches for a given competition. The *Events* dataset lists all events for a given match, including the type of event, the player performing the event, their team, etc. The *Lineups* dataset provides team compositions for various matches. Finally, the *360* dataset is StatsBomb's latest innovation and is currently only available for a few competitions. This dataset provides, for each event in a match, the position of all players visible on camera at the time of the event, whether they are players from the team of the player performing the action or players from the opposing team. This new feature represents a significant advancement, allowing for a better understanding and analysis of the game. By adding context to each event, StatsBomb 360 makes even more advanced performance analysis possible. The three images in Figure 1.4 demonstrate that StatsBomb 360 represents a significant advancement in understanding the context of each event occurring during a match. Indeed, prior to this new technology, we only had access to the position of the players closest to the player performing the action. Now, with the position of all players visible on camera, it is easier to have a comprehensive view of the context in

which an event takes place, facilitating the analysis of that event.

Definition 1.0.15 (Event). An event refers to any significant action that occurs during a match, such as a pass, a shot on goal, a goal, a foul, a card, a tactical substitution, the end of a half, etc.

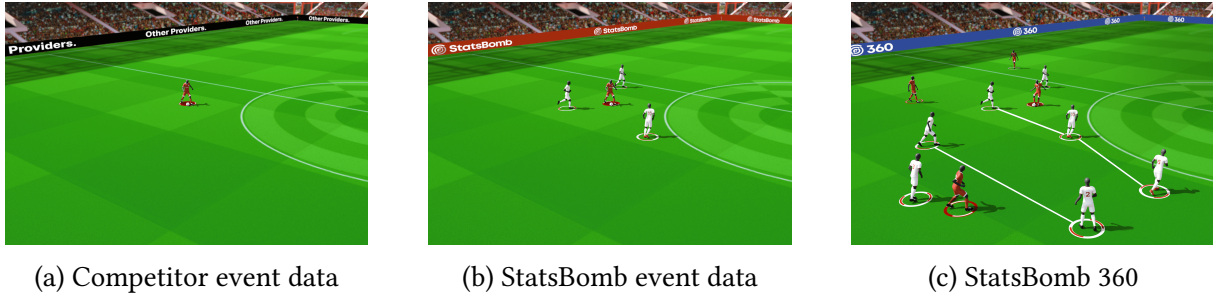


Figure 1.4: Images showing the progress made by StatsBomb 360 technology from [4]

Then, in order to analyze the characteristics of the players participating in the analyzed build-up phases, another dataset other than those provided by StatsBomb was used. This is a dataset found on Kaggle [21] providing over 100 attributes for over 19,000 football players. Obviously, we did not keep the entire dataset. We had to make a selection of the attributes and players that interested us to have a smaller dataset.

Finally, to build the code allowing the correspondence analysis of different characteristics of build-up phases, it was necessary to watch several matches. For this, we used the website Footballia [11]. It is a site containing thousands of complete football matches. The site includes over 36,000 matches with matches ranging from very old ones (1950s) to much more recent ones (2023/2024 season).

At present, there are no scientific articles on the specific subject of the success factors of a build-up phase in football. However, some articles make observations on certain characteristics of build-up phases, but these phases are not the sole subject of these articles. Instead, these articles focus on a specific aspect of the game and analyse this aspect in different phases of the game, including the build-up phase. For example, article [22] concluded that defenders have a very important role in build-up phases and that coaches should be aware of this. Article [23] looks at the average speed of players as a function of the phase of play, including the build-up phase. This article highlights the fact that players playing in attacking positions tend to walk during build-up phases. The article [24] looks at team formations in the German league as a function of the phase of play the team is in. The article explains which team formations are most commonly used during a build-up phase and explains the advantages of different team formations, but it does not explain which team formations are more likely to result in a successful build-up phase than others. We will look at this in Chapter 4. The specificity of this master thesis is therefore to focus solely on the build-up phase and to determine the success

and/or failure factors.

This master thesis will be divided into several chapters. Chapter 2 will provide an opportunity to define precisely what a build-up phase is. In this chapter, we will also define the different degrees of success of a build-up phase. In Chapter 3, we will present and define the two statistical methods used in this thesis to obtain the desired results. In Chapter 4, the different results obtained will be presented. Finally, in Chapter 5, we will look at two build-up phases resulting from the 2022 World Cup in order to observe whether the success factors obtained in Chapter 4 are well illustrated in these two phases.

CHAPTER 2

Definition of a successful build-up phase

The aim of this chapter is to define precisely the type of build-up phase that interests us in this thesis and also to define what constitutes a successful build-up phase. We will therefore begin by defining the build-up phase as it interests us in the context of this thesis. Next, we will define the success or failure of a build-up phase.

2.1 Definition of a build-up phase

A build-up phase, as used in this master thesis, can be defined by two characteristics. The first characteristic is the way in which the phase ends. The second characteristic is the way in which the phase begins.

2.1.1 End of a build-up phase

As mentioned in the introduction, we are interested here in the build-up phase of a team in its own half of the pitch. As a reminder, plot 2.1 illustrates how StatsBomb defines the coordinates of a football pitch. The top left corner is point $(x,y) = (0,0)$ while the bottom right corner is point $(x,y) = (120,80)$.

Once the ball has crossed the halfway line, the analysis is complete and we are not interested in the rest of the action. There are several ways for a team to pass the ball from its own half of the pitch to the opposing half. The most commonly used is the pass. When a pass starts in a team's own half and ends up in the other half, the build-up phase is considered to be complete. Figure 2.2a illustrates this with a pass whose start point is point $(x,y) = (40,45)$ and whose end point is point $(x,y) = (90,20)$. We consider that the team executing the build-up phase must score in the right goal. However, the pass illustrated on the Figure 2.2b does not

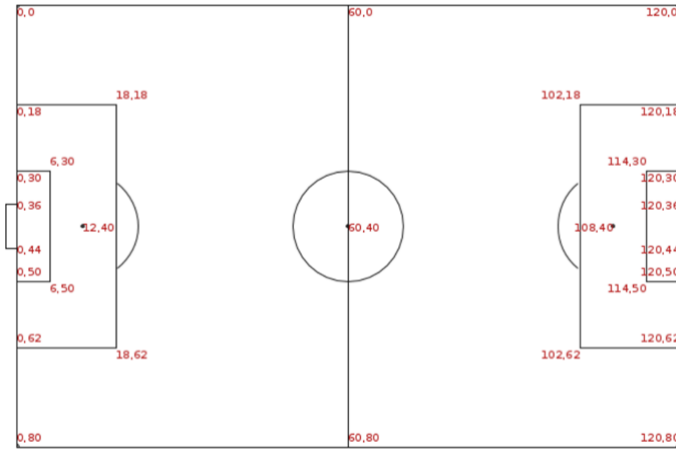
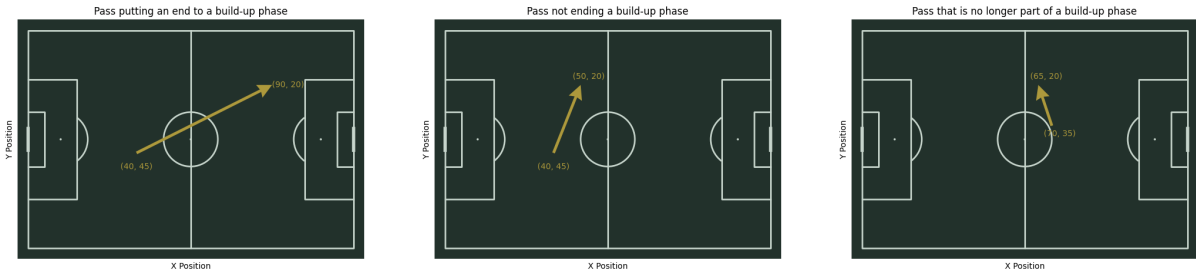


Figure 2.1: Football pitch dimensions by StatsBomb from [5]

end a build-up phase as the end point of the pass is located in the build-up phase team's half of the pitch. The pass illustrated on Figure 2.2c is no longer of interest to us, it is no longer part of a build-up phase as it takes place in the opponent's half of the pitch.



(a) Pass putting an end to a build-up phase (b) Pass not ending a build-up phase (c) Pass that is no longer part of a build-up phase

Figure 2.2: Images showing a pass that ends a build-up phase, a pass that is part of a build-up phase and a pass that is not part of a build-up phase

A second way of ending a build-up phase is when a player crosses the halfway line while carrying the ball.

Definition 2.1.1 (Carry). A carry is defined as any movement of the ball by a player which is greater than five metres from where they received the ball. [25]

A third way of ending a build-up phase is when a player crosses the halfway line of the pitch off the dribble. Finally, a build-up phase can also end when a player loses the ball before crossing the halfway line. It is also possible for a build-up phase to end with a goal kick, but this happens very rarely as the starting point of the shot must be in the shooter's own half of the pitch and the probability of scoring a shot from your own half of the pitch is very low.

To sum up, there are four different outcomes to a build-up phase (five if we include the shot, which only happens very rarely): the pass, the carry, the dribble and the loss of the ball.

2.1.2 Start of a build-up phase

Before defining the different ways of starting a build-up phase, it is important to explain the division of a football pitch into three parts¹. This division is illustrated in Figure 2.3. Each team has a first third and a last third depending on the direction of their play, then a central third which is shared between the two teams. As you can see from the image, the team in possession of the ball (the red team) has its final third at the far left of the pitch and its first third at the far right of the pitch. But it is exactly the opposite for the white team, attacking in the other direction. The build-up phase is therefore a team's construction phase when it is in its first third. [6]

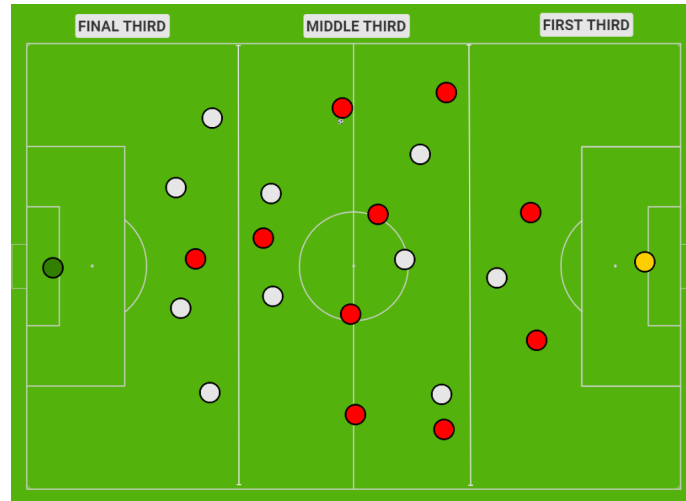


Figure 2.3: Division of a football pitch into thirds from [6]

For the purposes of this master thesis, we have decided to focus solely on build-up phases starting in the first quarter of a team's pitch, i.e. in the first 30 metres (between the coordinates $x = 0\text{m}$ and $x = 30\text{m}$), assuming here that the pitch is 120 metres long. We therefore decided not to take into account build-up phases starting between the 30-metre line and the 60-metre line to make sure we did not mix up the build-up phase with the progression phase.

Definition 2.1.2 (Progression phase). Progression phase involves the team's ability to progress up the pitch and play the ball around the middle third. [6]

Now we need to determine how these phases begin. The first way to start a build-up phase is with a goal kick. In this case, the build-up phase is started by the goalkeeper.

¹There are other ways of dividing the pitch.

Definition 2.1.3 (Goal kick). A goal kick is awarded to the defending team when the ball goes out of the field of play by crossing, either on the ground or in the air, the goal line, without a goal being scored, when the last player to touch the ball was a member of the attacking team. [26]

A second way of starting a build-up phase is to make a recovery in the zone bounded by the coordinates $x = 0\text{m}$ and $x = 30\text{m}$. The build-up phase is therefore started by the player who has recovered the ball from the opponent. He can pass, dribble, carry, etc. Figure 2.4a shows a ball recovery at $(x,y) = (25,50)$ which starts a build-up phase, whereas Figure 2.4b shows a ball recovery at $(x,y) = (45,20)$ which does not start a build-up phase because the recovery is made too high up the pitch.

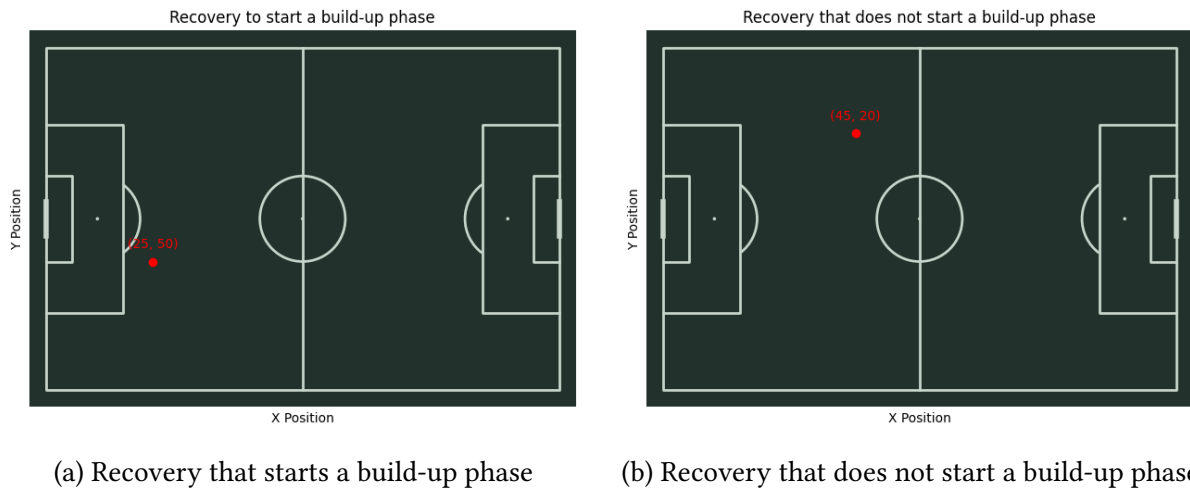


Figure 2.4: Images showing a ball recovery initiating a build-up phase and a ball recovery not initiating a build-up phase

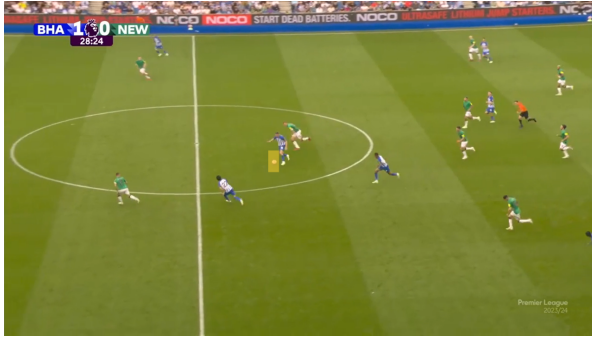
A third way to start a build-up phase is to have a throw-in in your favour within your first 30 metres of the pitch. The build-up phase is then started by the player taking the throw-in. As with ball recovery, if the throw-in occurs beyond the 30-metre line, we do not consider this to be the start of a build-up phase. A fourth way to start a build-up phase is to take a free-kick inside your own 30 metres. The build-up phase is then started by the player taking the free-kick. As with the throw-in, if the free-kick is beyond the 30-metre line, the build-up phase is not taken into account.

Definition 2.1.4 (Free kick). Penalty given to a team for a foul committed by one of its players. This penalty consists of giving the ball back to the team that committed the foul. [27]

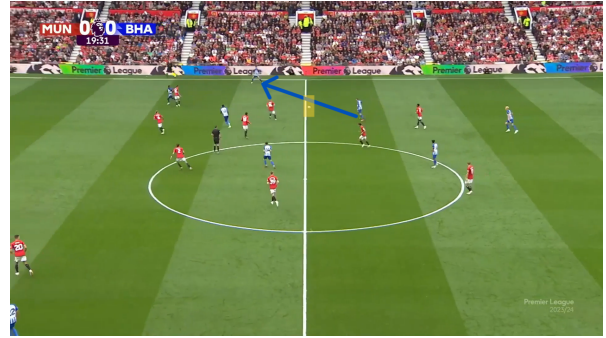
So there are four ways to start a build-up phase: with a goal kick, a ball recovery, a throw-in or a free kick.

2.2 Definition of the success of a build-up phase

Now that we have defined the build-up phases that interest us, we can define the success of these phases. At first glance, you might think that a build-up phase can either fail or succeed. However, this is not really the case. Do you think that the two build-up phases, whose endings are illustrated in the images below, have the same impact on the rest of the action? Is there not one of the two build-up phases that will allow Brighton to have more chances than the other thanks to the free space created?



(a) Manchester United - Brighton in Premier League 2023/2024



(b) Brighton - Newcastle in Premier League 2023/2024

Figure 2.5: Two build-up phases for Brighton in the 2023/2024 Premier League from [3]

In fact, if we compare the two images, we can see that the build-up phase in Figure 2.5a frees up more space than that in Figure 2.5b. In Figure 2.5a, the Brighton player who is about to cross the halfway line carrying the ball is in a situation of three attackers against two defenders. In Figure 2.5b, there are still seven Manchester United defenders between the Manchester United goalkeeper and the ball when the ball crosses the halfway line via the pass. It is therefore much more likely that Brighton will have a chance following the build-up phase in Figure 2.5a than following the build-up phase in Figure 2.5b, given the much smaller number of defenders remaining in case 2.5a.

We can therefore say that the success of a build-up phase is not binary, it is not either failed or successful. In fact, two phases can both be successful, i.e. the ball has crossed the median line without being lost, but not have the same impact on the rest of the action. There are therefore several degrees of success in a build-up phase. But how do you measure the degree of success of a build-up phase? One idea might have been to use a Possession Value model (Expected Threat, VAEP, On-Ball Value, etc.).

Definition 2.2.1 (Possession Value models). Statistical frameworks used to to objectively and quantitatively measure the value of each event on the pitch. Expected Threat (xT) is calculated by laying a 'value surface' over a football pitch to divide it into zones, where each zone has a value assigned to it based on how likely a goal is to be scored from that zone. Players can

then be credited for moving the ball from zone to zone. Unlike xT, VAEP values all actions performed by players - not just passes and carries, but shots and defensive actions too. It also considers the impact an action has on a team's chances of conceding as a result of the action - not just the impact on their chances of scoring. On-Ball Value (OBV) is an approach introduced by StatsBomb in 2021. It evaluates all actions on that take place on the pitch, containing Goals For and Goals Against components in the model to accurately measure the risk/reward of each action. [28]

The problem with these models is that they are fairly complex to implement for the StatsBomb data we have. For example, there are no characteristics giving the On-Ball Value of an event in the StatsBomb datasets. Writing a code giving us the On-Ball Value of each event present in the StatsBomb datasets would have taken too long. It would not have allowed us to concentrate on the main subject of the thesis: build-up phases. We have therefore chosen another metric for grading the success of a build-up phase: packing.

2.2.1 Packing

As explained above, the aim of the build-up phase is to find yourself at the end of it, i.e. at the halfway line of the pitch, with the most free space in front of you and therefore the most opponents eliminated. This is where the notion of packing comes into play.

Definition 2.2.2 (Packing). Packing is a statistic that can be used to quantify the utility of a pass or dribble. It is computed by counting the number of opposition players that a pass or dribble bypassed, leaving them behind the ball. [29]

The packing concept was invented and developed by Stefan Reinartz and Jens Hegeler, two former Bayer Leverkusen players [30]. Reinartz explained that, in order to find free space, a team needs a good passer but also a player capable of receiving a pass in an interesting free area of the pitch. The aim of the packing is to quantify this. Packing gives as much credit to the player who makes the pass as to the player who receives the pass. For a long time, much more credit was given to the passer than to the receiver. For example, in Figure 2.6, if player 2 passes to player 3, three players from the blue team will be eliminated by this pass. Player 2's packing value will be increased by three because he made the pass, while player 3's packing value will be increased by three because he received the pass. On the other hand, if player 2 makes the pass to player 1, no opposing player will be eliminated by the pass and the packing will therefore be zero.

Packing is a very interesting metric for determining which team has dominated a match. For example, at Euro 2016, 34 of the 51 matches were won by the team with the highest packing value. Only three matches ended in victory for the team with the lowest packing value. Another striking example is the famous Brazil-Germany match at the 2014 World Cup. Looking at the classic statistics for each team (possession, number of shots, corners, passes, etc.), it is hard to understand how Germany managed to win this match 7-1. However, if you look at each team's packing value (402 opponents taken out for Germany and 341 for Brazil),

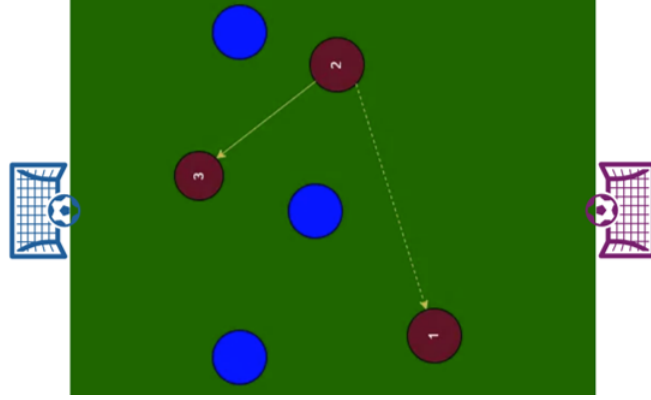


Figure 2.6: Image illustrating the packing metric concept

it is easier to understand the scoreline [31]. Packing is also a metric for evaluating players. It is used to determine which players are capable of finding passing lines that eliminate a large number of opponents and to determine which players are capable of receiving passes in advanced areas of the pitch where the density of opposing players is high. Toni Kroos has a very high packing value thanks to the large number of forward passes he makes, while a player like Eden Hazard had a very high packing value thanks to the number of pass receptions he made.

Now that we understand the concept of packing, we can easily apply it to a build-up phase to determine the different degrees of success possible:

$$\text{Packing value of a build-up phase} = \text{opponents taken out}_{\text{end}} - \text{opponents taken out}_{\text{start}}$$

To calculate the packing value for a build-up phase, simply subtract the number of opponents already eliminated at the start of the phase from the number of opponents eliminated at the end of the phase. To illustrate this, we can calculate the packing value of two build-up phases that actually took place. Let's start with the example in Figure 2.7. This is a Brighton build-up phase during the Brighton vs Newcastle match in the 2023/2024 Premier League. We can see in Figure 2.7a that the build-up phase begins with a goal kick by Billy Gilmour. At the moment of the goal kick, all the Newcastle players are between their goal and the ball, so none of them are eliminated. We have:

$$\text{opponents taken out}_{\text{start}} = 0$$

The build-up phase lasts seven seconds and ends with a carry by Evan Ferguson. By the time he crossed the halfway line, we can see in Figure 2.7b that there were only two Newcastle defenders between the goal and the ball, so eight players had been eliminated. Eliminated players are highlighted in blue (an eliminated player is not visible to the camera). So we have:

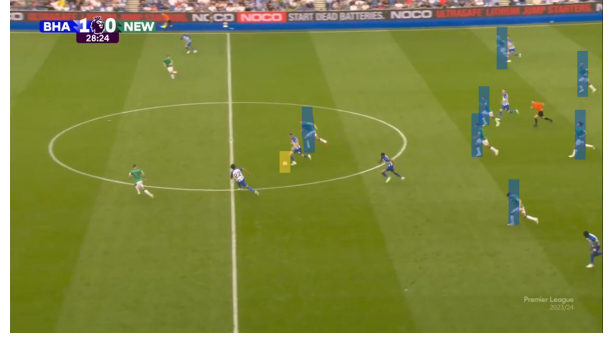
$$\text{opponents taken out}_{\text{end}} = 8$$

We can therefore calculate that the packing value for this build-up phase is:

$$\text{Packing value} = 8 - 0 = 8$$



(a) Start of the build-up phase



(b) End of the build-up phase

Figure 2.7: Build-up phase of Brighton during Brighton - Newcastle in Premier League 2023/2024 from [3]

We can calculate the packing value for another example. This is a build-up phase for Portugal during the Portugal-Spain match at the 2018 World Cup. Figure 2.8a shows that the build-up starts with Portugal's goalkeeper, Rui Patrício. The phase lasts 17 seconds and ends with a pass from William Carvalho to Cédric Soares. Figure 2.8a shows that:

$$\text{opponents taken out}_{\text{start}} = 0$$

At the end of the phase, Figure 2.8b shows the eliminated opponents highlighted in blue:

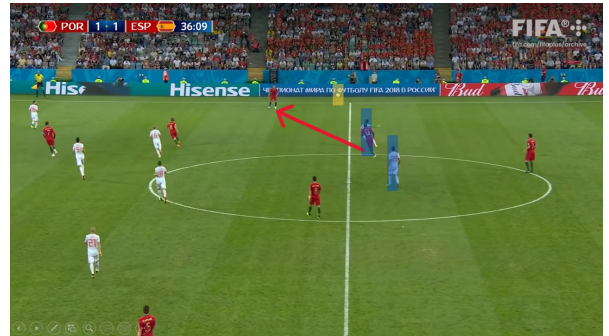
$$\text{opponents taken out}_{\text{end}} = 2$$

We can therefore calculate the packing value for this build-up phase:

$$\text{Packing value} = 2 - 0 = 2$$



(a) Start of the build-up phase



(b) End of the build-up phase

Figure 2.8: Build-up phase of Portugal during Portugal - Spain at World Cup 2018 from [7]

With the two examples above, we have two successful build-up phases which do not offer the same opportunities to the two teams for the rest of the possession phase. Indeed, with

build-up phase 2.7, the probability of Brighton creating an opportunity is higher than that of Portugal creating an opportunity with build-up phase 2.8. This is why, using the packing metric, we decided to create different categories of build-up phase success. If the build-up phase is unsuccessful, we obviously do not calculate the packing, as it is of no interest in the context of a failed phase. The table below sets out the six levels of success for a build-up phase.

Packing value of the build-up phase	Level of success of the phase
/	Failed
0 or 1	Neutral
2 or 3	Successful
4 or 5	Very successful
6 or 7	Exceptional
8 or 9	Perfect

Table 2.1: Different levels of success of a build-up phase

For the thesis, we decided not to create a category for build-up phases with a packing value of ten or eleven, as it is virtually impossible to have a phase with this packing value. Obviously, having a packing value of eleven is impossible as you would have to eliminate the opposing goalkeeper at the end of the build-up phase, which is impossible given the position of goalkeeper². It is also very complicated to have a packing value of ten, as this would require all the opposing outfield players to be in the build-up phase team's half of the pitch. This is virtually impossible given that it is a rather 'suicidal' way of defending.

Definition 2.2.3 (Outfield player). Any of the players on a football team other than the goalkeeper. [32]

To avoid taking counter-attacks into account, which are of no interest to us in this master thesis, we decided to exclude all phases in which five or more opponents had already been eliminated at the start of the phase.

Definition 2.2.4 (Counter attack). A counter attack is an attack made in response to or in defence against an attack made by another. Counter attacks take place when a team loses possession of the ball, with the opponent that has won possession of the ball forming an attack as they transition from defence to attack.

The difference between build-up phases and counter-attacks is that during a counter-attack, free space has already been created because there are quite a few opponents who are no longer in a position to defend. The objective of a counter-attack is therefore to attack the free space as quickly as possible. In a build-up phase, on the other hand, the aim is to create free space.

²The goalkeeper is generally positioned in his goal, well away from the halfway line.

CHAPTER 3

Statistical methods used

We have now defined what a build-up phase is and the different degrees of success it can have. The aim of this chapter is to explain how we are going to analyse these build-up phases in order to determine their success factors. The methods we are going to use are correspondence analysis and chi-square test of independence.

3.1 What will we apply our statistical methods to

Before explaining in detail how the two methods we are going to use work, it is essential to understand why we need these two methods. To find the factors that determine the success and/or failure of a build-up phase, we will apply our two statistical methods to contingency tables.

3.1.1 Definition of contingency table

Definition 3.1.1 (Contingency table). A contingency table is a method of representing data derived from counting, used to estimate the dependence between two categorical variables. It involves crossing two characteristics of a population (for example, an age group and a score) by counting the number of occurrences corresponding to the conjunction of 'characteristic 1' and 'characteristic 2'. The partial counts are gathered in a two-way table, with rows for the first characteristic and columns for the second characteristic: this is the 'contingency table'. The entries in the contingency table are therefore integers. [33]

Figure 3.1 is an example of contingency table. This table is useful for finding out whether there is a dependency between a person's favourite sport and their gender. As you will have understood, the different rows represent the different possible genders for a person and the columns represent the different possible sports preferences. The table can be interpreted as follows: among women, 35 prefer archery, 15 boxing and 50 cycling, while among men, 10 prefer archery, 30 boxing and 60 cycling. Of the 45 people whose favourite sport is archery,

35 are women. Of the 45 people whose favourite sport is boxing, 15 are women. Of the 110 people whose favourite sport is cycling, 50 are women.

		Sport Preference			
		Archery	Boxing	Cycling	
Gender	Female	35	15	50	100
	Male	10	30	60	100
		45	45	110	200

Figure 3.1: Example of a contingency table from [8]

3.1.2 Contingency table for a build-up phase feature

In this thesis, we will have several contingency tables, each corresponding to a feature that we wish to analyse. Obviously, we need data to complete these different tables. We are going to use the data provided by StatsBomb [20]. To analyse the build-up phases, we need matches for which StatsBomb provides 360 frame datasets. 360 frame technology is fairly recent, so there are not yet many datasets providing this kind of data. For our analysis, we decided to focus on one of the only competitions with 360 frame functionality: the 2022 World Cup. We therefore created a Python code, available at the following link [34]: https://github.com/Adri1simon18/Football_Build-up_Phase, to analyse the build-up phases of the 64 World Cup matches and their features.

The code is divided into four different functions, each focusing on a different way of starting build-up phases: on goal kick, on ball recovery, on throw-in and on free kick. For each feature we want to analyse, there is a contingency table whose rows correspond to the different degrees of success of a build-up phase and whose columns correspond to the different categories of the feature we want to analyse. Each element of the contingency table therefore corresponds to the number of times a category of a certain feature is identified in a build-up phase of a certain degree of success. The role of each of the four functions is therefore to fill in these contingency tables by iterating over each of the build-up phases of the matches we are analysing.

It is important to illustrate what a contingency table looks like in the context of this thesis. For example, if we want to analyse the impact of the type of passes made in a build-up phase, each column of the contingency table corresponds to a type of pass: the first to ground passes, the second to low passes and the third to high passes. If we have a build-up phase that began with

a ball recovery, ended in a failure and had two ground passes and three low passes, the *ball recovery* function will increase the element corresponding to the first row and first column by two and will increase the element corresponding to the first row and second column by three. When the *ball recovery* function adds this build-up phase to the type of pass contingency table, which was initially empty, we obtain Table 3.1.

	Ground	Low	High
Failed	2	3	0
Neutral	0	0	0
Successful	0	0	0
Very successful	0	0	0
Exceptional	0	0	0
Neutral	0	0	0

Table 3.1: Example of a contingency table for the type of pass feature

If we want to analyse the impact of the physical quality of the players in a build-up phase, each column of the contingency table corresponds to a category of physical quality of the players: 60-70 for the first, 70-80 for the second, 80-90 for the third and 90-100 for the last. Each row of the table corresponds to a degree of success in a build-up phase. When we analyse a feature concerning the players participating in a phase, each element of the contingency table corresponds to the number of touches by players belonging to the category corresponding to the column during the phases of the degree of success corresponding to the row in the table. If we have a build-up phase that began with a throw-in, whose outcome is Successful and whose number of touches by players with a physic of 60-70 is six and the number of touches by players with a physic of 80-90 is four, the *throw-in* function will increase the element corresponding to the third row of the table and the first column by six and will increase the element corresponding to the third row and third column by four. When the *throw-in* function adds this build-up phase to the physic of players contingency table, which was initially empty, we obtain Table 3.2.

	60-70	70-80	80-90	90-100
Failed	0	0	0	0
Neutral	0	0	0	0
Successful	6	0	4	0
Very successful	0	0	0	0
Exceptional	0	0	0	0
Neutral	0	0	0	0

Table 3.2: Example of a contingency table for the physic of players feature

The Chapter 4 will present the results of the various correspondence analyses carried out.

3.2 Correspondence analysis

Correspondence analysis is a statistical method for analysing data. It is used to analyse and prioritise the information contained in a contingency table, and is now particularly widely used to study the relationship between two qualitative variables [35]. This statistical tool is perfectly suited to our subject as we are seeking to analyse the relationships between two categorical variables. The first is the degree of success of a build-up phase and the second is one of the many characteristics of a build-up phase.

3.2.1 How correspondence analysis works

We will use Table 3.3, from [12], to explain how correspondence analysis works. Each row of the table represents a brand and each column of the table represents an attribute: tasty, aesthetic and economic. Each cell of the table contains the number of responses associating a particular brand with a specific attribute, as determined by the responses of participants to a survey.

	Tasty	Aesthetic	Economic
Butterbeer	5	7	2
Squishee	18	46	20
Slurm	19	29	39
Fizzy Lifting Drink	12	40	49
Brawndo	3	7	16

Table 3.3: Example of a contingency table to illustrate how a correspondence analysis works from [12]

We will therefore describe in detail, as explained in [36], how a correspondence analysis works on a contingency table such as Table 3.3. This can be seen as a matrix C of size $m \times n$, where m is the number of rows and n the number of columns. In this case, $m = 5$ and $n = 3$:

$$C = \begin{bmatrix} 5 & 7 & 2 \\ 18 & 46 & 20 \\ 19 & 29 & 39 \\ 12 & 40 & 49 \\ 3 & 7 & 16 \end{bmatrix}$$

In a correspondence analysis, we need to calculate the matrix of standardised residuals S defined as [36]:

$$S = W_m(P - w_m w_n)W_n$$

We need to compute W_m , W_n , P , w_m and w_n .

Definition 3.2.1 (Residual). A residual represents the difference between the value observed in a cell of the contingency table and the value expected under the assumption of independence between the rows and columns of this table.

Standardised residuals are measures used to assess the fit of observed data to an independence model. They are calculated for each cell of the contingency table and identify associations that are stronger or weaker than would be expected if the variables were independent of each other. The advantage of standardised residuals is that they are normalised, which means that they are comparable with each other and can be interpreted consistently. On the other hand, raw residuals can vary in size depending on the absolute frequency of observations, which can make their interpretation less straightforward.

The matrix P is defined as:

$$P = \frac{1}{n_C} C$$

where $n_C = \sum_{i=1}^n \sum_{j=1}^m C_{ij}$ is the sum of all cell values in matrix C . Matrix P is just the data matrix C transformed into portions i.e. each cell value is just the cell portion of the sum of the whole table. In our example, matrix P is:

$$P = \frac{1}{312} \begin{bmatrix} 5 & 7 & 2 \\ 18 & 46 & 20 \\ 19 & 29 & 39 \\ 12 & 40 & 49 \\ 3 & 7 & 16 \end{bmatrix} = \begin{bmatrix} 0.016 & 0.022 & 0.006 \\ 0.058 & 0.147 & 0.064 \\ 0.061 & 0.093 & 0.125 \\ 0.038 & 0.128 & 0.157 \\ 0.01 & 0.022 & 0.051 \end{bmatrix}$$

Now we can calculate the row masses vector w_m and the column masses vector w_n . These two vectors are defined as follows:

$$w_m = \frac{1}{n_C} C \mathbf{1} \text{ and } w_n = \frac{1}{n_C} \mathbf{1}^T C$$

where $\mathbf{1}$ is a column vector of ones. In our example, we have:

$$w_m = \frac{1}{312} \begin{bmatrix} 5 & 7 & 2 \\ 18 & 46 & 20 \\ 19 & 29 & 39 \\ 12 & 40 & 49 \\ 3 & 7 & 16 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.044 \\ 0.269 \\ 0.279 \\ 0.324 \\ 0.083 \end{bmatrix}$$

and:

$$w_n = \frac{1}{312} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 5 & 7 & 2 \\ 18 & 46 & 20 \\ 19 & 29 & 39 \\ 12 & 40 & 49 \\ 3 & 7 & 16 \end{bmatrix} = \begin{bmatrix} 0.182 & 0.413 & 0.404 \end{bmatrix}$$

So w_m is just a vector whose elements are the row sums of C divided by the sum of C and w_n is a vector whose elements are the column sums of C divided by the sum of C . Finally, the matrices W_m and W_n are defined as:

$$W_m = \text{diag}\left(\frac{1}{\sqrt{w_m}}\right) \text{ and } W_n = \text{diag}\left(\frac{1}{\sqrt{w_n}}\right)$$

where $\text{diag}(\cdot)$ is a diagonal matrix. W_m is a $m \times m$ matrix and W_n is a $n \times n$ matrix.

Definition 3.2.2 (Diagonal matrix). In linear algebra, a diagonal matrix is a matrix in which the entries outside the main diagonal are all zero.

In our example, we have:

$$W_m = \begin{bmatrix} \frac{1}{\sqrt{0.044}} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{0.269}} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{0.279}} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{0.324}} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\sqrt{0.083}} \end{bmatrix}$$

and:

$$W_n = \begin{bmatrix} \frac{1}{\sqrt{0.182}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{0.413}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{0.404}} \end{bmatrix}$$

We now have all the components we need to calculate the S matrix. The next step is to decompose the S matrix into its component parts using singular value decomposition:

$$S = U\Sigma V^*$$

where U and V are the left and right singular vectors of S and Σ is a square diagonal matrix with the singular values σ_i of S on the diagonal. V^* is the adjoint matrix of V . The singular values are the square roots of the eigenvalues of S^*S or SS^* .

Definition 3.2.3 (Adjoint matrix). The adjoint matrix of a matrix M with complex coefficients is the transpose matrix of the conjugate matrix of M .

Definition 3.2.4 (Conjugate matrix). The conjugate matrix of a matrix A is the matrix $\bar{A}$ made up of the conjugate elements of A . The conjugate of a complex number z is the complex number with the same real part as z but the opposite imaginary part.

Definition 3.2.5 (Transpose matrix). The transposed matrix of a matrix A is the matrix A^T , obtained by exchanging the rows and columns of A .

Definition 3.2.6 (Singular value decomposition). The linear algebra procedure of singular value decomposition (SVD) of a matrix is an important tool for factoring real or complex rectangular matrices.

Definition 3.2.7 (Eigenvalue). An eigenvalue of a square matrix is a scalar λ for which there exists a non-zero vector v such that the product of the matrix and this vector is equal to a multiple of this vector itself. In other words, an eigenvalue λ and a corresponding eigenvector v satisfy the equation:

$$Av = \lambda v$$

where A is the square matrix and v is the eigenvector associated with λ .

A singular value decomposition generates three outputs: the matrix U of size $m \times m$ containing the left singular vectors of S , the matrix Σ of size $m \times n$ containing the singular values of S on the diagonal and the matrix V of size $n \times n$ containing the right singular vectors of S . In the context of correspondence analysis, the left singular vectors capture the relationships between the modalities of the categorical variables represented by the rows of the contingency table (the brand in our example). The right singular vectors capture the relationships between the categories or responses represented by the columns of the table (the attribute in our example). These vectors are used to reduce the dimensionality of the data and visualise the relationships between the modalities and the categories in a reduced dimensional space. In our example, the U matrix is of size 5×5 , the Σ matrix is of size 5×3 and the V matrix is of size 3×3 . The matrix U is the following:

$$U = \begin{bmatrix} -0.439 & -0.424 & -0.084 & * & * \\ -0.652 & 0.355 & -0.626 & * & * \\ 0.16 & -0.0672 & -0.424 & * & * \\ 0.371 & 0.488 & -0.274 & * & * \\ 0.469 & -0.06 & -0.588 & * & * \end{bmatrix}$$

The matrix Σ is the following:

$$\Sigma = \begin{bmatrix} 0.265 & 0 & 0 \\ 0 & 0.114 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

The matrix V is the following:

$$V = \begin{bmatrix} -0.41 & -0.81 & -0.427 \\ -0.489 & 0.59 & -0.643 \\ 0.77 & -0.055 & -0.635 \end{bmatrix}$$

Note that, in our example, it is not necessary to calculate the last two columns of the U matrix, as the last two rows of Σ are zero. In the U matrix, each left singular vector¹ corresponds to a dimension in the space of the contingency table rows. The first column corresponds to the first dimension, the second column corresponds to the second dimension, etc. In the matrix V , each right singular vector corresponds to a dimension in the space of the contingency table columns. The first column corresponds to the first dimension, the second column corresponds to the second dimension, etc.

¹Each column of U corresponds to a left singular vector.

Definition 3.2.8 (Dimension). A dimension is a component or axis in a multidimensional space that captures some of the variability in the data, and each dimension represents a linear combination of the original variables. In the specific case of correspondence analysis, where we often work with contingency tables, the dimensions correspond to the axes in a space where the rows and columns of the table are projected.

The singular values of S , located on the diagonal of Σ , are placed in descending order. They measure the relative importance of each dimension in the data representation. The sum of all the squared singular values is called the total inertia:

$$I = \sum_{i=1}^p \sigma_i^2$$

where p is the dimension of Σ . Each dimension has its own inertia:

$$\varepsilon_i = \frac{\sigma_i^2}{\sum_{i=1}^p \sigma_i^2}$$

The inertia is a measure of the dispersion of the data in a reduced-dimensional space after performing a correspondence analysis or any other dimension reduction. It represents the proportion of the total variance of the data that is explained by each retained dimension. Thus, the higher the inertia of a dimension, the greater the proportion of variance it captures.

Definition 3.2.9 (Variance). The variance measures the dispersion of the values of a variable in relation to its mean. In the context of correspondence analysis, variance is used to measure the variability of the data along each retained dimension. The higher the variance along a dimension, the more variability or important information that dimension captures in the data.

In our example, we have:

$$I = 0.265^2 + 0.114^2 + 0^2 = 0.083221$$

The respective inertia of each dimension is therefore as follows:

- inertia of the first dimension = $\frac{0.265^2}{0.083221} = 0.845$
- inertia of the second dimension = $\frac{0.114^2}{0.083221} = 0.155$
- inertia of the third dimension = 0

In our example, we therefore obtain 84.5% of the variance expressed by our first dimension, and 15.5% in our second dimension (our third dimension explains 0% of the variance). The greater the proportion of variance explained by the first two dimensions, the more successful the correspondence analysis. The inertia of each dimension is presented in the form of a scree plot. A scree plot is a bar graph showing the percentage of variance explained by each dimension. Figure 3.2 is an example of a scree plot (note that this scree plot does not correspond to the example in the 3.3 contingency table).

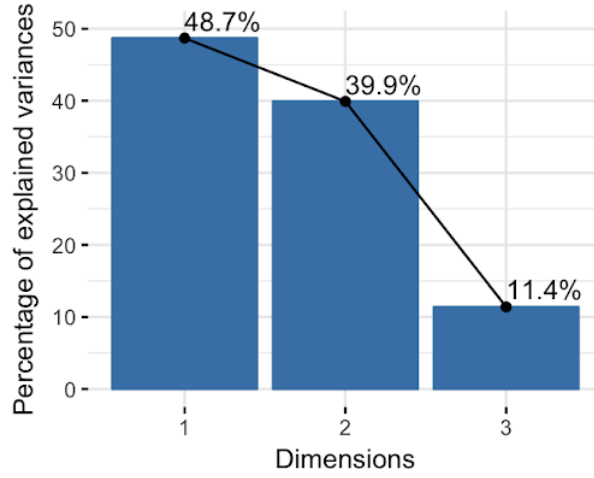


Figure 3.2: Example of scree plot from [9]

Now that we have the singular value decomposition of the S matrix, we can represent the results of the correspondence analysis using different plots. To do this, we need to transform the left and right singular vectors of S (contained in U and V) into coordinates. There are two types of coordinates: principal coordinates and standard coordinates. The principal coordinates for the rows of the C matrix are calculated as follows:

$$F_m = W_m U \Sigma$$

and the principal coordinates for the columns of the C matrix are calculated as follows:

$$F_n = W_n V \Sigma$$

The matrix F_m is therefore of size $m \times n$ while the matrix F_n is of size $n \times n$. The principal coordinates preserve the chi-square distances between the columns and rows in a correspondence analysis, meaning that the distances between the points in the space of the principal axes are proportional to the chi-square distances between the modalities of the rows and columns in the space of the original variables.

Definition 3.2.10 (Chi-square distance). The chi-square distance is a measure of dissimilarity or distance between the categorical terms of the rows and columns in a contingency table. It is often used to quantify the similarity or dissimilarity between terms in a correspondence analysis.

The standard coordinates for the rows are calculated as follows:

$$G_m = W_m U$$

and the standard coordinates for the columns are calculated as follows:

$$G_n = W_n V$$

The matrix G_m is of size $m \times m$ and the matrix G_n is of size $n \times n$. In our example, the matrix G_m is the following:

$$G_m = \begin{bmatrix} \frac{1}{\sqrt{0.044}} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{0.269}} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{0.279}} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{0.324}} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\sqrt{0.083}} \end{bmatrix} \begin{bmatrix} -0.439 & -0.424 & -0.084 & * & * \\ -0.652 & 0.355 & -0.626 & * & * \\ 0.16 & -0.0672 & -0.424 & * & * \\ 0.371 & 0.488 & -0.274 & * & * \\ 0.469 & -0.06 & -0.588 & * & * \end{bmatrix}$$

$$= \begin{bmatrix} -2.07 & -2 & -0.4 & * & * \\ -1.27 & 0.68 & -1.21 & * & * \\ 0.3 & -1.27 & -0.8 & * & * \\ 0.65 & 0.86 & -0.48 & * & * \\ 1.62 & -0.21 & -2.94 & * & * \end{bmatrix}$$

This matrix G_m therefore concerns the rows of the contingency table 3.3 in our example. The first column of this matrix G_m gives the standard coordinates of the different brands in our contingency table 3.3 (Butterbeer, Squishee, etc) in the first dimension, the second column gives the standard coordinates of the different brands in the second dimension and the third column gives the standard coordinates in the third dimension. In reality, we only need the first two columns of this G_m matrix as the first two dimensions explain 100% of the variability in the data. We can group together the coordinates in the first and second dimensions of the rows corresponding to the different brands of our contingency table 3.3:

	First dimension	Second dimension
Butterbeer	-2.07	-2
Squishee	-1.27	0.68
Slurm	0.3	-1.27
Fizzy Lifting Drink	0.65	0.86
Brawndo	1.62	-0.21

Table 3.4: Table of rows standard coordinates for our example illustrating how a correspondence analysis works

In our example, the matrix G_n is the following:

$$G_n = \begin{bmatrix} \frac{1}{\sqrt{0.182}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{0.413}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{0.404}} \end{bmatrix} \begin{bmatrix} -0.41 & -0.81 & -0.427 \\ -0.489 & 0.59 & -0.643 \\ 0.77 & -0.055 & -0.635 \end{bmatrix}$$

$$= \begin{bmatrix} -0.96 & -1.89 & -1 \\ -0.76 & 0.92 & -1 \\ 1.21 & -0.09 & -1 \end{bmatrix}$$

The matrix G_n concerns the columns of the contingency table 3.3 in our example. The first column of the matrix G_n gives the standard coordinates of the different attributes in our contingency table 3.3 (Tasty, Aesthetic and Economic) in the first dimension, the second column gives the standard coordinates of the different attributes in the second dimension and the third column gives the standard coordinates in the third dimension. As with the G_m matrix, we are only interested in the first two columns of G_n because the first two dimensions explain 100% of the variability in the data. We can group together the coordinates in the first and second dimensions of the columns corresponding to the different attributes of our contingency table 3.3:

	First dimension	Second dimension
Tasty	-0.96	-1.89
Aesthetic	-0.76	0.92
Economic	1.21	-0.09

Table 3.5: Table of columns standard coordinates for our example illustrating how a correspondence analysis works

3.2.2 Interpretation of the results of a correspondence analysis

With standard coordinates and principal coordinates, we can obtain different plots that we can interpret to obtain results. The plot of interest in this thesis is the asymmetric column plot. This plot displays column principal coordinates and row standardized coordinates in the same plot. The first dimension is on the x-axis and the second dimension is on the y-axis. As explained in [10], this plot is useful to look for relationships among the row and column categories and to help interpret the principal components. The closer a point for a column category is to a point for a row category, the higher the value of the column profile for the row category. Figure 3.3 is an example of an asymmetric column plot.

In this example, from [10], white squares correspond to the different academic subjects (biochemistry, physics, etc.), while the blue squares correspond to the different levels of funding (A is the highest level of funding and D is the lowest. E indicates no funding). Physics typically enjoys the highest level of funding because the blue square corresponding to funding level A (the highest level of funding) is the closest point to Physics of the different funding levels. Zoology typically receives the lowest because the blue square corresponding to funding level D (the lowest level of funding) is the closest point to Zoology of the different funding levels. Biochemistry tends to be unfunded, but when it is funded, it rarely falls into the lowest funding category. The plot shows that Biochemistry is closer to categories A, B and C than to category D.

It is also possible to analyse other plots such as the symmetric plot in standard coordinates and the asymmetric row plot (see Appendix B), but this is not necessary for the purposes of this master thesis, as the asymmetric column plot already gives us all the information we need.

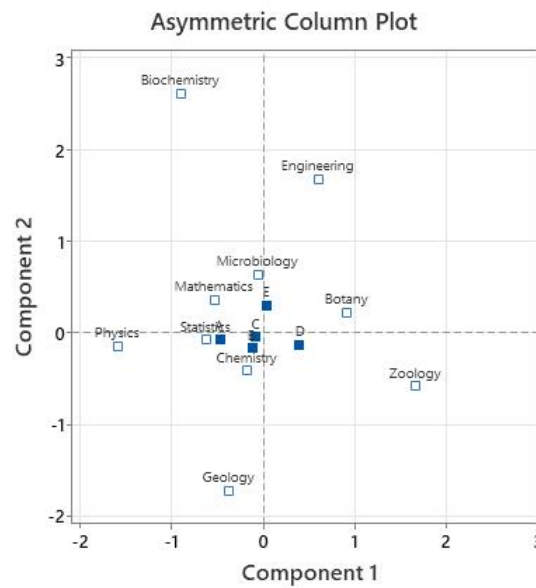


Figure 3.3: Example of an asymmetric column plot from [10]

3.3 Chi-square test of independence

We will use correspondence analysis to obtain results on the success factors of a build-up phase. However, correspondence analysis is a visualisation/exploration method. If this method suggests a dependency between the degree of success of a phase and a certain characteristic of that phase, we need to confirm this dependency using another method: the Chi-square test of independence. This test is a statistical hypothesis used to determine whether two categorical variables are likely to be related or not. This test can be applied to a contingency table. For the test to be valid, each combination of the categories of the two variables must have at least five expected values.

We will use an example, from [37], to explain how the chi-square test of independence works. The example is as follows: we have data for 600 people in a cinema. For each person, we know the type of film they watched and whether they bought a snack. Table 3.6 shows this data:

	Snacks	No snacks
Action	50	75
Comedy	125	175
Family	90	30
Horror	45	10

Table 3.6: Example of a contingency table to illustrate how a chi-square test of independence works

We now need to check that each combination of two categories of the two variables contains at least five expected values. Table 3.6 shows that each combination contains at least five values, but what about the expected values if the variables type of film and purchase of snacks are independent?

To calculate the expected values, we first need to calculate the total number of occurrences for each row and each column, as well as the total number of occurrences in the table. The following table summarises this information.

	Snacks	No snacks	Row totals
Action	50	75	125
Comedy	125	175	300
Family	90	30	120
Horror	45	10	55
Column totals	310	290	TOTAL = 600

Table 3.7: Example of a contingency table with rows and columns totals to illustrate how a chi-square test of independence works

Using Table 3.7, we can calculate the expected values for each combination of categories. For example, for the first combination Action/Snacks (first column and first row), we calculate the expected value by multiplying the total of the first row with the total of the first column and then dividing by the total number of occurrences:

$$\frac{125 \times 310}{600} = 64.58$$

We do the same with all the other combinations to obtain the following table of expected values:

	Snacks	No snacks
Action	64.58	60.42
Comedy	155	145
Family	62	58
Horror	28,42	26,58

Table 3.8: Table showing the expected counts for our example contingency table illustrating how a chi-square test of independence works

Table 3.8 shows that all the combinations have an expected value greater than five, so the test is valid for this example. We can now carry out the actual test. The idea of this test is to compare the observed values with the expected values. For each combination, we calculate the difference between the observed values and the expected values. We square this difference so

that cases where there are more expected than observed values have the same weight as cases where there are more observed than expected values. We then divide this squared difference by the expected value for the corresponding combination. For example, for the Action/Snacks combination, the calculation is as follows:

$$\frac{(50 - 64.58)^2}{64.58} = 3.29$$

We carry out the same calculation for all the combinations and obtain the following table:

	Snacks	No snacks
Action	3.29	3.52
Comedy	5.81	6.21
Family	12.65	13.52
Horror	9.68	10.35

Table 3.9: Preparation to calculate our test statistic of our example illustrating how a chi-square test of independence works

Our test statistic is the sum of the values in each cell of Table 3.9:

$$\text{Test statistic} = 65.03$$

Now that we have our test statistic, we can calculate the p-value. Generally speaking, this is calculated numerically, but it can also be calculated using a chi-square distribution table (see Appendix C for the table).

Definition 3.3.1 (P-value). In null-hypothesis significance testing, the p-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct. [38]

To determine whether our two categorical variables are dependent or not, we can use the test statistic or the p-value. First, we need to determine the significance level we want. In general, we choose $\alpha = 0.05$. This corresponds to the risk we are prepared to take of concluding that the two variables are not independent when in fact they are. We also need to determine the degrees of freedom:

$$df = (r - 1) \times (c - 1)$$

where r is the number of rows and c the number of columns. Using the significance level and the degrees of freedom, we can calculate the chi-square value using the chi-square distribution table (see Appendix C for the table). This is the value we would expect if the two variables were independent. For $\alpha = 0.05$ and $df = 3$, we obtain a chi-square value of 7.815. We compare the value of our test statistic (65.03) with the chi-square value. Since $65.03 > 7.815$, we reject the hypothesis that film type and snack purchases are independent. We conclude that there is some relationship between film type and snack purchases. We can reach the same conclusion by comparing the p-value with the significance level. If the p-value is lower than the significance

level, then we can reject the hypothesis that the two categorical variables are independent.

Now that we know how the chi-square independence test works, we can use it to obtain results that we will analyze in Chapter 4.

CHAPTER 4

Analysis of the features of a build-up phase

Once we have detailed how the statistical analysis we are going to use works, we can present the different results obtained thanks to these two methods. For a better understanding of the various plots presented in this chapter, please refer to Table 2.2.1.

4.1 Type of pass

As part of our statistical analysis, we have separated passes into three different categories: ground pass, low pass and high pass.

Definition 4.1.1 (Ground pass). Ball does not come off the ground.

Definition 4.1.2 (Low pass). Ball comes off the ground but is under shoulder level at peak height.

Definition 4.1.3 (High pass). Ball goes above shoulder level at peak height.

Plot 4.1 represents the result of the correspondence analysis for the contingency table between the degree of success of a build-up phase (rows) and the type of pass (columns) in the form of an asymmetric column plot. We can see from this plot that the first dimension explains 99.1% of the variability in the data, i.e. almost all the data. Dimension 1 best explains the row categories *Failed* and *Neutral*, and shows these two fields farthest from the origin, but with opposite signs. This means that these two categories have very different profiles. The sum of the entries in the contingency table is 5,081.

Thanks to the asymmetric column plot, we can say that favouring ground passes increases the chances of success in a build-up phase, while favouring high passes decreases the chances of success in the phase. In fact, the yellow point representing the *High* pass category is the

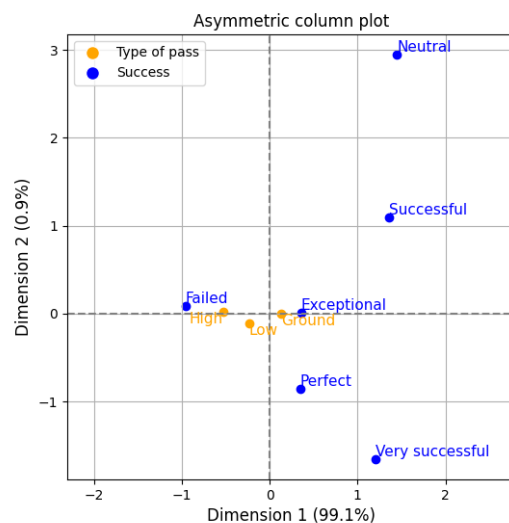


Figure 4.1: Asymmetric column plot for the correspondence analysis of type of pass

one of the three that is closest to the *Failed* category, whereas the yellow point representing the *Ground* pass category is the one that is closest to all the other build-up phase success categories.

Correspondence analysis suggests that the success of a build-up phase and the type of pass used are not independent. We will attempt to confirm it using a chi-square test of independence. By performing the test on the contingency table between the degree of success of a build-up phase (rows) and the type of pass (columns), we obtain a p-value below the significance level of 0.05 (p-value = $4.09e-65$). Given the low value of the p-value, the hypothesis that the degree of success of a build-up phase is independent of the type of pass used during that phase can be rejected with a high degree of certainty.

It is important to remember that it is not always because a player has made a high pass that a build-up phase has failed. It can be the other way round: a player makes a high pass because the build-up phase is going badly and is virtually doomed to failure. For example, if a player is under heavy pressure from the opposition and does not have time to think about the right action to take, the player will prefer to "get rid" of the ball by making a high pass towards the opposition half. The build-up phase will then fail, but at least the player will not have lost the ball close to his goal. On the other hand, there are also quite a few situations where a build-up phase fails because a team has made a high pass towards the opponent's half. This happens more often with weaker teams because these teams do not particularly have the players they need to be able to carry out build-ups using small ground passes. These teams therefore prefer to play long high passes to their forwards, who must try to win their duels against defenders who are often taller and more physical, which often leads to failure. Some teams that prefer

short ground passes sometimes use high passes to try and gain access to open space. In this case, the team has chosen to play a high pass, not to "get rid" of the ball.

4.2 Passing of players

We are now going to analyse the passing quality of the players taking part in a build-up phase. Passing quality is rated from 0 (worst quality) to 99 (best quality). We created five categories of passing quality: 50 to 60 (50-60), 61 to 70 (60-70), 71 to 80 (70-80), 81 to 90 (80-90) and 91 to 99 (90-100).

Plot 4.2 shows the result of the correspondence analysis for the contingency table between the success of a build-up phase (rows) and the passing quality of the players (columns). This figure is an asymmetric column plot. This plot shows that the first dimension explains 67.7% of the variability in the data and the second dimension 19.5%. The correspondence analysis is therefore fairly accurate, with the first two dimensions explaining more than 80% of the variability in the data. The contingency table contains 3,033 entries.

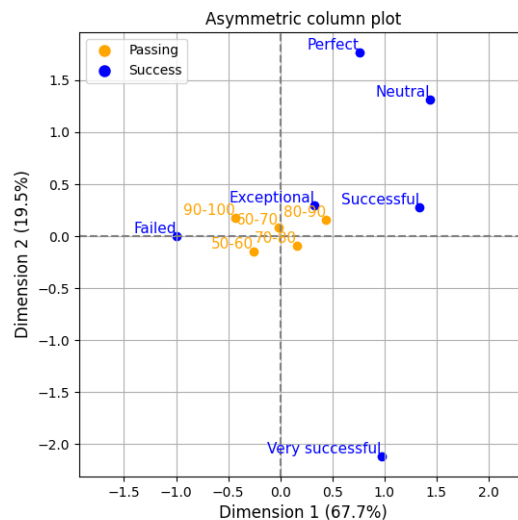


Figure 4.2: Asymmetric column plot for the correspondence analysis of passing of players

On the asymmetric column plot, we can see that as the passing quality of the players involved in the build-up phase increases, so will the chances of success in this phase. However, players with a pass quality between 91 and 99 tend to contradict this, as the 90-100 category is the closest to the *Failed* category. This result is quite astonishing. There is no reason why a player with an extremely good passing quality (between 91 and 99) should be a weak point for his team during a build-up phase. Having such a high passing quality can only be an advantage. This can perhaps be explained by the fact that the method is statistically unreliable for the

90-100 category, given that the sample for this category is very small compared with the rest of the contingency table (size 209 compared with 3,033). A small sample size can lead to an unrepresentative result, like the one we obtained with the 90-100 category.

The correspondence analysis suggests that the success of a build-up phase is related to the passing quality of players. We will attempt to confirm this using a chi-square test of independence. By performing the test on the contingency table between the degree of success of a build-up phase (rows) and the passing quality of players (columns), we obtain a p-value below the significance level of 0.05 (p-value = $2.34e-31$). Given the low value of the p-value, the hypothesis that the degree of success of a build-up phase is independent of the passing quality of players can be rejected with a high degree of certainty. So it is fair to say that having players with good passing quality will help you succeed in a build-up phase.

Pep Guardiola, the manager of Manchester City, has well understood that having players with good passing quality is essential for a successful build-up phase. Indeed, his team, known as one of the best in the world in build-up phases, has many players with very good passing quality. Kevin De Bruyne, Rodri, and Bernardo Silva have excellent passing skills. As for the defenders, Guardiola likes to have players capable of contributing to the build-up play with their passing quality (Ruben Dias, Manuel Akanji, and Nathan Aké). Guardiola also benefits from having a goalkeeper, Ederson, with very good passing ability.

4.3 Weak foot of players

The characteristic we are now going to analyse is the players' weak foot quality. The quality of the players' weak foot is divided into five categories: 1 out of 5 for the players with the worst quality right through to 5 out of 5 for the players with the best quality. However, we have decided to combine categories 1 out of 5 and 2 out of 5 into a single category called $<3/5$ and categories 4 out of 5 and 5 out of 5 into a single category called $>3/5$. We did this because categories 1 out of 5 and 5 out of 5 had some cells that did not contain at least five occurrences, which made it impossible to perform a chi-square test of independence.

Plot 4.3 represents the result of the correspondence analysis for the contingency table between the degree of success of a build-up phase (rows) and the weak foot quality of the players (columns). This figure is an asymmetric column plot. The plot shows that the first dimension explains 90% of the variability in the data and that the second dimension explains 10% of the variability. The correspondence analysis is therefore of very high quality. The sum of the entries in the contingency table is 3,993.

The asymmetric column plot shows us that the $>3/5$ category is the one to favour, as it is the closest to the *Exceptional*, *Very successful* and *Perfect* categories. This category is also the

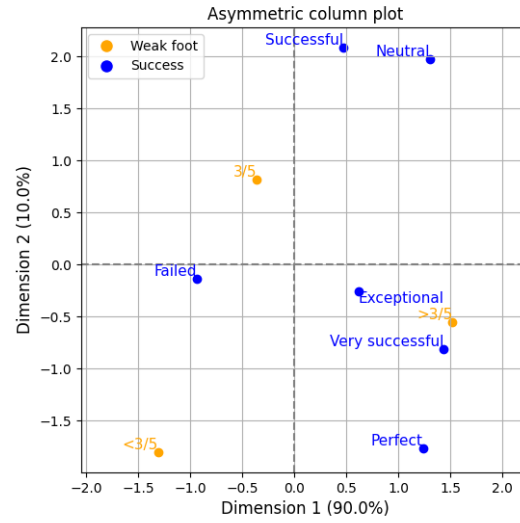


Figure 4.3: Asymmetric column plot for the correspondence analysis of weak foot of players

furthest removed from the *Failed* category. Having players with a weak foot quality of 4 or 5 out of 5 therefore increases the probability of not failing a build-up phase.

The correspondence analysis suggests that the success of a build-up phase is related to the weak foot quality of players. We will attempt to confirm this using a chi-square test of independence. By performing the test on the contingency table between the degree of success of a build-up phase (rows) and the weak foot quality of players (columns), we obtain a p-value below the significance level of 0.05 (p-value = $2.74e-13$). Given the low value of the p-value, the hypothesis that the degree of success of a build-up phase is independent of the weak foot quality of players can be rejected with a high degree of certainty. So it is fair to say that having players with good weak foot quality will help you succeed in a build-up phase.

But why does having players with very good weak foot quality help in having excellent build-up phases? Firstly, a player who is very good with both feet is more unpredictable and therefore more difficult to defend against. Secondly, having good quality with both feet allows for more passing options. The player can find multiple directions and angles for passes while maintaining high accuracy. Another advantage is that it allows for a higher tempo in the build-up phase. Indeed, a player who is comfortable with both feet can use either foot to control and pass the ball. On the contrary, a player who is not comfortable with their weak foot will lose time adjusting to use their strong foot. Finally, being able to easily play with both feet allows a player to more easily escape intense opposition pressing.

4.4 Preferred foot of players

The characteristic we are now going to analyse is the preferred foot of the players taking part in the build-up phase. Figure 4.4 represents the result of the correspondence analysis of the contingency table between the degree of success of a build-up phase (rows) and the players' preferred foot (columns). This figure is an asymmetric column plot.

The plot shows that the first dimension explains 100% of the variability in the data. This can be explained by the fact that the contingency table contains only two columns, since a player can only be left- or right-footed. The sum of the entries in the contingency table is 3,995.

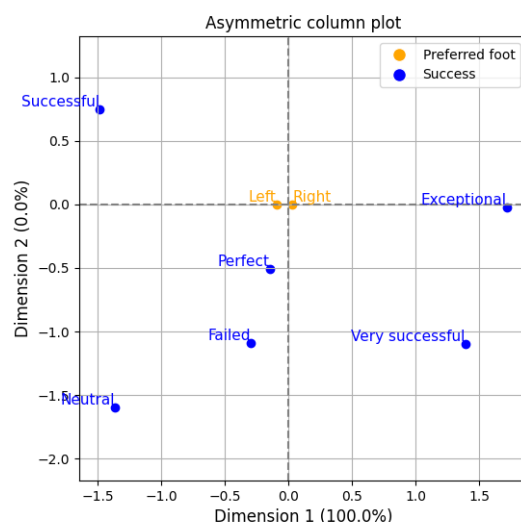


Figure 4.4: Asymmetric column plot for the correspondence analysis of preferred foot of players

As can be seen from the asymmetric column plot, the *Right* category tends to be closer to the *Very successful* and *Exceptional* categories than the *Left* category. We can also see that the *Left* category is closer to the *Failed*, *Successful* and *Perfect* category than the *Right* category. However, the distance between *Left* and *Right* is very small.

The correspondence analysis suggests that the success of a build-up phase is related to the preferred foot of players. We will attempt to confirm this using a chi-square test of independence. By performing the test on the contingency table between the degree of success of a build-up phase (rows) and the preferred foot of players (columns), we obtain a p-value below the significance level of 0.05 (p-value = 0.037). The hypothesis that the degree of success of a build-up phase is independent of the weak foot quality of players can be rejected.

We can therefore say that the success rate of a build-up phase is not independent of the preferred foot of the players involved. But we cannot say that it is more advantageous to have a right-footed or left-footed player. Indeed, having left-footed players tends to very slightly increase the probability of having a *Successful* or *Perfect* build-up phase, while having right-footed players tends to very slightly increase the probability of having a *Very Successful* or *Exceptional* build-up phase. Thus, a player's preferred foot is not a determining factor in the success of a build-up phase.

4.5 Skills of players

The characteristic we will be analysing in this section is the players' skills moves. Skills refer to a player's technical ability and are rated from 1 out of 5, the lowest, to 5 out of 5, the highest. We did not include the 5 out of 5 category in our correspondence analysis because the sample of players with 5 out of 5 skills moves is very limited. The results may therefore be influenced.

Figure 4.5 represents the asymmetric column plot of the correspondence analysis of the contingency table. The rows of the table are the different degrees of success in a build-up phase and the columns of which are the different skill move categories of the players. The plot shows that the first dimension explains 85.7% of the variability in the data, while the second dimension explains 9.8%. The correspondence analysis is therefore of very high quality. The plot shows that the distance between the points in the *Neutral* and *Very successful* categories is small. This indicates that these two categories have two very similar profiles. The *Neutral* and *Very successful* build-up phases therefore use players with virtually the same skill moves. The sum of the entries in the contingency table is 3,973.

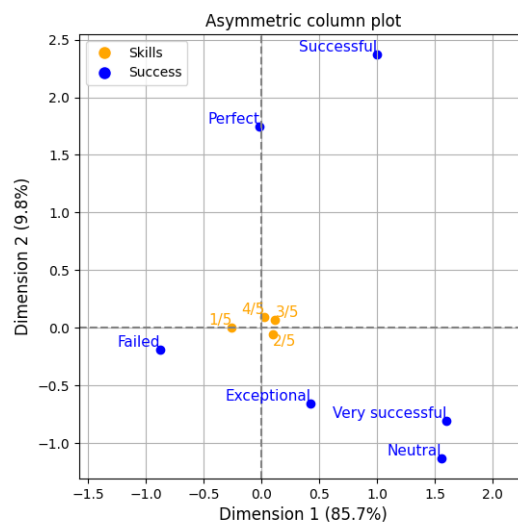


Figure 4.5: Asymmetric column plot for the correspondence analysis of skills of players

On the asymmetric column plot, we can see that the 1/5 category is closer to the *Failed* category than the other three skill move categories, which are all very close to each other. We can therefore conclude that avoiding players with a skill moves score of 1 out of 5 reduces the probability of failure in a build-up phase. However, having players with a score of 4 out of 5 will not significantly increase the chances of a successful build-up phase. However, we should be wary of this last point given that the sample of players with 4 out of 5 skills is considerably smaller than the other three skills category samples. The sum of the entries in column 4/5 is 140, which is very low compared to the total sum of the table (3,973).

The correspondence analysis suggests that the success of a build-up phase is related to the skills moves of players. We will attempt to confirm this using a chi-square test of independence. By performing the test on the contingency table between the degree of success of a build-up phase (rows) and the skills moves of players (columns), we obtain a p-value below the significance level of 0.05 (p-value = 1.44×10^{-18}). Given the low value of the p-value, the hypothesis that the degree of success of a build-up phase is independent of the skills moves of players can be rejected with a high degree of certainty. We can therefore say that having players with good skills moves favours the success of a build-up phase.

How does having players with a high skill moves rating help in the success of a build-up phase? The main advantage of these players during a build-up phase is that they are capable of escaping intense opposition pressing with a dribble. Using dribbling can also allow a player to create space, which is one of the main objectives of a build-up phase.

4.6 Positioning of players

The characteristic analysed in this section is the quality of a player's positioning. This is rated from 0 to 99, i.e. from the weakest to the strongest.

Definition 4.6.1 (Positioning). A footballer's positioning quality refer to his ability to occupy the right areas of the pitch, read the play and anticipate the movements of team-mates and opponents, while contributing effectively to the team.

We created five categories for player positioning quality: below 50 (<50), 50 to 60 (50-60), 61 to 70 (60-70), 71 to 80 (70-80), and 81 to 99 (80-100). Figure 4.6 shows the asymmetric column plot of the correspondence analysis of the contingency table, where the rows represent different levels of success in the build-up phase, and the columns represent different categories of player positioning quality.

The plot indicates that the first two dimensions account for 71.5% and 20.5% of the data variability, respectively, demonstrating the high quality of the correspondence analysis. The plot also shows that the distance between the *Perfect* and *Very successful* categories is small,

indicating that these two categories have very similar profiles. The sum of the entries in the contingency table is 3,995.

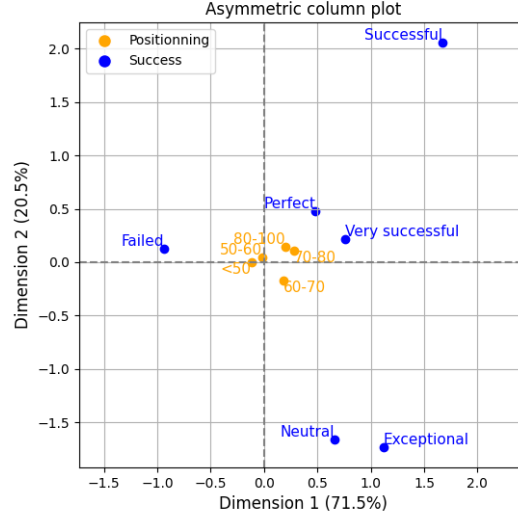


Figure 4.6: Asymmetric column plot for the correspondence analysis of positioning of players

On the asymmetric column plot, we can see that the two categories *<50* and *50-60* are closest to the *Failed* category. For a successful build-up phase, it is therefore preferable not to have any players with a position quality of less than 60. However, we can see from the plot that having a very good positioning quality does not necessarily guarantee a perfectly successful build-up phase. In fact, the *80-100* category is not much closer to the *Very successful*, *Exceptional* and *Perfect* categories than the *60-70* and *70-80* categories. However, we cannot confirm this with 100% certainty, given that the *80-100* sample is much smaller than the samples for the other four categories. The *80-100* category column contains only 101 elements out of the 3,995 in the contingency table.

The correspondence analysis suggests that the success of a build-up phase is related to the positioning quality of players. We will attempt to confirm this using a chi-square test of independence. By performing the test on the contingency table between the degree of success of a build-up phase (rows) and the categories of player positioning quality (columns), we obtain a p-value below the significance level of 0.05 (p-value = $5.01e-17$). Given the low value of the p-value, the hypothesis that the degree of success of a build-up phase is independent of the positioning quality of players can be rejected with a high degree of certainty. We can therefore say that having players with good positioning quality favours the success of a build-up phase.

Players with good positioning quality are able to provide good passing options to their teammates. These players also have the ability to understand the positioning that the coach wants to implement, which facilitates the execution of the coach's game plan. Having players

with good positioning therefore allows the implementation of desired passing patterns.

4.7 Vision of players

The characteristic we are going to analyse in this section is the vision of the game of the players taking part in the build-up phase. This is rated from 0, for the players with the worst vision of the game, to 99, for the players with the best vision of the game.

Definition 4.7.1 (Vision). Football vision can be understood as a player's ability to anticipate the actions of team-mates and opponents, as well as perceiving tactical opportunities on the pitch. It encompasses reading the game, understanding the team's tactical schemes and the ability to make quick, judicious decisions to influence the course of the match.

As part of our analysis, we created five categories for a player's vision of the game: below 50 (<50), 50 to 60 (50-60), 61 to 70 (60-70), 71 to 80 (70-80), and 81 to 99 (80-100). Due to the small sample size in the 91 to 99 category, we merged the 81 to 90 and 91 to 99 categories. Since some cells in the 91 to 99 category had fewer than five occurrences, we could not perform a chi-square test of independence.

Figure 4.7 shows the asymmetric column plot of the correspondence analysis of the contingency table. The rows represent different degrees of success in a build-up phase and the columns represent the different categories of a player's vision of the game. The plot indicates that the first dimension explains 63.1% of the variability in the data, while the second dimension explains 21.1%. This analysis is of good quality, as these two dimensions together account for more than 80% of the data variability. The sum of the entries in the contingency table is 3,995.

On the asymmetric column plot, there are two 'groups' of game vision categories: the first includes categories <50, 50-60 and 60-70 and the second categories 70-80 and 80-100. The first is closer to the *Failed* category than the second, which implies that having players with poor game vision increases the probability of failure in the build-up phase. Within the second group, the 70-80 category is closer to the *Successful* and *Very successful* categories, while the 80-100 category is closer to the *Exceptional*, *Neutral* and *Perfect* categories. This means that having players whose vision of the game is very very good (between 81 and 99) increases the probability of having an *Exceptional* or *Perfect* build-up phase but also increases the risk of having a *Neutral* build-up phase. On the other hand, having players in the 70-80 category is less risky in the sense that the probability of having a *Neutral* phase build-up is lower but the probability of having an *Exceptional* or *Perfect* phase is also lower.

The correspondence analysis suggests that the success of a build-up phase is related to the vision of players. We will attempt to confirm this using a chi-square test of independence. By performing the test on the contingency table between the degree of success of a build-up phase

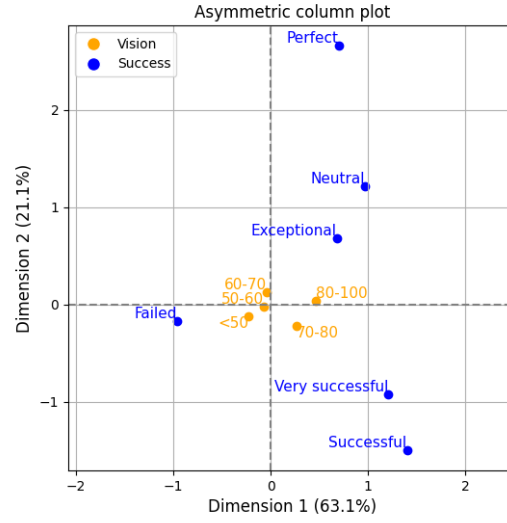


Figure 4.7: Asymmetric column plot for the correspondence analysis of vision of players

(rows) and the categories of player vision (columns), we obtain a p-value below the significance level of 0.05 (p-value = $1.41e-39$). Given the low value of the p-value, the hypothesis that the degree of success of a build-up phase is independent of the vision of players can be rejected with a high degree of certainty. We can therefore say that having players with good vision favours the success of a build-up phase.

Players with very good vision are therefore an advantage for successfully completing a build-up phase because they can anticipate the movements of their teammates and opponents. This allows them to spot opportunities that other players might miss. They are also able to see passes that other players would not, with these passes breaking several lines and eliminating a large number of opposing players. With their vision, these players know when to play long, when to play short, when to accelerate the game, and so on.

4.8 Pace of players

In this section, we will analyze the players' pace, which is rated from 0 to 99, with 0 being the slowest and 99 being the fastest. We created six categories for players' pace: below 50 (<50), 50 to 60 (50-60), 61 to 70 (60-70), 71 to 80 (70-80), 81 to 90 (80-90), and 91 to 99 (90-100). Figure 4.8 shows the asymmetric column plot of the correspondence analysis of the contingency table, where the rows represent different degrees of success in a build-up phase, and the columns represent the different categories of a player's pace.

The plot indicates that the first dimension explains 69.3% of the variability in the data, while

the second dimension explains 26.1%. This correspondence analysis is of excellent quality, as the first two dimensions together explain more than 95% of the data variability. The sum of the entries in the contingency table is 3,995.

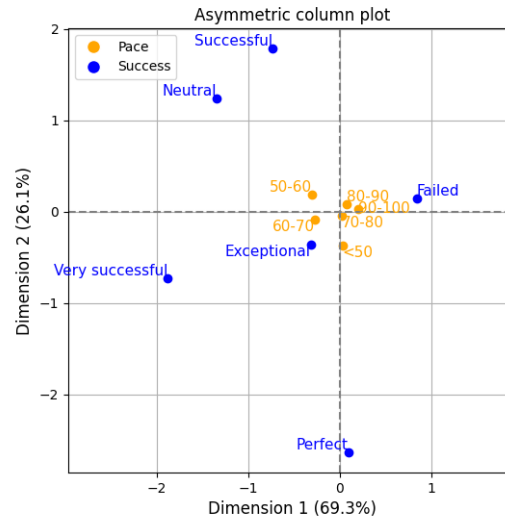


Figure 4.8: Asymmetric column plot for the correspondence analysis of pace of players

On the asymmetric column plot, it is observed that the 70-80, 80-90, and 90-100 categories are the three closest to the *Failed* category. This means that having players with a pace exceeding 70 increases the probabilities of a build-up phase failure. It is also noticeable that the <50 category is the one closest to the *Perfect* category. Therefore, it can be said that having particularly slow players increases the probabilities of having a perfectly successful build-up phase.

The correspondence analysis suggests that the success of a build-up phase is related to the pace of players. We will attempt to confirm this using a chi-square test of independence. By performing the test on the contingency table between the degree of success of a build-up phase (rows) and the categories of player pace (columns), we obtain a p-value below the significance level of 0.05 (p-value = 1.36e-31). Given the low value of the p-value, the hypothesis that the degree of success of a build-up phase is independent of the pace of players can be rejected with a high degree of certainty. We can therefore say that having players with low pace favours the success of a build-up phase.

This result is quite surprising. Indeed, one might think that having only very fast players is an advantage, but this is not necessarily the case. Obviously, very fast players have several advantages, but the build-up phase is not a phase where pace is the primary quality sought. This phase is, most of the time, a phase where one must be patient and wait for a space to open up. Some very fast players can sometimes be too reliant on their pace and attempt

quick breakthroughs when the situation calls for more patience and control. This can lead to turnovers.

4.9 Ball control of players

In this section, we will analyze the quality of players' ball control, which is rated from 0 to 99. We created five categories for ball control quality: below 50 (<50), 50 to 60 (50-60), 61 to 70 (60-70), 71 to 80 (70-80), and 81 to 99 (80-100). Figure 4.9 shows the asymmetric column plot of the correspondence analysis of the contingency table, where the rows represent different degrees of success in a build-up phase, and the columns represent the different categories of a player's ball control.

The plot indicates that the first dimension explains 79.7% of the variability in the data, while the second dimension explains 15.2%. This correspondence analysis is of excellent quality, as the first two dimensions together explain 94.9% of the data variability. The sum of the entries in the contingency table is 3,995.

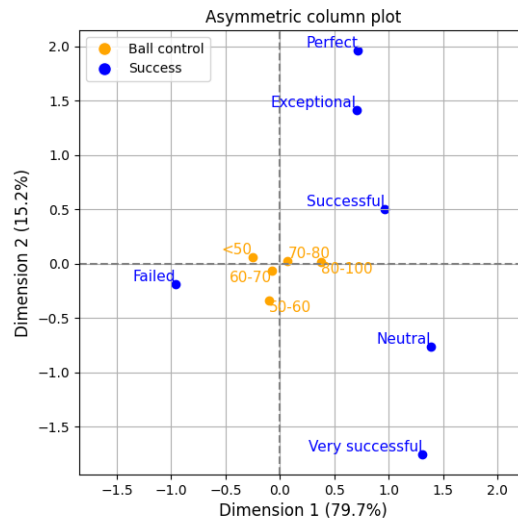


Figure 4.9: Asymmetric column plot for the correspondence analysis of ball control of players

On the asymmetric column plot, it is observed that the 80-100 category is the furthest from the *Failed* category. This means that having players with a ball control rating above 80 greatly decreases the probabilities of a build-up phase failure. It is also noticeable that the worst players to have for a build-up phase are those whose ball control quality belongs to the 50-60 category. Indeed, this category is quite close to both the *Failed* and *Neutral* categories. We can see that the <50 category is very slightly closer to the *Failed* category than the 50-60 category, but the latter is closer to the *Neutral* category than the <50 category. These two categories

should be avoided if you want a successful build-up phase.

The correspondence analysis suggests that the success of a build-up phase is related to the ball control quality of players. We will attempt to confirm this using a chi-square test of independence. By performing the test on the contingency table between the degree of success of a build-up phase (rows) and the categories of player's ball control quality (columns), we obtain a p-value below the significance level of 0.05 (p-value = $4.55e-37$). Given the low value of the p-value, the hypothesis that the degree of success of a build-up phase is independent of the ball control quality of players can be rejected with a high degree of certainty. We can therefore say that having players with good ball control quality favours the success of a build-up phase. Having players with very good control quality is therefore an advantage in the context of a build-up phase. Through their controls, these players are able to escape intense opposition pressing or create space.

4.10 Physic of players

In this section, we will analyze the physical qualities of players, rated from 0 to 99. We created four categories for physical qualities: 61 to 70 (60-70), 71 to 80 (70-80), 81 to 90 (80-90), and 91 to 99 (90-100). Figure 4.10 shows the asymmetric column plot of the correspondence analysis of the contingency table, where the rows represent different degrees of success in a build-up phase, and the columns represent the different categories of a player's physical qualities.

The plot indicates that the first dimension explains 89.1% of the variability in the data, while the second dimension explains 9.3%. This correspondence analysis is of excellent quality, as the first two dimensions together explain 98.4% of the data variability. The sum of the entries in the contingency table is 3,982.

On the asymmetric column plot, it is observed that the two categories closest to *Failed* are the 60-70 and 90-100 categories. Therefore, it can be said that to avoid failing a build-up phase, it is necessary to avoid having players with too little physicality, but it is also necessary to avoid players with too much physicality. It is also observed that the category closest to the *Exceptional* and *Perfect* success degrees is the 70-80 category, but it is closely followed by the 80-90 category. The ideal physical qualities for a player participating in a build-up phase are therefore all those rated between 71 and 80.

The correspondence analysis suggests that the success of a build-up phase is related to the physical quality of players. We will attempt to confirm this using a chi-square test of independence. By performing the test on the contingency table between the degree of success of a build-up phase (rows) and the categories of player's physical quality (columns), we obtain a p-value below the significance level of 0.05 (p-value = $2.51e-18$). Given the low value of the

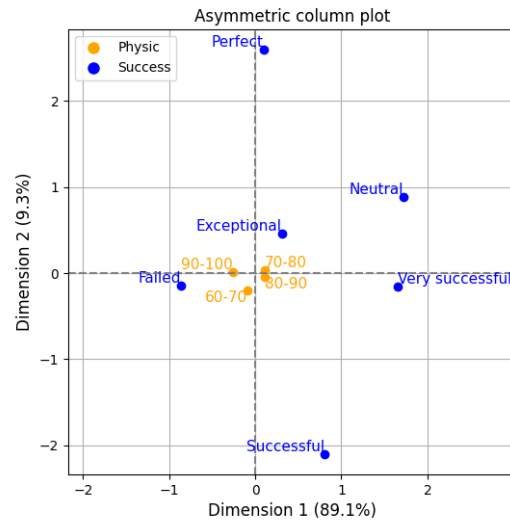


Figure 4.10: Asymmetric column plot for the correspondence analysis of physic of players

p-value, the hypothesis that the degree of success of a build-up phase is independent of the physical quality of players can be rejected with a high degree of certainty.

The physical quality of a player is based on several aspects: speed, strength, endurance, etc. It is quite easy to understand that players with low physical quality are not ideal for a build-up phase. However, for players with a physical quality rating above 90, it is less straightforward to comprehend. In fact, players who are too physical (with a rating above 90) may lack mobility and agility due to their significant muscle mass. To illustrate the point, we can observe that the most renowned players in the world for a build-up phase have a physical quality rating ranging between 70 and 90: Kevin De Bruyne has a rating of 78, Frenkie de Jong also, Virgil van Dijk has a rating of 84, Trent Alexander-Arnold has a rating of 72, etc.

4.11 Agility of players

In this section, we will analyze the agility of players, rated from 0 to 99. We created five categories for agility: below 50 (<50), 50 to 60 (50-60), 61 to 70 (60-70), 71 to 80 (70-80), and 81 to 99 (80-100). Since some cells in the 81 to 90 and 91 to 99 categories had fewer than five occurrences, we merged these two categories to perform a chi-square test of independence.

Figure 4.11 shows the asymmetric column plot of the correspondence analysis of the contingency table. The rows represent different degrees of success in a build-up phase, and the columns represent different categories of a player's agility. The plot indicates that the first dimension explains 72.9% of the variability in the data, while the second dimension explains

20.4%. This correspondence analysis is of excellent quality, as the first two dimensions together explain more than 90% of the data variability.

The plot also shows that the *Failed* and *Perfect* categories, as well as the *Exceptional* and *Very successful* categories, are very close to each other. This means that these pairs of categories have very similar profiles in terms of players' agility. The sum of the entries in the contingency table is 3,995.

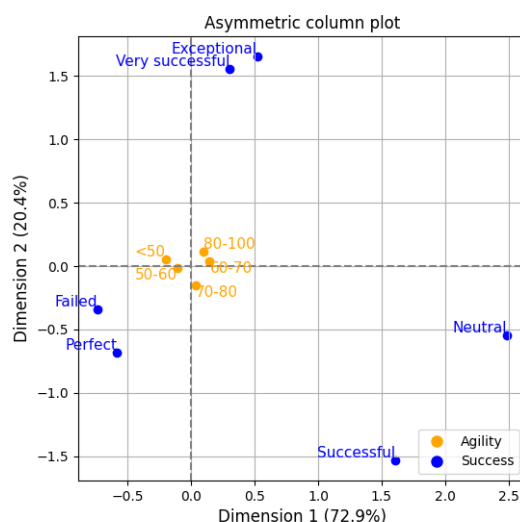


Figure 4.11: Asymmetric column plot for the correspondence analysis of agility of players

On the asymmetric column plot, it is observed that the success degree categories of a build-up phase are separated into three 'groups': *Failed* with *Perfect*, *Exceptional* with *Very successful*, and *Successful* with *Neutral*. The *80-100* category is closest to the *Very successful* and *Exceptional* groups. Therefore, this category is the most interesting for players in a build-up phase. Indeed, the *Successful* and *Neutral* group is less interesting because these two categories have a lower success degree, while the *Failed* and *Perfect* group is also less interesting because there is a risk of the build-up phase failing. Therefore, it can be said that the ideal agility for a player in a build-up phase is between 81 and 99. However, we must be wary of the *80-100* category, as the sample in this category is much smaller than those in the other categories. The *80-100* category column contains only 142 elements out of the 3,995 in the contingency table.

The correspondence analysis therefore suggests that the success of a build-up phase is related to the agility of players. We will attempt to confirm this using a chi-square test of independence. By performing the test on the contingency table between the degree of success of a build-up phase (rows) and the categories of player's agility (columns), we obtain a p-value below the significance level of 0.05 (p-value = $1.82e-13$). Given the low value of the p-value, the hypothesis that the degree of success of a build-up phase is independent of the agility of

players can be rejected with a high degree of certainty.

Similarly to players with good skill moves, agile players are valuable in a build-up phase because they are often capable of executing dribbles to bypass opponents. They are thus crucial for creating space and avoiding opposition pressing. Moreover, due to their agility, they can easily change direction and therefore outmaneuver their defender, enabling them to provide a good passing option to their teammates.

4.12 Work rate of players

The characteristic we are going to analyze in this section is the work rate of players.

Definition 4.12.1 (Work rate). The work rate of a football player refers to their level of activity and effort on the field, particularly regarding their willingness to pursue, defend, score, and contribute to the team. The work rate is expressed in two parts: first, the player's offensive involvement, and then their defensive involvement. For example, a player with a High/Low work rate is highly involved in their team's offensive actions but minimally involved in defensive actions.

For our analysis, we have identified seven types of work rates: High/High (*H/H*), High/Medium (*H/M*), High/Low (*H/L*), Medium/High (*M/H*), Medium/Medium (*M/M*), Low/High (*L/H*), and Low/Medium (*L/M*). Figure 4.12 is the asymmetric column plot of the correspondence analysis of the contingency table. The rows are the different degrees of success in a build-up phase and the columns are the different work rate. The plot shows that the first dimension explains 62.8% of the variability in the data, while the second explains 22.1%. The correspondence analysis is of very good quality since the first two dimensions explain more than 80% of the data variability. The sum of the entries in the contingency table is 3,986.

On the asymmetric column plot, we observe that the *L/M* category is the closest to *Failed*. This means that having players with a Low/Medium work rate increases the chances of failure in a build-up phase. This is quite logical since, by definition, these players have low offensive involvement. Note that you need to be careful with the *H/L* and *L/M* categories, as the samples in these two categories are fairly small (size 33 and 67 respectively). To succeed in a build-up phase, it is therefore preferable to have players with a high offensive contribution.

The correspondence analysis therefore suggests that the success of a build-up phase is related to the work rate of the players. We will attempt to confirm this dependence using a chi-square test of independence. By performing the test on the contingency table between the degree of success of a build-up phase (rows) and the different work rate (columns), we obtain a p-value below the significance level of 0.05 (p-value = 8.99e-27). Given the low value of the p-value, the hypothesis that the degree of success of a build-up phase is independent of the work rate

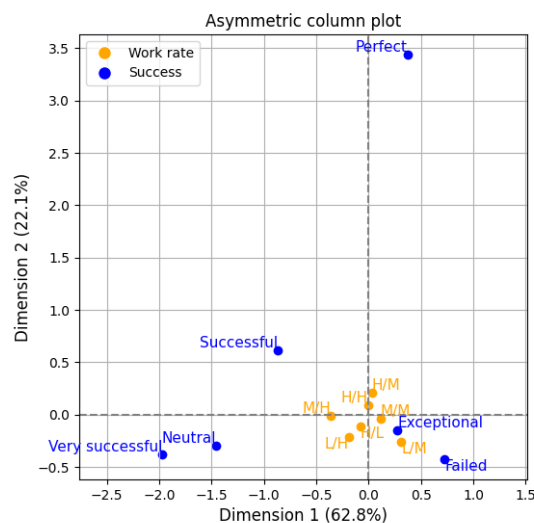


Figure 4.12: Asymmetric column plot for the correspondence analysis of work rate of players

of the player can be rejected with a high degree of certainty. We can therefore say that the success of a build-up phase is related to the work rate of the player.

4.13 Tactical system

The characteristic we will analyze in this section is the team's tactical system. We have identified seven different formations for this analysis: 4-2-3-1, 4-3-3, 4-4-2, 3-4-2-1, 3-5-2, 4-1-4-1, and 3-4-3. Each number indicates the number of players in each line, starting with defense, then midfield, and finally attack. For example, in the 4-3-3 formation, there are four defenders, three midfielders, and three attackers.

Figure 4.13a is the asymmetric column plot of the correspondence analysis of the contingency table. The rows are the different degrees of success in a build-up phase and the columns are the different formations. The plot shows that the first dimension explains 68.9% of the variability in the data, while the second explains 23.7%. The correspondence analysis is of excellent quality since the first two dimensions explain more than 90% of the data variability. Figure 4.13b is a zoom of the center of the asymmetric column plot to facilitate a clearer interpretation of this plot. The sum of the entries in the contingency table is 2,128.

Thanks to the zoom of the asymmetric column plot, despite the distances between the different formations being quite small, it can be observed that the formation closest to the *Failed* category is the 4-4-2. Therefore, the 4-4-2 formation should be avoided to decrease the chances of failure in the build-up phase. The asymmetric column plot also shows that tactical systems

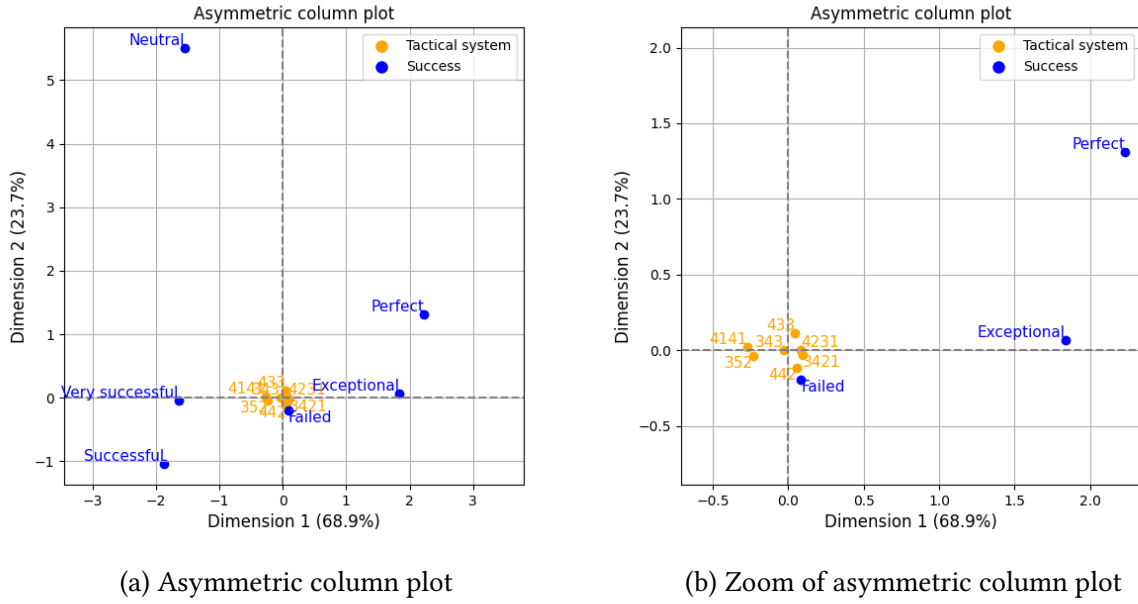


Figure 4.13: Asymmetric column plots for the correspondence analysis of tactical system

4-2-3-1 and 3-4-2-1 are the most likely to give rise to an *Exceptional* or *Perfect* build-up phase, when they do not give rise to a *Failed* phase. However, they are among the tactical systems closest to the *Failed* category. It is therefore risky to use these systems.

The correspondence analysis therefore suggests that the success of a build-up phase is related to the tactical system. We will attempt to confirm this using a chi-square test of independence. By performing the test on the contingency table between the degree of success of a build-up phase (rows) and the formations (columns), we obtain a p-value below the significance level of 0.05 (p-value = 0.032). The hypothesis that the degree of success of a build-up phase is independent of the tactical system can be rejected. We can therefore say that the success of a build-up phase is related to the tactical system.

This result confirms the reputation of the 4-4-2 system. Indeed, this system is known to be ideal for playing on the counter-attack [39]. When playing in a 4-4-2 formation, it is difficult to control the game. It is not a system designed to maintain ball possession. In fact, in this system, there are only two central midfielders. The team can therefore often find itself outnumbered against teams with three central midfielders. The idea behind this tactical system is therefore at odds with the success of a build-up phase where the principle is to be able to keep the ball for a long time, waiting for space to open up. We can also talk about the 4-2-3-1, the preferred system of Manchester City under Pep Guardiola during the 2023/2024 season. Although it is risky, it provides numerous passing lines and different angles to play from the back. This system is renowned for playing a possession-based game [40]. Manchester City is such a

strong team in the build-up phase that they manage to leverage the positives of the 4-2-3-1 while minimizing the risk of failure associated with this system.

4.14 League

The characteristic we are going to analyze in this section is the league in which the players play. We took into account the five biggest leagues: *Ligue 1* (French league), *Serie A* (Italian league), *La Liga* (Spanish league), *Bundesliga* (German league) and the *Premier League* (English league).

Figure 4.14 is the asymmetric column plot of the correspondence analysis of the contingency table. The rows are the different degrees of success in a build-up phase and the columns are the different leagues. The plot shows that the first dimension explains 41.5% of the variability in the data, while the second explains 35.7%. The correspondence analysis is of good quality since the first two dimensions explain more than 75% of the data variability. The sum of the entries in the contingency table is 3,063.

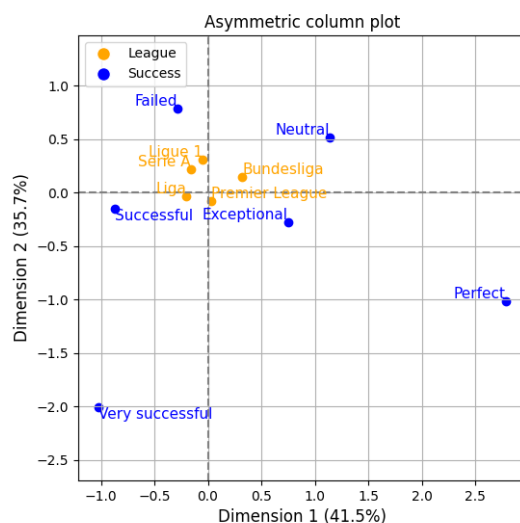


Figure 4.14: Asymmetric column plot for the correspondence analysis of league

On the asymmetric column plot, we can see that the two leagues closest to the *Exceptional* and *Perfect* categories are the *Bundesliga* and the *Premier League*. We can therefore say that these two leagues are the best in terms of build-up phase success. The plot also shows that the two leagues closest to the *Failed* category are *Ligue 1* and *Serie A*. Players from these two leagues should therefore be avoided during a build-up phase.

The correspondence analysis therefore suggests that the success of a build-up phase is related to the league. We will attempt to confirm this using a chi-square test of independence. By performing the test on the contingency table between the degree of success of a build-up phase (rows) and the different leagues (columns), we obtain a p-value below the significance level of 0.05 (p-value = $1.38e-19$). Given the low value of the p-value, the hypothesis that the degree of success of a build-up phase is independent of the league can be rejected with a high degree of certainty. We can therefore say that the success of a build-up phase is related to the league in which the players play.

La Liga is historically known for its possession and build-up phases, thanks in particular to Pep Guardiola's FC Barcelona in the early 2010s. It is no surprise, then, to find this league among the top three in terms of successful build-up phases. Historically, the Premier League is not particularly renowned for its build-up phases. However, in recent years, the arrival of certain coaches (Pep Guardiola at Manchester City, Roberto De Zerbi at Brighton, etc.) has greatly contributed to the development of build-up phases in this league. In the Bundesliga, build-up phases are known to be quicker and more direct than in other leagues.

4.15 Preferred foot of the left center back

The characteristic we are going to analyze in this section is the preferred foot of the left center back. There are only two possible categories for this feature: *Left-footed* or *Right-footed*. We are not interested in the preferred foot of the right centre back, as all the right centre backs in our sample of players are right-footed.

Figure 4.15 is the asymmetric column plot of the correspondence analysis of the contingency table, where the rows are the different degrees of success in a build-up phase and the columns are *Left-footed* and *Right-footed*. The plot shows that the first dimension explains 100% of the variability in the data. This can be explained by the fact that the contingency table contains only two columns, since a player can only be left- or right-footed. The correspondence analysis is of excellent quality. The sum of the entries in the contingency table is 838.

On the asymmetric column plot, we can see that the *Right-footed* category is closer to the *Very Successful*, *Exceptional* and *Perfect* categories than the *Left-footed* category. We can also see that the *Left-footed* category is slightly closer to the *Failed* category than the *Right-footed* category. We can therefore say that to obtain a good degree of success during a recovery phase, it is preferable to play a right-footed player in the left centre back position.

The correspondence analysis therefore suggests that the success of a build-up phase is related to the preferred foot of the left center back. We will attempt to confirm this dependence using a chi-square test of independence. By performing the test on the contingency table between the

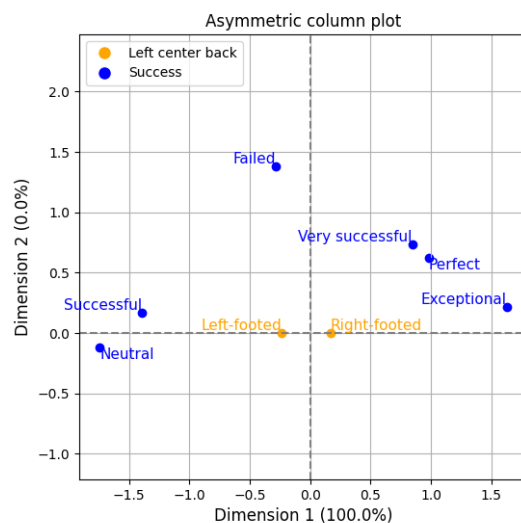


Figure 4.15: Asymmetric column plot for the correspondence analysis of preferred foot of the left center back

degree of success of a build-up phase (rows) and the left-footed and right-footed columns, we obtain a p-value below the significance level of 0.05 ($p\text{-value} = 4.71e-06$). Given the low value of the p-value, the hypothesis that the degree of success of a build-up phase is independent of the preferred foot of the left center back can be rejected with a high degree of certainty. We can therefore say that the success of a build-up phase is related to the preferred foot of the left center back.

This result contradicts the historical view that it is preferable to play a left-footed central defender. For example, Manchester United coach Erik Ten Hag says that left-footed central defenders have "an advantage in possession... because the left-footed are better orientated on the left side, in possession, you then have better angles for playing out". Mikel Arteta, Arsenal coach, said that left-footed central defenders give him more options and more solutions [41]. This result tells us that a right-footed player who is comfortable on the left foot will be just as effective, if not more so, than a left-footed player in the position of left center back. Indeed, some coaches do not hesitate to play a right-footed player at left center back, and they are right to do so: Guardiola with Manuel Akanji and Xavi (Barcelona coach) with Pau Cubarsi.

4.16 Preferred foot of the left back

The characteristic we are going to analyze in this section is the preferred foot of the left back. There are only two possible categories for this feature: *Left-footed* or *Right-footed*. We are not interested in the preferred foot of the right back, as all the right backs in our sample of players are right-footed.

Figure 4.16 is the asymmetric column plot of the correspondence analysis of the contingency table. The rows are the different degrees of success in a build-up phase and the columns are left-footed and right-footed. The plot shows that the first dimension explains 100% of the variability in the data. This can be explained by the fact that the contingency table contains only two columns, since a player can only be left- or right-footed. The correspondence analysis is of excellent quality. The sum of the entries in the contingency table is 281.

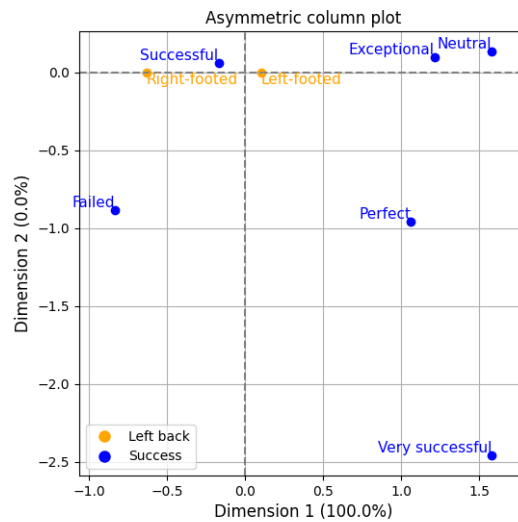


Figure 4.16: Asymmetric column plot for the correspondence analysis of preferred foot of the left back

On the asymmetric column plot, we observe that the *Right-footed* category is closer to *Failed* than *Left-footed*, and that *Left-footed* is closer to the other success categories of a build-up phase than *Right-footed*. Therefore, it can be said that it is preferable to have a left-footed player at the left-back position. However, it is important to note that this interpretation should be treated cautiously since the number of right-footed players playing at the left-back position is quite low. Indeed, the column of the table corresponding to *Right-footed* players contains only 39 items.

The correspondence analysis therefore suggests that the success of a build-up phase is related to the preferred foot of the left back but we cannot perform a chi-square test of independence because some of the cases in the contingency table do not have at least five occurrences.

In fact, the choice of playing a left- or right-footed player at left back depends on the style of build-up phases the coach wants to implement. If the coach wants to play mainly through the flanks, it will be more interesting to have a left-footed left back, as he will find it easier to pass along the line thanks to his left foot. On the other hand, if the coach prefers his build-up

phases to run through the centre of the pitch, it is preferable to have a right-footed left back, who will be more likely to play inside thanks to his right foot.

4.17 Summary of success features for a build-up phase

Table 4.1 summarises all the factors for the success and/or failure of a build-up phase analysed in this chapter.

Feature	Comments
Type of pass	Favouring ground passes is a key to success.
Passing	Favouring players with good passing skills encourages success.
Weak foot	Favouring players with 4 or 5 out of 5 weak foot increases the chances of having a Very successful phase, or better.
Skills	Favouring players with good skills moves is a key to success.
Positioning	Favouring players with poor positioning skills increases the likelihood of failure.
Vision	Favouring players with poor vision increases the likelihood of failure.
Pace	Favouring slow players is a factor for success.
Ball control	Favouring players with good ball control skills is a key to success.
Physic	Favouring players with physical qualities between 70 and 90 is a factor for success.
Agility	Favouring agile players is a key to success.
Work rate	Favouring players with a low offensive contribution is a factor of failure.
Tactical system	Using the 4-4-2 formation is a factor of failure.
League	League 1 and Serie A players should be avoided during a build-up phase.
Left center back	Contradiction between historical thinking and statistical analysis.
Left back	It is preferable to have a left-footed player, but that may also depend on the game plan.

Table 4.1: Table of success features for a build-up phase

CHAPTER 5

Illustration with real build-up phases

In this chapter, we will analyse several build-up phases of the 2022 World Cup to illustrate the features of a successful or unsuccessful build-up phase.

5.1 Example of a Very successful build-up phase

The build-up phase we are going to analyse in this section is a *Very successful* build-up phase carried out by Spain during the Spain-Germany match in the group phase of the 2022 World Cup. The phase lasts 29 second and consists of eight passes. Seven different players touch the ball. Figure 5.1 shows six images of the build-up phase. These images come from the Footballia website [11].

Of the eight passes in the build-up phase, seven are ground passes and only one is a high pass. This illustrates what was said in Section 4.1: ground passes are the most effective passes for a build-up phase. The average passing quality of the seven Spanish players who touched the ball was 77. This illustrates what Figure 4.2 shows: the category of pass quality that is closest to the *Very successful* category on the plot is the *70-80* category. As far as the quality of the players' weak foot is concerned, it is inconsistent to analyse it in this phase given that all the players used their strong foot.

With regard to the quality of positioning and vision of the players who touched the ball during the build-up phase, the average was 72 and 74.5 respectively. This illustrates what was observed in Figures 4.6 and 4.7: the *70-80* categories are the closest to the *Very successful* categories. The average pace of the Spanish players who took part in the phase was 73. This illustrates what was said in Section 4.8: you do not necessarily need very fast players to succeed in a build-up phase.

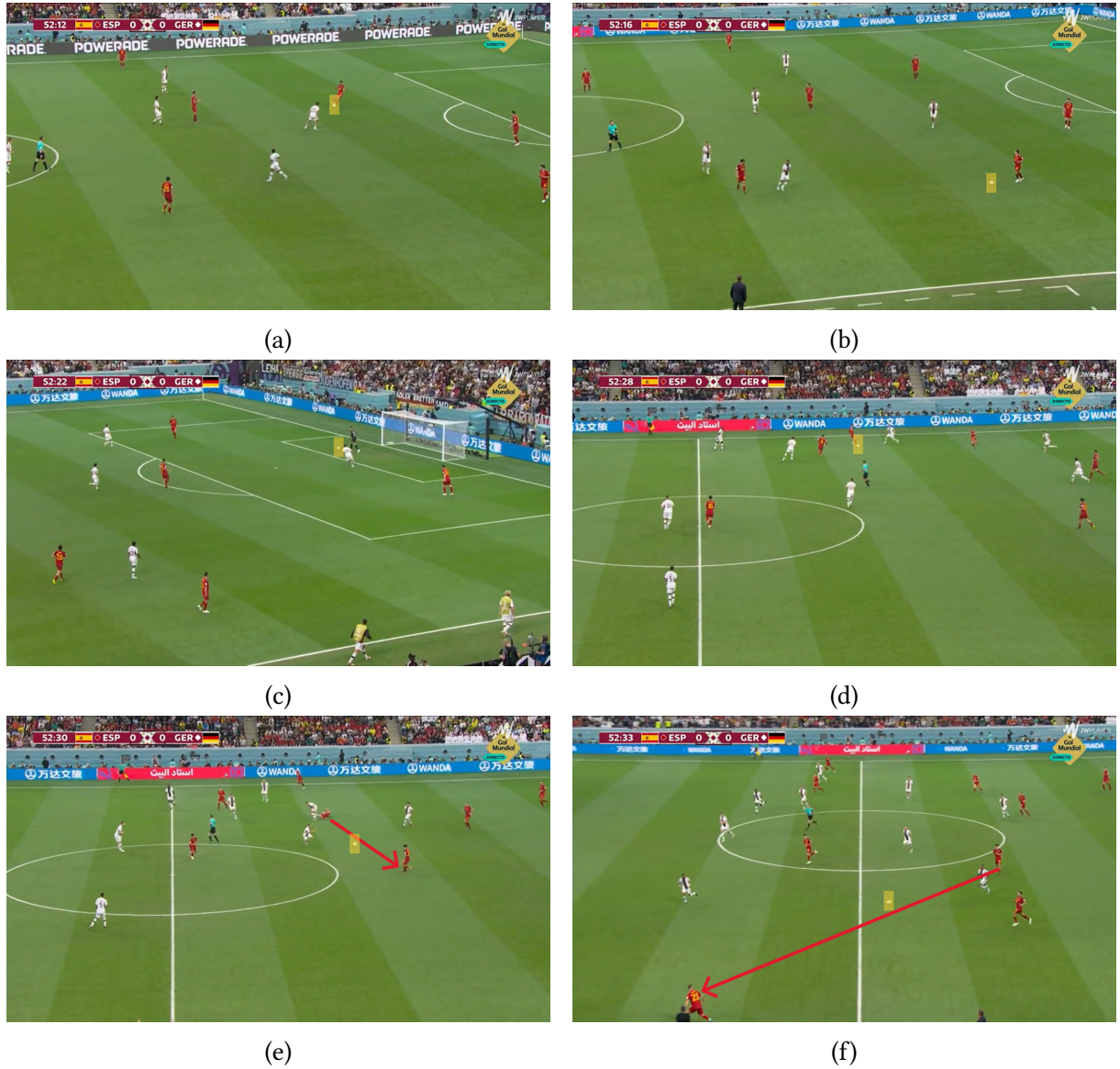


Figure 5.1: Spain's build-up phase during the Spain-Germany match at the 2022 World Cup from [11]

The average ball control quality of the Spanish players who took part in the phase was 82. This illustrates the fact that it is better to have players with good ball quality to avoid failing a build-up phase (see Figure 4.9). The average physical quality of the Spanish players who took part in the phase was 78.25. Once again, this illustrates what was said in Section 4.10: the ideal build-up phase is to have players whose physical quality varies between 70 and 90. The average agility of the players who took part in the phase was 72.75. This slightly contradicts the analysis in Figure 4.11. On this plot, the 70-80 category is one of the furthest from the *Very Successful* category.

As for the leagues in which the players taking part in the phase play, we have four players playing in La Liga, two players playing in the Premier League and one player playing in the Bundesliga. This supports Figure 4.14's analysis that these three leagues are the three best in terms of build-up phase. The left-back who took part in the build-up phase (Jordi Alba) is left-footed. This illustrates the fact that having a left-back is a factor in the success of a build-up phase. However, the left central defender (Aymeric Laporte) is also left-footed. This contradicts Figure 4.15, which indicates that it is preferable to have a right-handed player in this position.

5.2 Example of a Failed build-up phase

The build-up phase we are going to analyse in this section is a *Failed* build-up phase carried out by Canada during the Belgium-Canada match in the group phase of the 2022 World Cup. The phase lasts 20 second and consists of six passes. Four different players touch the ball. Figure 5.2 shows six images of the build-up phase. These images come from the Footballia website [11].

Of the six passes made during the build-up phase, five were ground passes and one was a high pass. It is precisely this last pass that leads to the failure of the phase. High passes should therefore be avoided, as explained in Section 4.1. The average passing quality of the Canadian players who took part in the phase was 62, which is not great. This clearly illustrates that having players with low passing quality increases the likelihood of failure in a build-up phase. The players taking part in the phase had an average quality of skills moves of 1.66 out of 5. This clearly illustrates what Figure 4.5 shows: having low quality skills moves increases the probability of failing the build-up phase.

The average positioning quality of the Canadian players taking part in the build-up phase is 40. This illustrates Figure 4.6: having players with a positioning quality of less than 40 increases the probability of failure in a build-up phase. The average vision of the players who took part in the phase was 57.33. This is a good illustration of Figure 4.7: having players with poor vision increases the probability of failure of the phase. The average pace of the players taking part in the phase is 66.5. This contradicts Figure 4.8. The plot shows that the 60-70 pace category is the furthest away from the *Failed* category. This shows that the correspondence analysis method is a good visualisation/exploration method but is not an exact science.

The average ball control quality of the players taking part in the phase was 51.66. This clearly illustrates Figure 4.9: having players with a low quality of ball control increases the risk of failing a build-up phase. The average physical quality of the players participating in the phase is 75.5. This does not particularly illustrate Figure 4.10. In fact, we can see on this plot that the 70-80 physical quality category is not the one most likely to result in a failed build-up phase.



Figure 5.2: Canada's build-up phase during the Belgium-Canada match at the 2022 World Cup from [11]

The tactical system used by Canada in this build-up phase is the 3-4-2-1. This is illustrated in Figure 4.13a, where we can see that the 3-4-2-1 is the second tactical system closest to the *Failed* category (after the 4-4-2). The 3-4-2-1 is not a tactical system conducive to a successful build-up phase. In terms of the leagues in which the players play, we cannot analyse the case of this Canadian relaunch as none of the players who took part in the phase are part of one of the five major leagues discussed in Section 4.13. As for the preferred footing of the left-back and the left central defender, we cannot illustrate anything with this build-up phase of Canada given that the tactical system with which Canada plays does not have a real left-back or left central defender.

CHAPTER 6

Conclusion

The objective of this thesis was to determine the success and failure factors of a build-up phase in football. To begin with, it was necessary to set the context for this thesis. We therefore briefly described the history of data analysis in sport and more specifically in football. We also explained why build-up phases were essential in football. It was also essential to define various key football concepts so that the reader could fully understand this thesis. This was done in Chapter 1.

After defining the context of the thesis, Chapter 2 provided a detailed definition of a build-up phase. It was important to define where and how a build-up phase could begin. In the context of this thesis, a build-up phase could begin in four different ways: by a goal kick from the goalkeeper, by a ball recovery in the first quarter of the field, by a throw-in in the first quarter of the field or by a free kick in the first quarter of the field. We also had to define how a build-up phase ends. A build-up phase could end in one of two ways: either when the team carrying out the build-up phase loses the ball, or when the team manages to pass through the half. In this case, the build-up phase was over. These definitions were made in Section 2.1. In Section 2.2, we defined the different degrees of success of a build-up phase. To do this, we used the Packing metric. We have defined six degrees of success for a build-up phase: Failed, Neutral, Successful, Very successful, Exceptional and Perfect.

After defining the build-up phase and its various degrees of success, we presented, in Chapter 3, the statistical methods we would use to analyze the build-up phases. In this chapter, we first introduced contingency tables before explaining, in Section 3.2, the detailed functioning of the correspondence analysis. In this section, we first explain how a correspondence analysis works in detail, before explaining how to interpret the results of such a method. In Section 3.3, we presented the second method used: the chi-square test of independence. This method was used to confirm certain dependency relationships between the degree of success of a build-up phase and certain features, these relationships having been suggested by the correspondence analysis.

Chapter 4 presents all the results obtained. Each section of this chapter corresponds to a feature. We will not go into detail on each section but can provide a general comment on the findings of this chapter. The correspondence analysis showed that the passes to favor during a build-up phase are ground passes, while high passes should be avoided. We also observed that the quality of players' passes had an influence on the degree of success of a phase. Another observation is that having players with a weak foot rating of 4 out of 5 or higher is ideal for a build-up phase. Regarding positioning and vision, we can say that having players with a rating above 70 reduces the probability of failure. We also observed that the degree of success of a build-up phase was related to the pace of the players involved: having slower players reduces the probability of failure. We concluded that the degree of success of a build-up phase was also related to other features such as ball control quality, physical quality, agility and work rate of the players. Concerning the physical quality of the players, we noticed that having players who are either too physical or not physical enough increases the probability of failure. Another feature influencing the success of a build-up phase is the team's tactical system. The tactical system to avoid is the 4-4-2. The degree of success of a build-up phase is also related to the league in which the players play. The leagues to avoid are Ligue 1 and Serie A, while the leagues to favor are the Bundesliga and the Premier League. Finally, we found that it is preferable to have a right-footed player in the left center-back position and a left-footed player in the left-back position.

In Chapter 5, we illustrated the results obtained in Chapter 4 using two build-up phases executed by Spain and Canada during the 2022 World Cup. These are two very different phases since Spain's phase is a *Very successful* one, while Canada's is a *Failed* phase. We observed that the results obtained in Chapter 4 were quite well illustrated through these two phases.

Before this thesis, there already existed Possession Value models such as Expected Threat, VAEP, and On-Ball Value that allowed the evaluation of each action taking place during a football match. These models enabled the assessment of actions such as dribbles, carrying, tackles, and interceptions by assigning a value to each event based on whether they increased or decreased the probabilities of scoring or conceding a goal. However, no study had been conducted on the build-up phases analyzed in this thesis. This thesis innovated by determining several factors of success and/or failure of a build-up phase by combining statistical methods (correspondence analysis and chi-square test of independence) with the Packing metric. This thesis thus allows for progress in the study of build-up phases in football.

However, this thesis has some limitations. The first limitation is the samples on which the correspondence analyses were conducted. The samples consist of all the matches from the 2022 World Cup. The results might have been slightly different if other competitions had been considered in addition to the 2022 World Cup. For that, the 360 Frame technology would have needed to be more developed. Indeed, it is currently available for only a very few competitions. Another limitation is that some of the players taking part in the build-up phases of the 2022 World Cup that we analysed are not present in the player dataset (see Appendix A.2). We

therefore have no information on these players and they are not taken into account in the results.

To conclude, we can present the various areas for improvement in this thesis. The first area for improvement would be to increase the sample size by carrying out the analyses on more matches and competitions than the 2022 World Cup. To do this, the 360 Frame technology would have to be developed and used on more competitions than it is currently.

A second area for improvement would be to differentiate between two types of build-up phase: phases with opposing pressure and phases without opposing pressure. For the moment, no such distinction has been made. If we did this, we could observe whether opposing pressure has an influence on the success and failure factors of a build-up phase. To achieve this, we would need to apply the two statistical methods¹ to phases with opposing pressure on the one hand and phases without opposing pressure on the other. It might even be possible to quantify opposing pressure and create several categories of pressure intensity. For example, there could be the following categories: "no pressing", "low pressing", "medium pressing" and "intense pressing". To measure this pressure, one idea might be to use the variable *Under_pressure* from the Events dataset in StatsBomb. The idea would be to calculate, for each build-up phase, a ratio of the number of build-up phase events carried out under pressure to the total number of events carried out during the build-up phase. The higher this ratio, the more intense the opposing pressure is considered to be.

A final point for improvement would simply be to analyse and find other factors in the success and/or failure of a build-up phase. These factors could be player characteristics. We could also analyse whether the current score of the match has an influence on the success or failure of a phase. We could try to determine which areas of the pitch are the most interesting to exploit during a build-up phase. In conclusion, this thesis opens up various avenues for exploring build-up phases in football.

¹Correspondence analysis and chi-square independence test.

CHAPTER A

Appendix : Datasets

A.1 StatsBomb

The first datasets we used were those from StatsBomb [20]. Of the five types of dataset provided by the company, we mainly used three:

- Events dataset
- Lineups dataset
- 360 frame dataset

A.1.1 Events dataset

This type of dataset provides event data for thousands of matches. During a single match, more than 3,000 events are recorded. For each event, there are more than 150 attributes. Each attribute corresponds to a column in the dataset, while each row corresponds to an event in the match. We have one dataset per match. Obviously, not all the attributes are necessary for the purposes of this thesis. Here is a non-exhaustive list of the most commonly used attributes in this thesis :

- **Index of the event** : Sequence notation for the ordering of events within each match. This attribute is useful for searching for a particular event.
- **Period of the event** : The part of the match the timestamp relates to (1 = first half, 2 = second half). This attribute is useful for differentiating between two events with the same timestamp but in two different periods.
- **Timestamp of the event** : Time in the match the event takes place, recorded to the millisecond. This attribute is useful for ordering events.
- **Type of the event** : This is the name of the event type. It's useful for knowing what type of event we're dealing with (pass, dribble, ball receipt, etc.).

- **Team of the event** : Name of the team this event relates to.
- **Player of the event** : Name of the player this event relates to. This attribute is useful for finding out which player has performed a given event and then analysing that player's characteristics.
- **Location of the event** : Array containing two integer values. These are the x and y coordinates of the event. We have an attribute for the start location of the event and an attribute for the end location of the event. These two attributes are useful for delimiting our build-up phases.
- **Duration of the event** : If relevant, the length in seconds the event lasted. This attribute is useful to have the total duration of a build-up phase.
- **Under_pressure** : This attribute is useful for determining whether or not an event is carried out under pressure.

A.1.2 Lineups dataset

This type of dataset provides information on team compositions for thousands of matches. For each match, there is a dataset in which each row corresponds to a player in the match and each column corresponds to an attribute. There are seven attributes in total, but we only use 2 of them :

- **Player_id** : The unique identifier for this player. This attribute is useful for finding the player who took part in the event.
- **Positions** : This attribute is a list. If the player has not played in the match, this list is empty. If the player has played at the same position throughout the match, this list contains a dictionary containing several keys (position_id, position, etc.). If the player played in different positions during the match, the list contains one dictionary per position. This attribute is useful for finding players who played in certain positions for certain matches that we want to analyse.

A.1.3 360 frame dataset

This type of dataset provides information about the position of players on the pitch for each event that takes place during a match. For each match, we have a dataset with each row corresponding to an event and each column corresponding to an attribute. The attributes we are interested in are as follows:

- **Id** : This attribute corresponds to the event identifier. It allows us to find a specific event within another type of dataset.
- **Location** : For each event, this attribute gives a list containing lists of coordinates [x, y] giving the position of each player visible to the camera at the time the event takes

place. This attribute is useful for calculating the packing of a build-up phase. We need the position of the players in order to calculate the packing.

- **Teammate** : For each event performed by a player, this attribute indicates whether the other players on the pitch are team-mates of the player who performed the event. This is also useful for calculating the packing of a build-up phase. For packing purposes, we need to know which players are the defenders and therefore the opponents of a team.

A.2 Players dataset

The second dataset we used was a csv containing information on over 19,000 players. This is the file *players_22.csv* from [21]. Each row of this file corresponds to a player and each column corresponds to an attribute (there are more than 100 attributes). For the purposes of this thesis, we kept only the rows and columns we needed, in order to increase the speed of code execution. The dataset we are using contains only around 850 players and the following attributes:

- **long_name** : This attribute gives the player's full name. This makes it possible to find the player in the StatsBomb event dataset.
- **league_name** : This attribute indicates the league in which a player plays and is useful for analysing the influence of the leagues in which players play during a build-up phase.
- **club_position** : This attribute indicates the position in which a player plays for his club.
- **preferred_foot** : This attribute indicates a player's strong foot and is useful for analysing the influence of a player's strong foot in a build-up phase.
- **weak_foot** : This attribute gives a score from 1 to 5 for a player's weak foot. The lowest score is 1 and the highest is 5. This attribute is used to analyse the impact of a player's weak foot in a build-up phase.
- **skill_moves** : This attribute gives a score from 1 to 5 for a player's technical ability. The lowest score is 1 and the highest is 5. This attribute is used to analyse the impact of a player's skills moves in a build-up phase.
- **work_rate** : This attribute gives the player's work rate and can be used to analyse the impact of the work rate of players in a build-up phase.
- **pace** : This attribute gives a score from 0 to 99, evaluating the player's pace and enabling the impact of the players' pace in a build-up phase to be analysed.
- **passing** : This attribute gives a score from 0 to 99, evaluating the player's passing quality, and is used to analyse the impact of players' passing quality in a build-up phase.
- **physic** : This attribute gives a score from 0 to 99, assessing the player's physical quality and enabling the impact of this to be analysed in a build-up phase.

- **skill_ball_control** : This attribute gives a score from 0 to 99, evaluating the quality of the player's ball control and analysing its impact in a build-up phase.
- **movement_agility** : This attribute gives a score from 0 to 99, evaluating the player's agility and enabling the impact of agility in a build-up phase to be analysed.
- **mentality_positioning** : This attribute gives a score from 0 to 99, evaluating the quality of the player's positioning and enabling the impact of this to be analysed in a build-up phase.
- **mentality_vision** : This attribute gives a score from 0 to 99, evaluating the player's vision of the game and analysing its impact in a build-up phase.

CHAPTER **B**

Appendix : Correspondence analysis plots

This appendix shows the symmetric plot in standard coordinates and asymmetric row plots of the correspondence analyses of the different features of a build-up phase.

B.1 Interpreting the symmetric plot and the asymmetric row plot

Before showing all the plots, it is essential to explain how to interpret these two types of plots.

The first plot is the symmetric plot in standard coordinates. The first dimension is on the x-axis and the second dimension is on the y-axis. The symmetric plot is a valuable tool for visualising and interpreting the relationships between the rows and columns of a contingency table in a space of reduced dimensions. Plot B.1 is an example of a symmetric plot in standard coordinates which has been constructed using Table 3.4 and 3.5.

The asymmetric row plot displays row principal coordinates and column standardized coordinates in the same plot. This plot is useful to look for relationships among the row and column categories and to help interpret the principal components. The closer a point for a row category is to a point for a column category, the higher the value of the row profile for the column category. Figure B.2 is an example of an asymmetric row plot.

In this example, from [10], red dots correspond to the different academic subjects (biochemistry, physics, etc.), while the blue squares correspond to the different levels of funding (A is the highest level of funding and D is the lowest. E indicates no funding). This plot can be interpreted as follows: Physics tends to have the highest funding level because the red

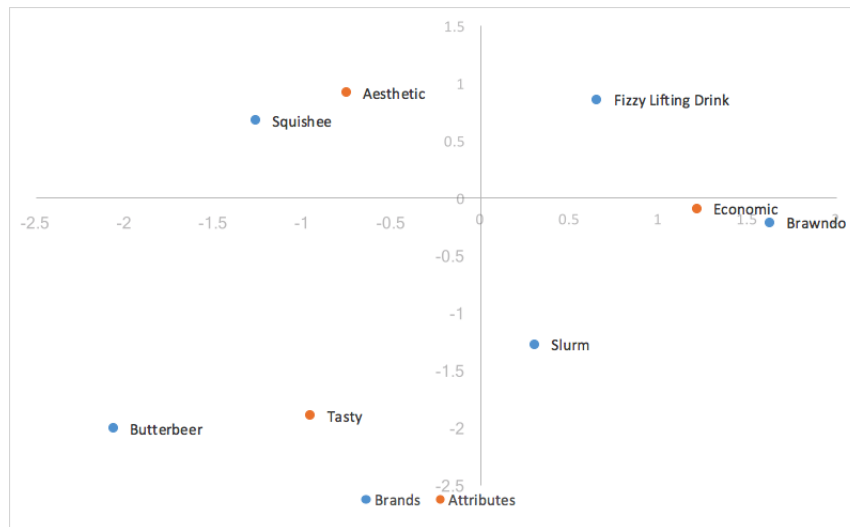


Figure B.1: Symmetric plot in standard coordinates of our example illustrating how a correspondence analysis works from [12]

point corresponding to the Physics category is the red point closest to point A. Zoology tends to have the lowest because the Zoology point is closest to category D, which represents the lowest level of funding. Biochemistry tends to be unfunded, but when it is funded, it rarely falls into the lowest funding category. The plot shows that Biochemistry is closer to categories B and C than to category D.

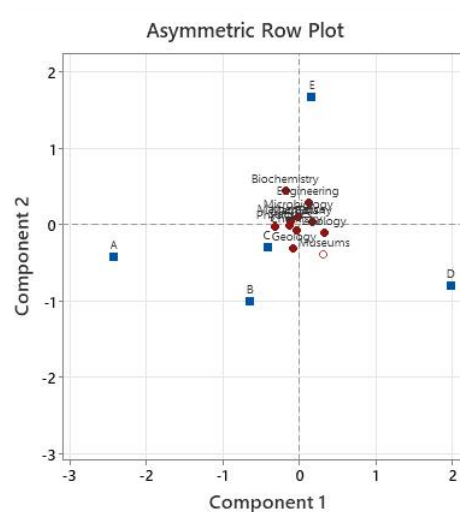
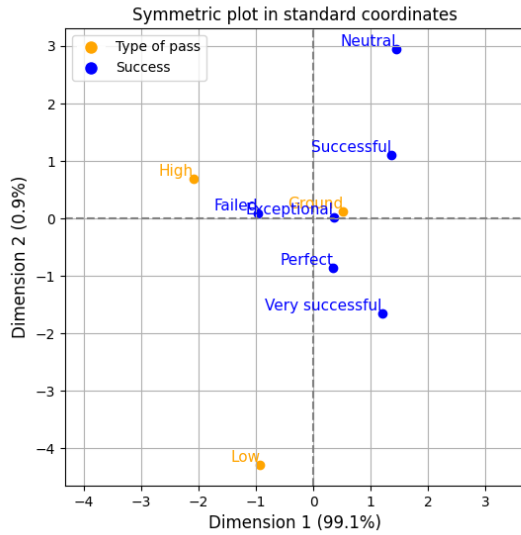
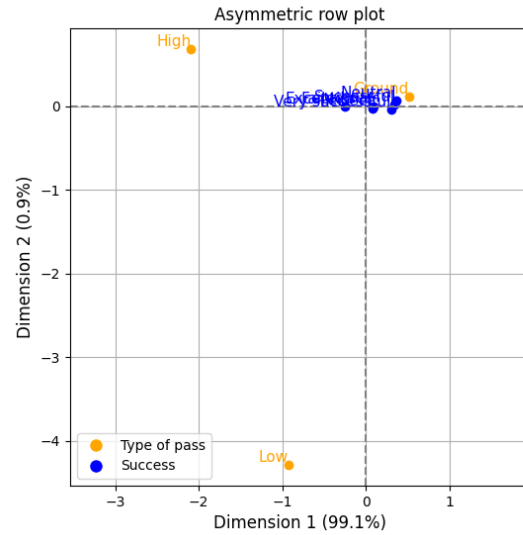


Figure B.2: Example of an asymmetric row plot from [10]

B.2 Type of pass



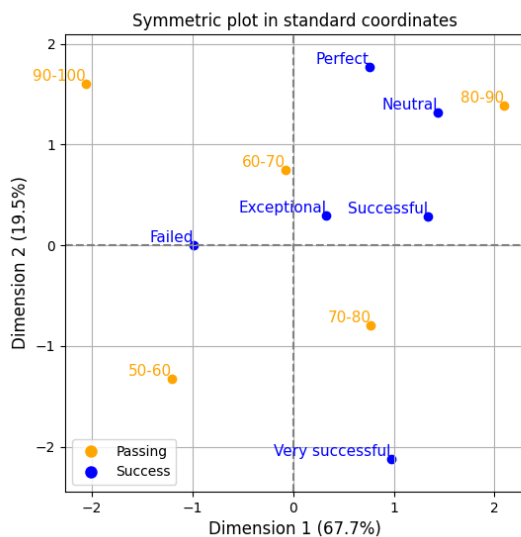
(a) Symmetric plot



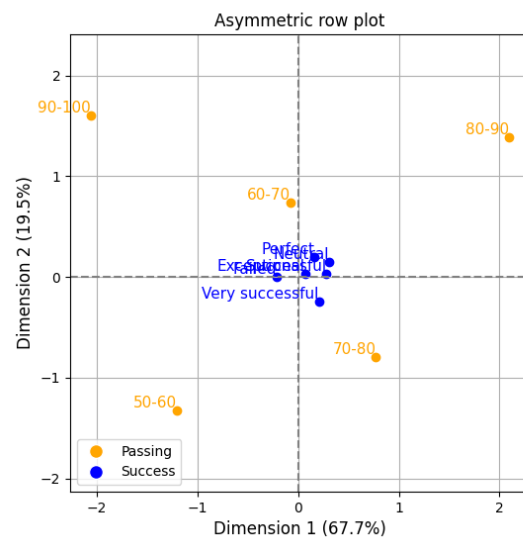
(b) Asymmetric row plot

Figure B.3: Plots resulting from correspondence analysis of the type of pass

B.3 Passing of players



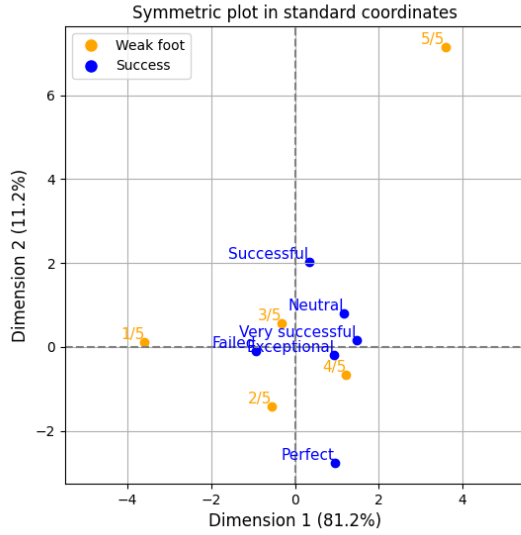
(a) Symmetric plot



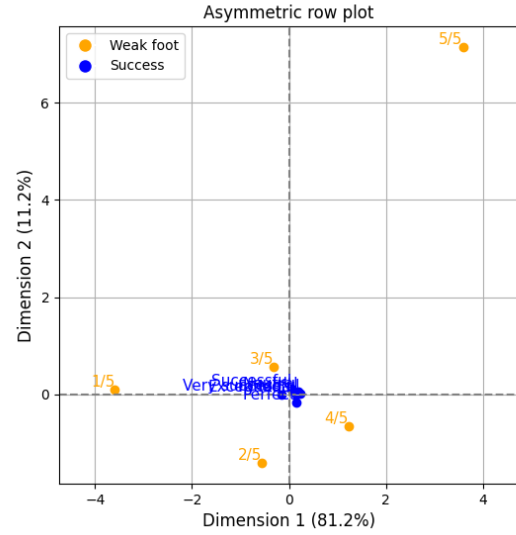
(b) Asymmetric row plot

Figure B.4: Plots for Passing of players

B.4 Weak foot of players



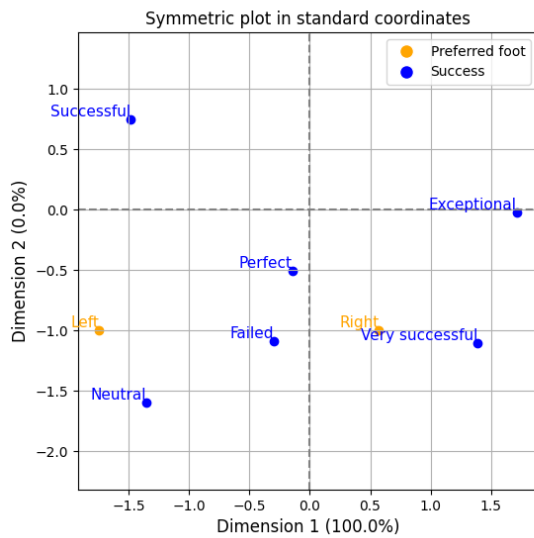
(a) Symmetric plot



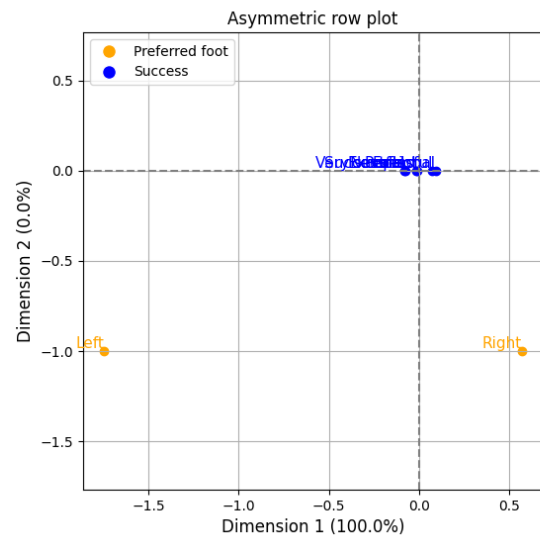
(b) Asymmetric row plot

Figure B.5: Plots resulting from correspondence analysis of weak foot of players

B.5 Preferred foot of players



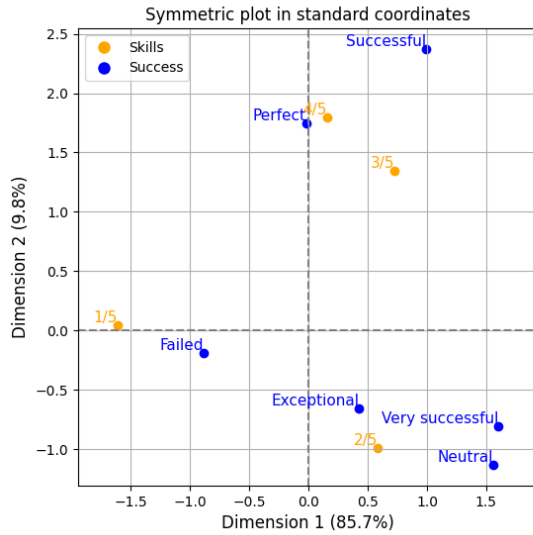
(a) Symmetric plot



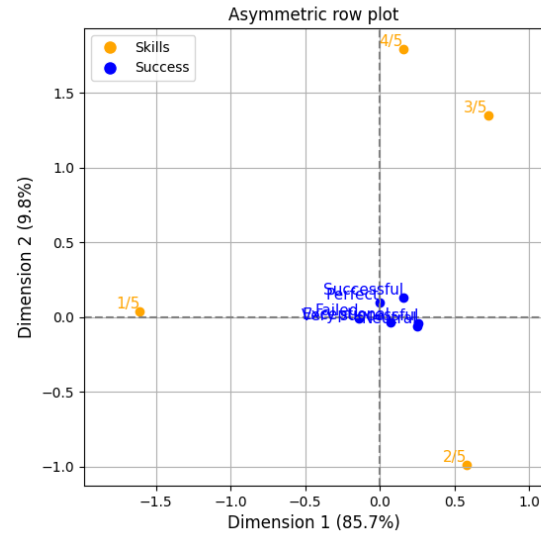
(b) Asymmetric row plot

Figure B.6: Plots resulting from correspondence analysis of preferred foot of players

B.6 Skills of players



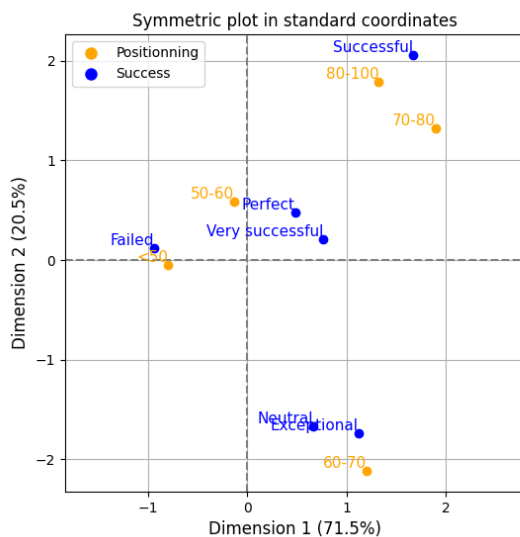
(a) Symmetric plot



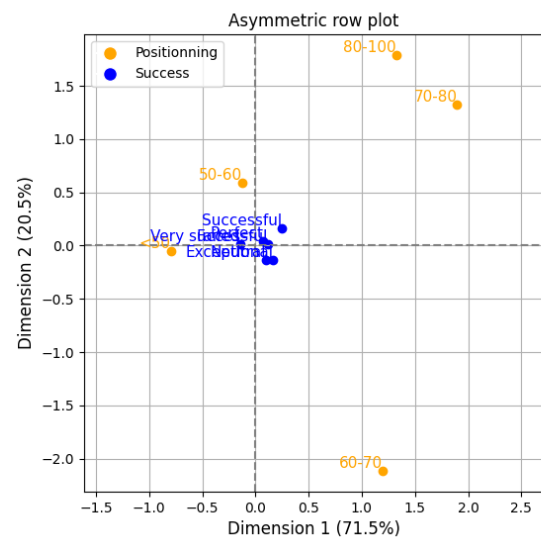
(b) Asymmetric row plot

Figure B.7: Plots resulting from correspondence analysis of skills of players

B.7 Positioning of players



(a) Symmetric plot



(b) Asymmetric row plot

Figure B.8: Plots resulting from correspondence analysis of positioning of players

B.8 Vision of players

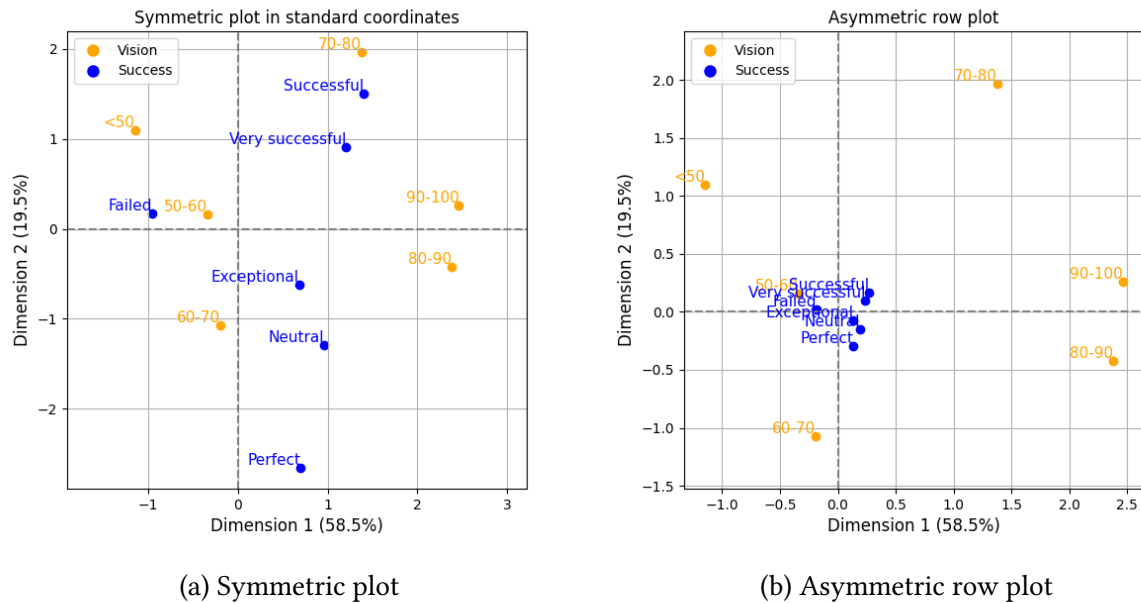


Figure B.9: Plots resulting from correspondence analysis of vision of players

B.9 Pace of players

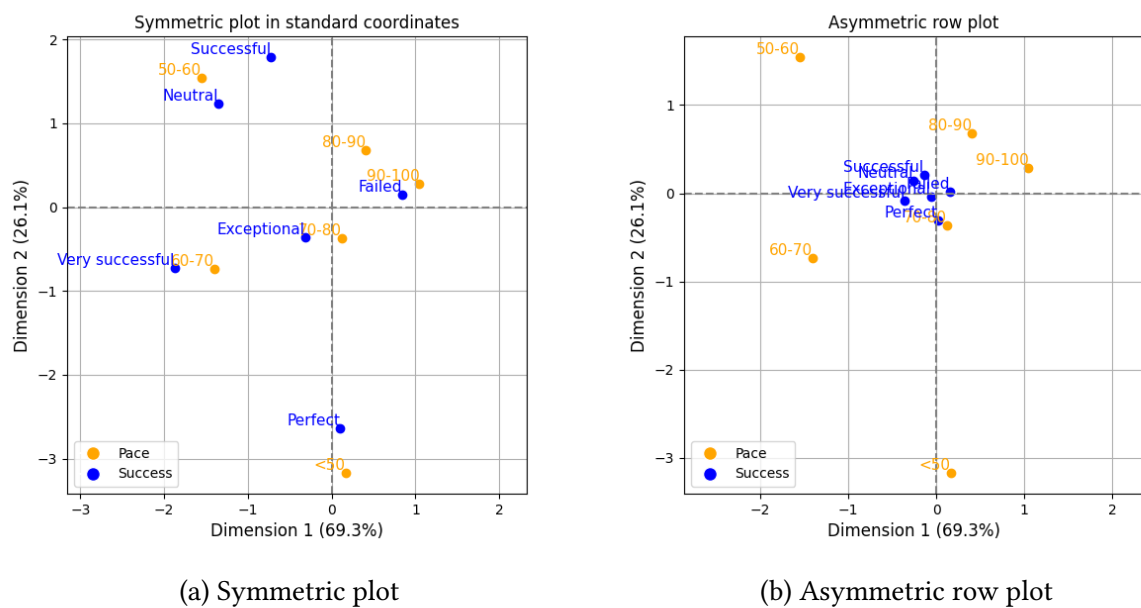
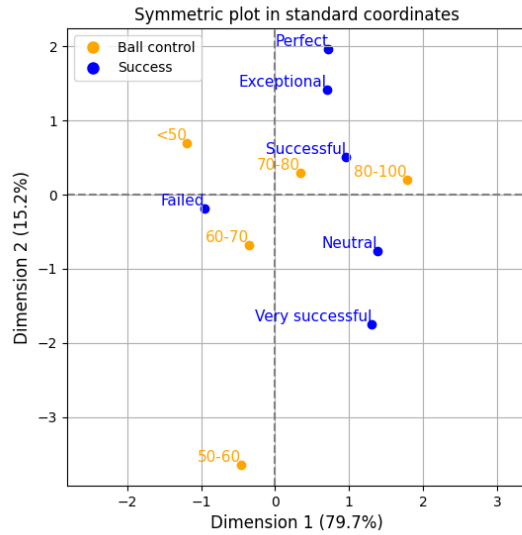
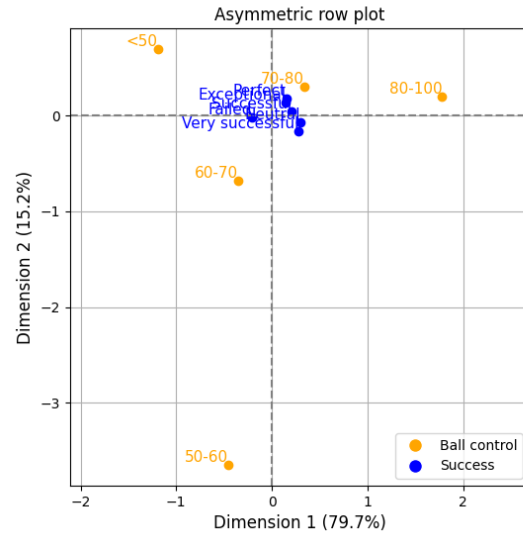


Figure B.10: Plots resulting from correspondence analysis of pace of players

B.10 Ball control of players



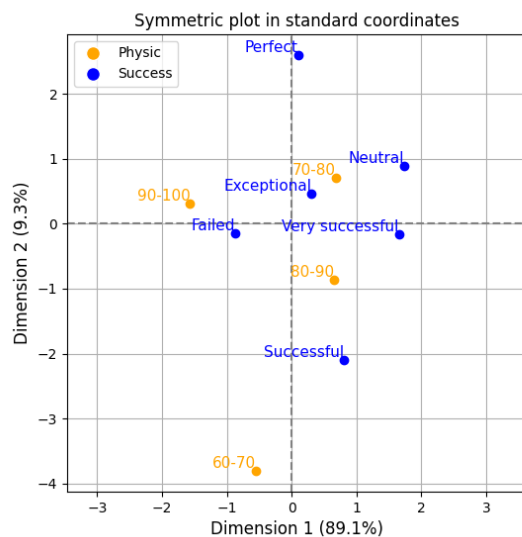
(a) Symmetric plot



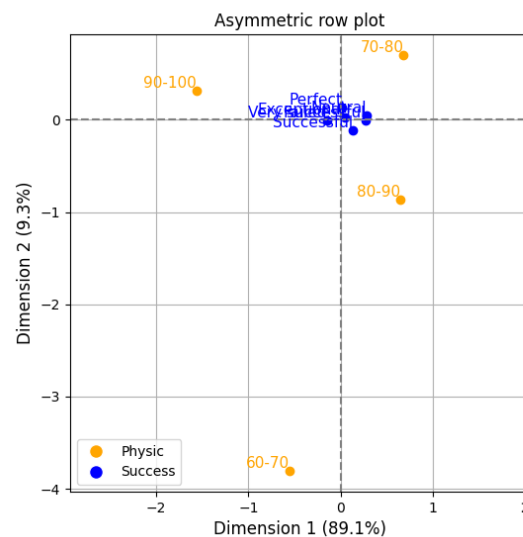
(b) Asymmetric row plot

Figure B.11: Plots resulting from correspondence analysis of ball control of players

B.11 Physic of players



(a) Symmetric plot



(b) Asymmetric row plot

Figure B.12: Plots resulting from correspondence analysis of physic of players

B.12 Agility of players

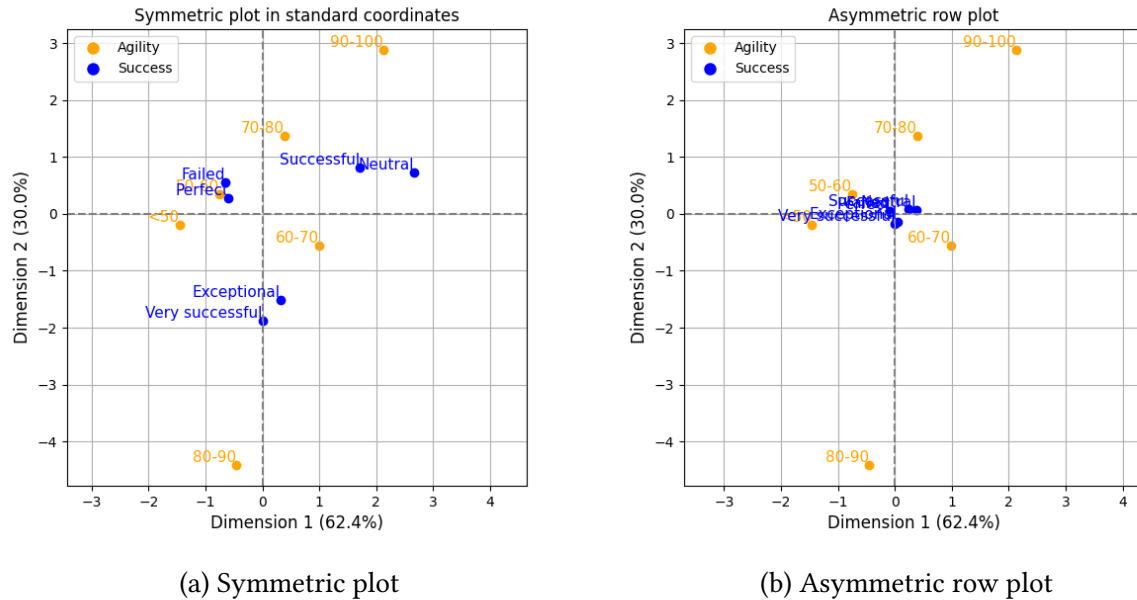


Figure B.13: Plots resulting from correspondence analysis of agility of players

B.13 Work rate of players

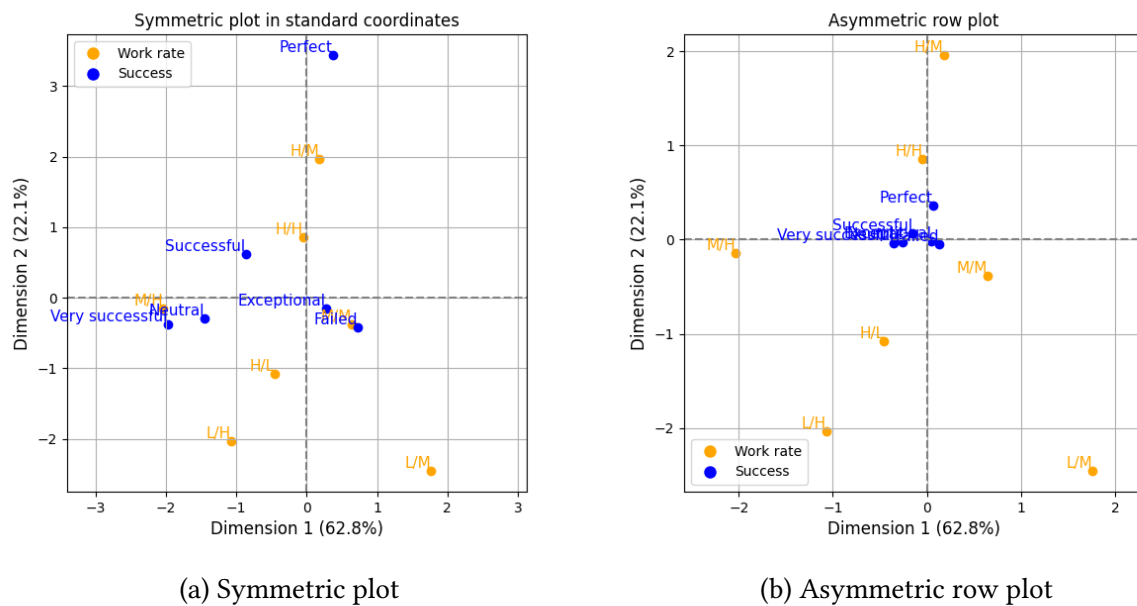


Figure B.14: Plots resulting from correspondence analysis of work rate of players

B.14 Tactical system

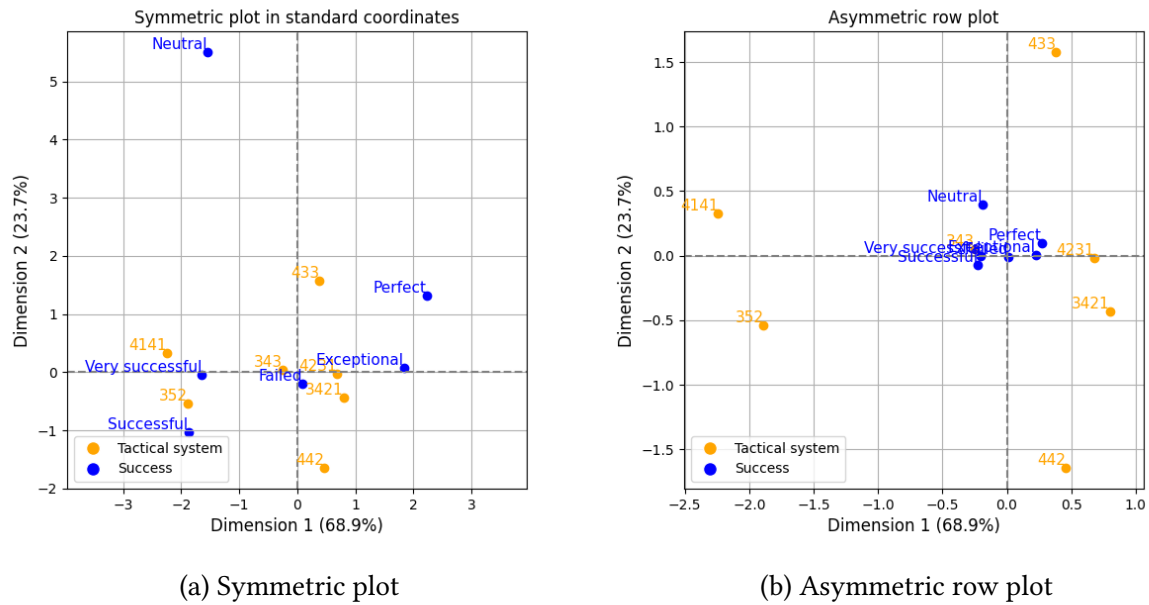


Figure B.15: Plots resulting from correspondence analysis of tactical system

B.15 League

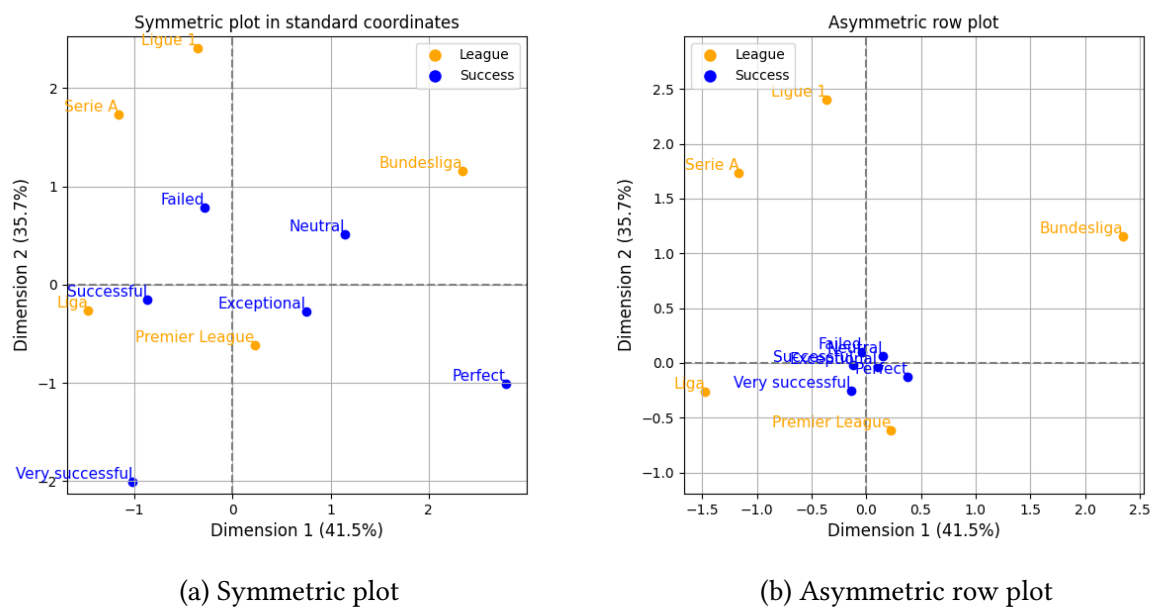


Figure B.16: Plots resulting from correspondence analysis of league

B.16 Preferred foot of the left center back

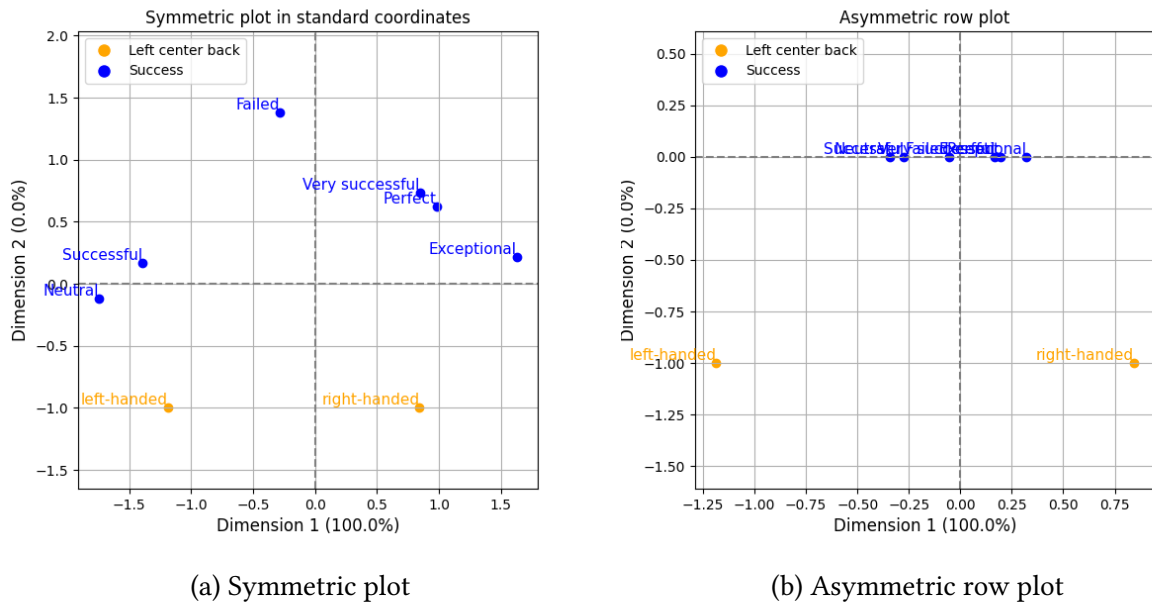


Figure B.17: Plots from correspondence analysis of preferred foot of the left center back

B.17 Preferred foot of the left back

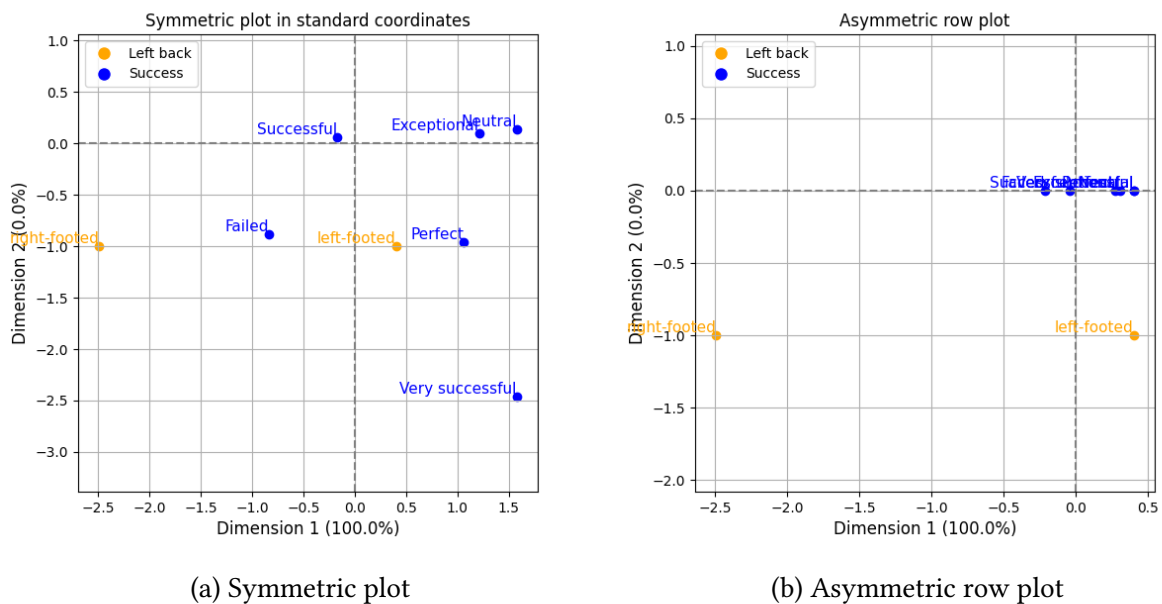


Figure B.18: Plots resulting from correspondence analysis of preferred foot of the left back

CHAPTER C

Appendix : Chi-square distribution table

<i>df</i>	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.82	9.35	11.35	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.54	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.66	23.59
10	2.15	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.75
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.21	28.30
13	3.56	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.69	26.12	29.14	31.31
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.15
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.56	40.00
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.93	36.78	40.29	42.80
23	9.26	10.19	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.88	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.37	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.32	46.93
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29
27	11.80	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.20	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34
30	13.78	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
40	20.67	22.14	24.42	26.51	29.06	33.67	39.34	45.61	51.80	55.75	59.34	63.71	66.80
50	27.96	29.68	32.35	34.76	37.69	42.95	49.34	56.33	63.16	67.50	71.42	76.17	79.52
60	35.50	37.46	40.47	43.19	46.46	52.30	59.34	66.98	74.39	79.08	83.30	88.40	91.98
70	43.25	45.42	48.75	51.74	55.33	61.70	69.34	77.57	85.52	90.53	95.03	100.44	104.24
80	51.14	53.52	57.15	60.39	64.28	71.15	79.34	88.13	96.57	101.88	106.63	112.34	116.35
90	59.17	61.74	65.64	69.13	73.29	80.63	89.33	98.65	107.56	113.14	118.14	124.13	128.32
100	67.30	70.05	74.22	77.93	82.36	90.14	99.33	109.14	118.49	124.34	129.56	135.82	140.19

Figure C.1: Chi-square distribution table from [13]

Bibliography

- [1] L. Mariaul, “Markov decision process modelling of football attacks and analysis of the optimal actions,” Master’s thesis, UCLouvain, 2022.
- [2] Scott Martin, “It’s impossible to stop Manchester City’s build-up play...so don’t – tactical analysis.” <https://totalfootballanalysis.com/team-analysis/manchester-city-2022-23-how-to-stop-their-build-up-play-scout-report-tactical-analysis-tactics>, May 2023.
- [3] Didar Kafroshi, “How Brighton build-up and progress : De Zerbi’s Brighton.” <https://medium.com/@didarkafroshi/how-brighton-build-up-and-progress-de-zerbis-brighton-438192b07ff0>, October 2023.
- [4] StatsBomb, “Game-changing Tactical Event Data.” <https://statsbomb.com/what-we-do/soccer-data/360-2/>, 2024.
- [5] Futbol AnalysR, “Tutorial: Examining shots by different pitch zones.” <https://futbolanalysr.com/post/tutorial-examining-shots-by-different-pitch-zones/>, June 2019.
- [6] Rhys Desmond, “Breaking down the phases of the game.” <https://themastermindsite.com/2022/08/26/breaking-down-the-phases-of-the-game/>, August 2022.
- [7] FIFA, “FULL MATCH: Portugal v Spain | 2018 FIFA World Cup.” <https://www.youtube.com/watch?v=OFbyNU6UQQs&t=802s>, August 2022.
- [8] Vedabrata Basu, “Contingency Analysis using R.” <https://www.datacamp.com/tutorial/contingency-analysis-r>, October 2018.
- [9] kassambara, “Principal Component Methods in R: Practical Guide.” <http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/113-ca-correspondence-analysis-in-r-essentials/>, September 2017.

- [10] Minitab, “Graphs for Simple Correspondence Analysis.” <https://support.minitab.com/en-us/minitab/help-and-how-to/statistical-modeling/multivariate/how-to/simple-correspondence-analysis/interpret-the-results/all-statistics-and-graphs/graphs/#:~:text=perform%20the%20analysis.-,Interpretation,categories%20that%20are%20more%20influential.>
- [11] Footballia, “Don’t just read about history, experience it!” <https://footballia.net/fr>, 2024.
- [12] Jeffrey Hao, “Correspondence Analysis: What is it, and how can I use it to measure my Brand? (Part 1 of 2).” <https://www.qualtrics.com/eng/correspondence-analysis-what-is-it-and-how-can-i-use-it-to-measure-my-brand-part-1-of-2/>, January 2019.
- [13] Bienvenudansladata, “Détargeting – Optimisation des leviers de conversion.” <http://bienvenudansladata.com/>, May 2016.
- [14] Wikipedia, “Sports analytics.” https://en.wikipedia.org/wiki/Sports_analytics, April 2024.
- [15] Nabil M Abbas, “Nba data analytics: Changing the game.” <https://towardsdatascience.com/nba-data-analytics-changing-the-game-a9ad59d1f116>, August 2019.
- [16] Tidiany M’Bo, “Comment la data a envahi le foot.” <https://www.konbini.com/sports/comment-la-data-a-envahi-le-foot/>, December 2021.
- [17] Jonny Whitmore, “Sequences and Possessions in football.” <https://theanalyst.com/eu/2021/03/possessions-and-sequences-in-football/>, March 2021.
- [18] Total Football Analysis, “Tactical theory: build-up.” <https://totalfootballanalysis.com/tactical-theory-build-up-tactical-analysis-tactics#:~:text=The%20tactical%20term%20build%20up,first%20third%20of%20the%20pitch.>, March 2022.
- [19] ThomasCvrK, “L’évolution du gardien de but.” <https://www.footmercato.net/a7216657453134100555-levolution-du-gardien>, June 2021.
- [20] StatsBomb, “Free Data.” <https://github.com/statsbomb/open-data>, 2024.
- [21] Stefano Leone, “FIFA 22 complete player dataset.” https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset?select=players_22.csv, 2022.

- [22] Vasilis Armatas, Joaquin Gonzalez-Rodenas, Evdokia Varamenti and Dimitris Gkoumas, “Effect of team ranking on defenders’ offensive performance in Greek top football league.” https://www.researchgate.net/profile/Vasilis-Armatas/publication/344445875_Effect_of_team_ranking_on_defenders'_offensive_performance_in_Greek_top_football_league/links/5f764574458515b7cf601ea0/Effect-of-team-ranking-on-defenders-offensive-performance-in-Greek-top-football-league.pdf, September 2020.
- [23] Sam Gregory, Sam Robertson, Robert Aughey and Grant Duthie, “The influence of tactical and match context on player movement in football.” <https://www.tandfonline.com/doi/full/10.1080/02640414.2022.2046938>, March 2022.
- [24] Pascal Bauera, Gabriel Anzera and Laurie Shawd, “Putting team formations in association football into context.” <https://content.iospress.com/download/journal-of-sports-analytics/jsa220620?id=journal-of-sports-analytics%2Fjsa220620>, March 2023.
- [25] Peter McKeever, “Identifying progressive ball carriers.” <https://www.statsperform.com/resource/identifying-progressive-ball-carriers/#:~:text=A%20carry%20is%20defined%20as,by%20Tom%20Worville%20in%202016.,> .
- [26] Wikipedia, “Goal Kick.” https://en.wikipedia.org/wiki/Goal_kick, December 2023.
- [27] L’internaute, “Définition Coup franc.” <https://www.linternaute.fr/dictionnaire/fr/definition/coup-franc/>, January 2021.
- [28] StatsBomb, “What is Expected Threat (xT)? Possession Value models explained.” <https://statsbomb.com/soccer-metrics/possession-value-models-explained/>.
- [29] Building Blocks, “Football Stats: PPDA and Packing.” <https://medium.com/@buildingblocks/football-stats-ppda-and-packing-a750a0df18ef>, January 2023.
- [30] Logan MacLeod, “Packing: The football metric that explains how good Brighton’s Adam Lallana and Pascal Gross are.” <https://www.sussexexpress.co.uk/sport/football/brighton-and-hove-albion/packing-the-football-metric-that-explains-how-good-brightons-adam-lallana-and-pascal-gross-are-3181219>, March 2021.
- [31] Ashwin Raman, “From Footballer to Moneyballer: The Story of Stefan Reinartz.” <https://thefutebolist.wordpress.com/2016/09/14/from-footballer-to-moneyballer-the-story-of-stefan-reinartz/>, September 2016.
- [32] Cambridge Dictionary, “Outfield player definition.” <https://dictionary.cambridge.org/fr/dictionnaire/anglais/outfield-player>.

- [33] Setex Société, “Tableau de contingence : définition, utilité et marketing.” <https://fiches-pratiques.e-marketing.fr/Thematique/agences-1329/FichePratique/a-quoi-sert-un-tableau-de-contingence--365354.htm>, March 2023.
- [34] “Gitbut containing the codes.” https://github.com/Adri1simon18/Football_Build-up_Phase, May 2024.
- [35] Wikipédia, “Analyse factorielle des correspondances.” https://en.wikipedia.org/wiki/Correspondence_analysis, October 2023.
- [36] Wikipedia, “Correspondence analysis.” https://en.wikipedia.org/wiki/Correspondence_analysis, November 2023.
- [37] JMP statistical discovery, “Chi-Square Test of Independence.” https://www.jmp.com/en_au/statistics-knowledge-portal/chi-square-test/chi-square-test-of-independence.html#:~:text=The%20Chi%2Dsquare%20test%20of%20independence%20checks%20whether%20two%20variables,idea%20is%20plausible%20or%20not.
- [38] Wikipedia, “P-value.” <https://en.wikipedia.org/wiki/P-value>, May 2024.
- [39] Jobs in football, “The 4-4-2 Formation: Its Uses, Benefits, and Weaknesses.” <https://jobsinfootball.com/blog/tactics/4-4-2-formation/>, May 2022.
- [40] Ali Tweedale, “The 4-2-3-1: football tactics explained.” <https://www.coachesvoice.com/cv/the-4-2-3-1-football-tactics-pochettino-guardiola-flick-southgate/>.
- [41] Liam Tharme, “How tactics impact transfers: Left-footed centre-backs, inverted wingers and ball-playing goalkeepers.” <https://www.nytimes.com/athletic/4636234/2023/07/04/tactics-transfers/>, July 2023.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain

Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/epl