PhD thesis

New methodological approaches to better understanding the attacking process during open play moments in football

Gwon, U.

___

Full bibliographic citation: Gwon, U. 2022. New methodological approaches to better understanding the attacking process during open play moments in football. PhD thesis Middlesex University

Year: 2022

Publisher: Middlesex University Research Repository

Available online: https://repository.mdx.ac.uk/item/10yzv1

___

NEW METHODOLOGICAL APPROACHES TO BETTER UNDERSTANDING THE ATTACKING PROCESS DURING OPEN PLAY MOMENTS IN FOOTBALL

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNGHO GWON

LONDON SPORT INSTITUTE
FACULTY OF SCIENCE & TECHNOLOGY
MIDDLESEX UNIVERSITY

November 2022

ABSTRACT

Attacking team possessions in football has been classified in terms of the number of passes, duration or using descriptive labels such as build up and direct play. However, little consideration has been given to the fact that teams probably utilise all possession types at various times and can be a consequence of teams playing with a high press or low block. In addition, playing styles are normally described only when football teams attack during open play moments i.e. not when a shot occurs directly from a set piece as these are very different situations involving specialist formations for each team. Furthermore, not all possessions will necessarily aid the description of playing style e.g. low duration ones, whereas some may be more complex than to be categorised into one style. The three studies in this thesis aimed to describe the attacking process using novel methodologies, during open play moments which started when the ball was regained or from set pieces that developed into open plan situations. All studies assessed all matches involving the first and second placed teams (Manchester City and Liverpool) in the 2018-19 English Premier League season. Study 1 defined indeterminate team possessions as possessions that did not allow the attacking intent to be determined. These accounted for 25% of the data and were excluded from the two step cluster analyses. This analysis detailed the attacking process in terms of where possession started, how the ball progressed forward and where the possession ended, including whether a resultant shot or goal occurred. Manchester City and Liverpool differed on 6 of the 16 derived clusters with results showing that Manchester City tended to build up play from the back and had more possessions that started in the final third of the pitch compared to Liverpool. The second study used an additional variable to classify how a team tried to break the opponent's defensive line. If this was successful, but the team subsequently passed the ball back behind the opponent's

defensive line, this was coded as a new phase of the same team possession. Data mining was used to produce association rules for the sequential possession data and decision trees then used to determine which sequential patterns produced the most shots at goal. Association rules allowed the identification of repeated possession sequences for different specific situations. For example, Manchester City had 34 different attacking patterns (Liverpool 33) after patterns with rule support of less than 1% were removed. Manchester City's most used team possession was build-up play culminating in a pass to the wide area. Liverpool utilised both fast transitions from defence to attack and build-up play with the significantly greater use of long passes from behind than Manchester City. The third study assessed the attacking strategies of the two teams under the combined influence of all levels of four contextual variables (match status, match venue, opposition quality and time period). Decision tree models for each team, confirmed that match status was the most important contextual variable, followed by match venue. Post hoc Node diagrams illustrated how the possessional phases changed in these decision tree models. Both Manchester City and Liverpool typically maintained similar possessional phases throughout matches, suggested to reflect their ability to control matches. Manchester City's did increase their use of direct play at the expense of fast breaks during losing situations in the second half of matches, albeit a relatively unusual occurrence. Liverpool tended to have more fast breaks in winning situations and during the last 30 minutes of matches there was an increased use of direct play. The fine-grained analysis presented in this thesis has aided the closing of the so-called theory-practice gap. However, future studies should consider the most appropriate methods to examine multiple individual teams within an analysis that maintains practical significance.

ACKNOWLEDGEMENTS

First, my appreciation is to my admirable parents, my father, Seokmook Gwon and my mother, Jaegyo Jeong have always given their boundless love and unconditional support during my PhD journey. They have encouraged and made me feel not alone during this difficult journey, particularly when I was frustrated during the difficult times of study. I owe them a huge debt of kindness and support.

Secondly, I really thank my best friends, Jungu Kang and Youngwon Shim. They have cheered me up during low moments and always encouraged me to be confident in my abilities. They also helped and took care of many things when I had to be in South Korea. At many times I could not have concentrated on my studies without their help.

I would also like to thank Nimai, my second supervisor. Nimai always found the weaknesses in my studies and gave critical feedback, this was very helpful for improving the quality of my studies. In addition, he managed the application process for conferences, helped with administrative paperwork and gave me the opportunity to supervise an MSc student.

Finally, I am most grateful to my supervisor, Professor Nic James. If he were not my supervisor, I would not be able to finish this journey. Over the four years, he was willing to help and gave time to me so that I could adjust to the PhD life, particularly when I was struggling with speaking and writing English. He always gave me big support and courage whenever I had problems. In addition, I have learned many things and developed my research skills from his insight into the knowledge and logical approaches to research. It was an honour and a great experience to receive his guidance during the PhD journey and I hope I can also be a doctor like him in the future.

PUBLICATIONS FROM THIS THESIS

Journal submission (under review)

1. Ungho Gwon, Nic James, Jongwon Kim, Goran Vučković & Nimai Parmar (2022). Discriminating possession types in football to reflect influencing factors.

Conference presentations

1. Ungho Gwon, Nic James & Nimai Parmar (2021). The Impact of indeterminate possession on assessments of team playing styles in football. Research paper presented at the 2021 MDX Research Students' Summer Conference (RSSC), London, UK, June. Published in the book of Abstracts of the 2021 MDX Research Students' Summer Conference, London, UK, pp. 111.
2. Ungho Gwon, Nic James, Jongwon Kim, Goran Vučković & Nimai Parmar (2022). A new phase approach to analysing possession in football. WCPASS 2022 & IACSS 2022 Join-Congress, Vienna, Austria, September.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

OVERVIEW OF THESIS

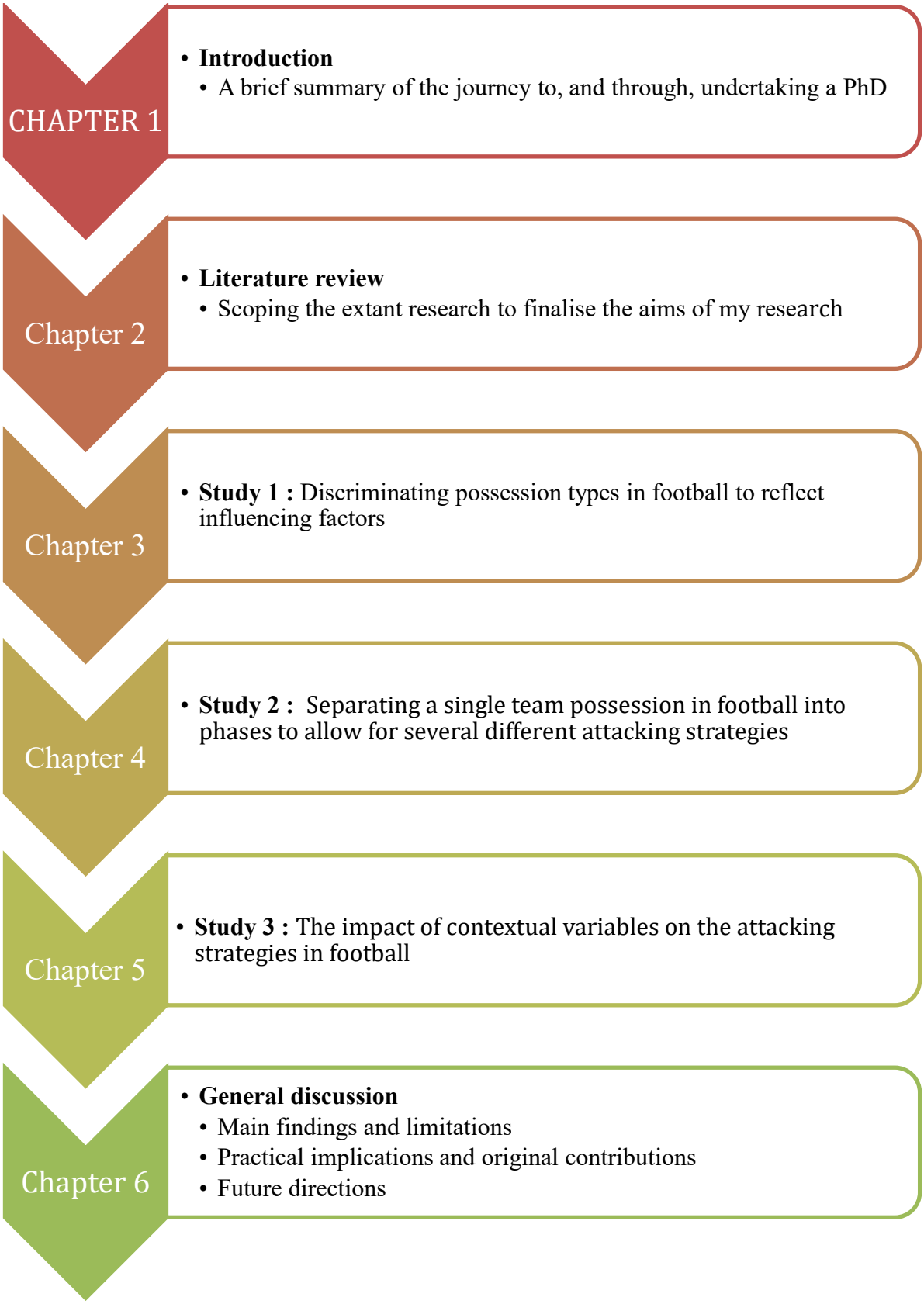


Figure 0.1. A brief overview of the chapters of the thesis

Chapter 1: Introduction

It was clear that Performance Analysis, as a discipline, was undergoing change, with the advancement of computer science, particularly big data methodologies, being discussed and utilised more frequently. Given the complexity of the attacking process and the relatively simplistic methods used in the extant literature, it was decided to explore the use of novel methodologies (developed in computer science) to better understand this complex process. At this early stage, it was clear that a relatively large number of team possessions did not conform to a playing style e.g. when ball possession was quickly lost without the chance of an attacking behavior taking place. This phenomenon was not clear in the research papers and the first study was devised to clarify this within a classification of possession styles using cluster analysis. Hence, the first study (Chapter 3) defined indeterminate possessions and presented the difference in the classification of possession styles depending on whether indeterminate possessions were included or not.

Following the first study it was considered that defining each single team possession into one style of play was not sufficient. It was recognised that sometimes a team's attack, using one style of play, could end and rather than lose possession the team could "reset", and start the attack again, often using a different style of play. To try to validate this belief, a new concept was defined called possessional phases where a "reset" signified a new phase of possession. Thus study 2 (Chapter 4) presents this new approach using Association rules to reveal these sequential patterns.

The final study (Chapter 5), tried to provide a more practical outcome for the approaches used in the first two studies. Using combinations of all levels of four situation variables this study sought to address the original aim of closing the theory gap in football. This study used Decision tree models with post hoc Node diagrams to

illustrate how the possessional phases changed in the different levels of the situation variables.

Whilst this short summary of my PhD journey sounds logical and relatively straight forward it was, of course, not the case. Many difficulties had to be overcome, defining the variables was an arduous task; discerning, learning and utilising the most appropriate statistics was challenging and a global pandemic was thrown into the mix for good measure! However, these hurdles, ultimately made the completion of the journey more pleasurable and worthwhile.

Chapter 2: Review of literature

2.1 Introduction

Football is a team invasion game where the objective is to score more goals than the opponent. In the simplest terms, this means performance can be divided into offensive, or attacking play, in possession of the ball trying to score and defensive play, without possession trying to prevent the opponents from scoring. Most performance analysis research in football has considered attacking play from the perspective of trying to understand how goals are scored and what the best methods for doing so are. This type of research has differentiated "build-up play" where the ball is passed between teammates as the ball progresses up the pitch relatively slowly from "direct play" where the emphasis is on taking a shot at goal quickly. Wade (1996) suggested that build up play by midfield and defensive players could be deemed the preparation phase of the game and hence a third phase of the game. Whilst this view has some merit, the fact that this distinction is not apparent in either common football terminology or academic research suggests that each possession tends to be described as one of the two attacking methods (build up and direct). Whilst different names have been used for these playing styles, they have been used so often in both research and applied contexts, although definitions vary somewhat, that they are no longer questioned as being accurate depictions of a "style of play", although they are not the only methods of attack (Figure 2.1).

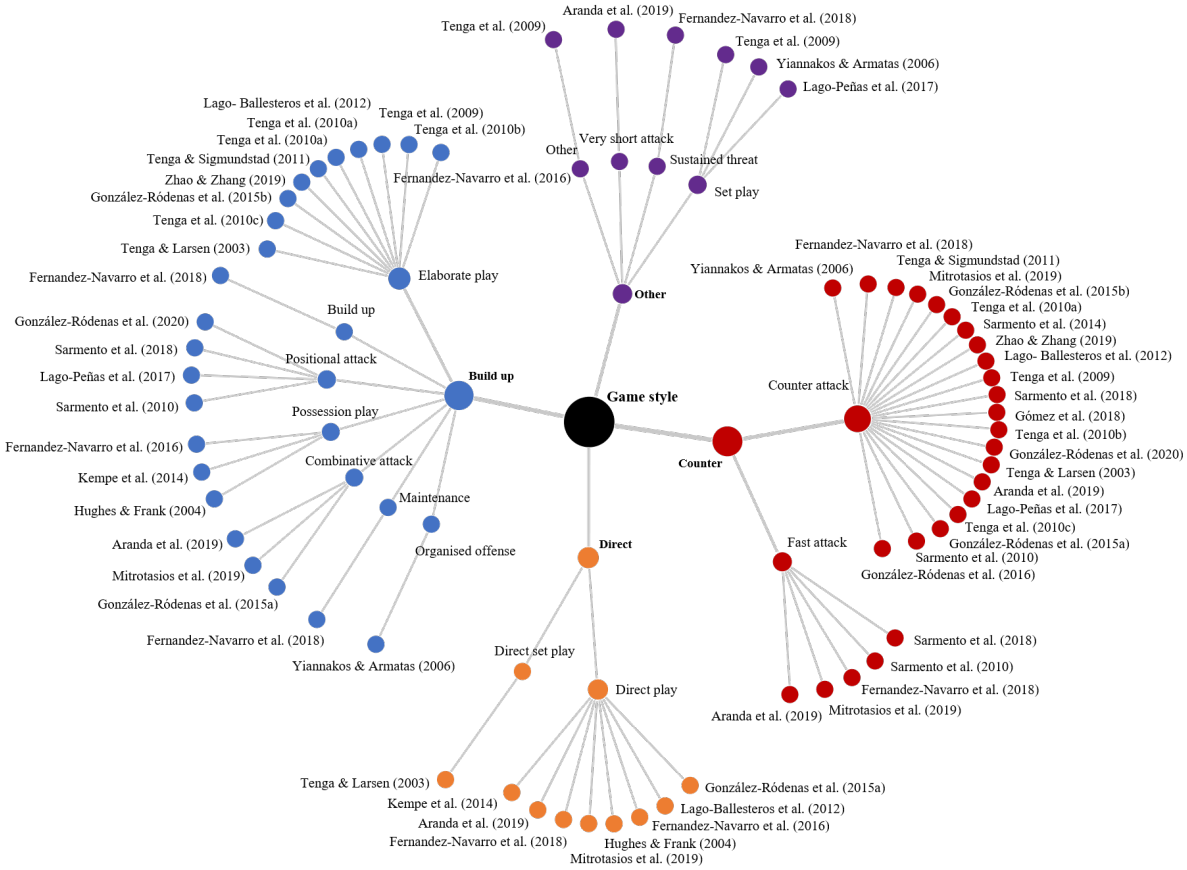


Figure 2.1 The categorisation of game styles used in previous studies

Recently, Hewitt *et al*. (2016) presented the case that football should be considered within the context of different moments of play. The suggestion was that elaborate attacks could, and should, be differentiated from transitions from defence to attack. Initially one could be excused for thinking that elaborate attacks are simply another name for build-up play and transitions simply direct play or counter attacks. However, the authors maintained that all attacking play is influenced by the corresponding defensive play of the opponent. To elucidate, if a team regains possession of the ball from a defender in their own half it is likely that the team will try to score as quickly as possible i.e. to take advantage of the advantageous situation. How this "transition" moment manifests itself is highly determined by where the transition takes place. For example, Vogelbein *et al*. (2014) found that ball recoveries

close to the opponent's goal produced seven times as many goals compared to recoveries in the defensive areas. It therefore follows that teams who play with a "high press" defensive strategy i.e. attacking players are committed to chasing down the ball carrier whilst trying to anticipate likely pass directions to regain the ball in the opponent's defensive areas, are likely to have relatively many direct play opportunities. Similarly, a defensive strategy of "sitting back" i.e. defending deep inside their own half (commonly referred to as a low block) would tend to result in a low total possession percentage and more opportunities for counter attacks (since the opponents would be largely in forward positions). However, these possessions could be characterised as "long ball" i.e. direct plays or counter attacks, depending on the operational definitions used.

Hewitt *et al*. (2016) also presented the case for set piece plays to be considered as a separate moment of the game. This is logical since set pieces (free kicks, corners, throw-ins and penalty kicks) often allow both teams to position themselves specifically for the type of set piece e.g. a defensive wall for a free kick close to the penalty area or man marking and zonal positions at corners. Set pieces account for approximately one third of all goals in elite football (Yiannakos & Armatas, 2006; Wright *et al*., 2011) and set pieces that take place in the opposition half are widely regarded as potential goal scoring opportunities. Indeed, professional teams spend a lot of time considering how to set up the players for these situations using formations that are very different to other moments of the game. Further evidence lies in the fact that league tables for performance on set pieces are routinely used in the media (e.g. Opta) to highlight perceived strengths and weaknesses of teams in this specific moment of the game.

The plan for this research is to discern the different playing styles exhibited by two highly ranked English Premier League (EPL) teams using moments of the game identified during ball possession. Secondly, a rigorous selection procedure is required to determine the most appropriate variables for identifying playing style of teams, potentially within each of the different moments. The following literature review has provided evidence to support the proposed variables with acknowledgement of Hewitt *et al*.'s (2016) definition of a game (playing) style: -

> "Game style is the characteristic playing pattern demonstrated by a team during games. It will be regularly repeated in specific situational contexts such that measurement of variables reflecting game style will be relatively stable. Variables of importance are player and ball movements, interaction of players, and will generally involve elements of speed, time and space (location)".

2.2 Literature search

The Web of Science and PubMed databases were searched for relevant papers up until October, 2022. The key words football OR soccer AND performance AND possession were used to only include journals written in English. When duplicate articles were removed 583 studies remained. A two-step screening process involved reading the title and abstract, which excluded 341 non-relevant studies and the remaining 242 studies read fully to exclude a further 193 studies. Hence the 49 articles that formed the basis of the literature review were finalised (Figure 2.2). Additional studies were identified through the reference lists and papers that cited the selected papers and manual searches of relevant sports science journals' web pages e.g. Journal of Sports Science, International journal of Performance Analysis in Sport and the European journal of Sport Science (N = 37).

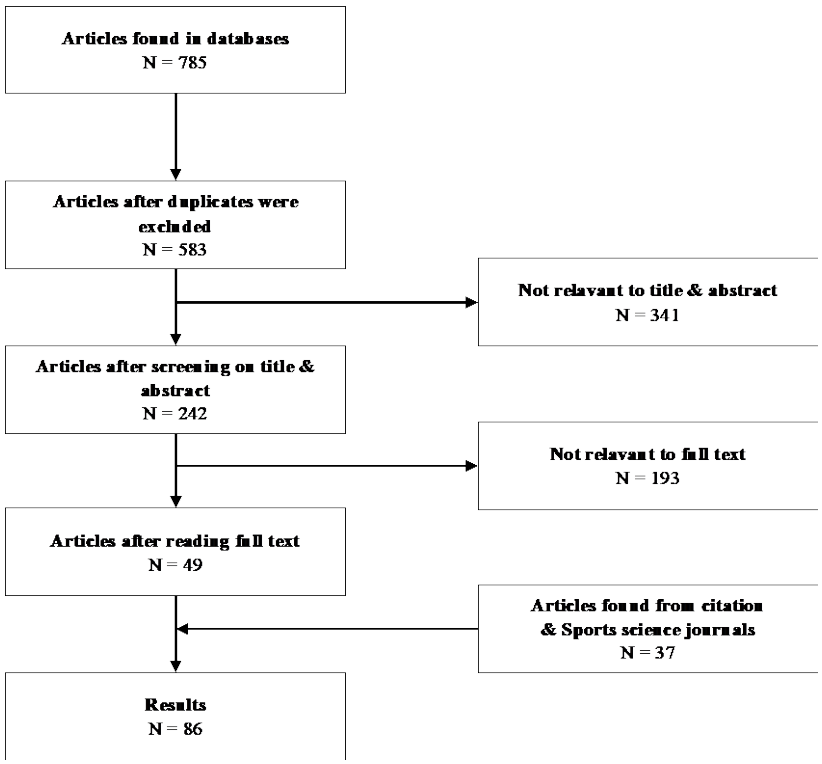


Figure 2.2 Flow chart of systematic search for relevant literature

2.3 Research on how to play football successfully.

In football, understanding the patterns of play exhibited within a game can help coaching be more specific and objective to improve tactical performances of teams (Tenga *et al*., 2015). A "pattern of play" may be considered as behaviours that are exhibited by a team, to obtain both attacking and defensive objectives, on multiple occasions during a match (Fernandez-Navarro *et al*., 2016). Since the objective of coaching is to improve performance, an understanding of a team's patterns of play can aid development of tactical strategies to enhance a team's performance (James *et al*., 2002). This can be methods to counteract an opponent's important patterns of play as well as to enhance and develop one's own patterns of play to create more goal scoring opportunities. Similar terminology has been used in the literature to depict the fact that teams use repeated patterns of play, some of which are encapsulated as a "playing style". For example, a "long ball team" would imply that the team likes to play long passes from defensive to attacking areas, bypassing the midfield players. Different playing styles have been the cause of great controversy over the history of football and have been considered important factors which can influence the outcome of matches (Yiannakos & Armatas, 2006).

Previous studies have typically analysed patterns of play in fairly superficial ways. For example, a team's playing style, suggested as representing the overall tactical approach, has been measured using the number of passes (Reep & Benjamin, 1968; Bate, 1988; Hughes & Franks, 2005) or the duration (James *et al*., 2002; Jones *et al*., 2004, Lago-Peñas & Martin, 2007; Lago-Peñas, 2009; Lago-Peñas & Dellal, 2010; Garratt *et al*., 2017) of each team possession. However, these are fairly crude measures which only distinguish playing styles to a limited extent. They fail to acknowledge that teams can utilise a variety of different possession lengths and

provide no detail of, for example, the types of passes, movement characteristics and pitch areas exploited.

Other studies have focused on how goals were scored to try to discriminate, successful or otherwise, patterns of play. For example, Yiannakos and Armatas (2006) stated more goals were scored from organised offences (44.1%) than set piece (35.6%) or counter attacks (20.3%) in the 2004 European Championships. This study only analysed passages of play resulting in goals and hence no information was generated regarding how successful or otherwise different types of possession were. This point was succinctly made by Hughes and Bartlett (2002) who specified that performance indicators need to be contextualised e.g. 10 goals from 30 set pieces (1/3) is more effective than 20 goals from 80 (1/4) open play sequences. A second issue related to analysing competition data relates to the fact that the external validity of the results (how representative they are of other football matches) is questionable. This is because International competitions consist of relatively small numbers of matches played against teams of variable quality (See Table 2.1). Hence if significant players did not play e.g. injured or rested for less important matches such as a group stage match, performance variability in these matches would have a more dramatic impact on the results than for larger data sets.

Previous studies analysed domestic league, competition or mixed events to analyse team performance (Table 2.1). The usefulness of each approach can be considered varied particularly in relation to sample sizes (number of matches analysed). Studies analysing competitions typically assessed a smaller number of matches compared to those looking at domestic leagues, but the mixture of teams was greater. Hence, team level findings from competitions are unlikely to be very useful given that the low frequency of events means that teams change between events.

However, if the purpose of the study was to examine the way in which football was played, then from a general perspective, varied samples are useful.

Table 2.1. Sample size of previous studies

	Matches	Domestic League		Competition		Mixed (n=5 papers cited)
		Single league (n=25 papers cited)	Multi Leagues (n=5 papers cited)	Club (n=6 papers cited)	International (n=8 papers cited)	
Reference (n=49)	< 30 (n=13 papers cited)	Jones *et al*. (2004), Lago (2009), Sarmento *et al*. (2010), Lago *et al*. (2012), Gama *et al.* (2014), Gonzalez-Rodenas *et al.* (2015b), Kawasaki *et al.* (2019), Fernandez-Navarro *et al*. (2020)	-	Paixão *et al*. (2015), Hughes and Lovell (2019)	Casal *et al*. (2017), Aranda *et al*. (2019)	James *et al.* (2002)
	< 70 (n=11 papers cited)	Taylor *et al.* (2008), Bradley *et al.* (2014)	Sarmento *et al*. (2014)	-	Yiannakos & Armatas (2006) Clemente *et al.* (2015b), Clemente (2018), Konefał *et al*. (2018), Aquino (2019), Alves *et al*. (2019)	Sarmento *et al*. (2014), Sarmento *et al*. (2018)
	<240 (n=12 papers cited)	Redwood-Brown (2008), Tenga *et al*. (2010a), Tenga, *et al*. (2010b), Tenga *et al*. (2010c), Paratas *et al*. (2016), Garratt *et al*. (2017), Lago-Peñas *et al*. (2017b)	Fernandez-Navarro *et al*. (2016), Lago-Peñas *et al*. (2017a), Mitotasios *et al* (2019), Gonzalez-Rodenas *et al* (2020)	-	Hughes and Franks (2005)	-
	240≤ (n=13 papers cited)	Lago and Dellal (2010), Gómez *et al*. (2012), Gómez *et al*. (2018), Gollan *et al*. (2018), Fernandez-Navarro *et al*. (2018), Bilek and Ulas (2019), Zhao and Zhang (2019), Lepschy *et al*. (2020)	-	Liu *et al*. (2015), Yi *et al* (2019), Yi *et al*. (2020)	-	Reep and Benjamin (1968), Kempe *et al*. (2014)

Some studies have looked at passages of play immediately prior to and following goals being scored in comparison to the rest of the passages of play. The idea being that teams may change something in their play which results in the goal being scored. Redwood-Brown (2008) found that scoring teams had significantly higher passing accuracy in the 5 minutes before a goal was scored compared to their

average for the half. This also corresponded to the conceding team playing significantly less passes than compared to their average for the half. This study was to be commended in that it was the first to split match events up in this way. It clearly showed that events prior to a goal being scored were different to other events although it could not be ascertained whether the changes that resulted in the goal being scored were a tactical ploy, an isolated piece of skill or luck or perhaps a mistake by the defending team. As with the International competition studies, the use of very small data samples, in this case 5 minutes of play where a goal resulted, increased the impact that the successful play had on the sample. To some extent this meant that the results were inevitable and perhaps not as informative as first thought.

In terms of sample size, the first, and probably most influential, football research paper had a data set consisting of 3213 matches originating in 1953 (Reep & Benjamin, 1968). This paper has widely been regarded as determining that direct play, also known as the long ball game, was superior to elaborate (possession football). Since this style of play is now thought less effective, many people deem Reep and Benjamin's work flawed although it is likely that it was the interpretation that was erroneous rather than the work itself. Passing distributions i.e. the number of passes in a single team possession, for a large number of possessions were analysed and when plotted, exhibited a negative binomial distribution (Figure 2.2).

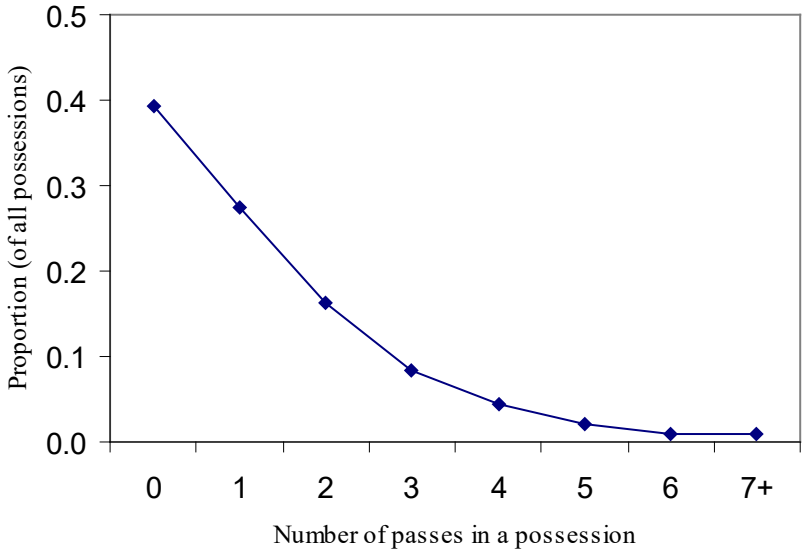


Figure 2.3 Relative incidence of possessions according to the number of passes played

Of critical importance to the interpretation of these results is the observation that small samples do not necessarily follow this distribution (as shown by Reep & Benjamin for 12 matches played by Arsenal during 1961-2, Table 2, page 582). This detail tends to be overlooked in comparison to the finding that events in football matches are very predictable when large data sets are used. Statisticians call this the "law of large numbers" (first proved by Swiss mathematician James Bernoulli in 1713). Reep and Benjamin thus produced their best-known findings that it takes on average 10 shots to score one goal, 50% of goals are scored from possessions that involve one pass or less (Note: zero pass possessions include penalties and free kicks) and regaining possession in the opponent's half provides many goal scoring opportunities. Whilst football has changed dramatically since this data was collected these key findings remain pretty similar. However, it was the way in which these results were interpreted that caused much consternation. Many football coaches and researchers deemed that because more goals were scored from possessions with low numbers of

passes it was therefore advantageous to have more short pass possessions. This was also believed to be true by Charles Reep who devised, sometimes very successful, tactics around this logic. However, this logic is flawed, since if 80% of possessions involve 3 passes or less and if 80% of goals are scored from these possessions then it is logical that the number of passes in a possession is irrelevant to the prospect of a goal being scored. Hughes and Franks (2005) attempted to provide evidence for this logic by analysing all matches played at the 1990 and 1994 World cups. Recognising that there were more possessions with low numbers of passes in comparison to those with high numbers they calculated goal scoring frequencies per 1000 possessions. Goal scoring frequencies were dramatically altered such that very similar rates of goal scoring were found for all pass possession lengths. Hughes and Franks (2005) suggested that their work supported the notion that direct play (short passing sequences) would be more efficient for less skilful teams whereas more skilful teams would favour elaborate play (long passing sequences). This is logical given that less skilful players are more likely to lose possession of the ball compared to more skilful players and hence short passing sequences offer less opportunity to lose the ball.

The measures used to distinguish playing styles have been limited by their reductionist nature (Mackenzie & Cushion, 2013) i.e. trying to understand tactical intent (which is complex and differs during different phases of the game) using a limited selection of variables. Furthermore, there is limited information in previous research that recognises that not all possessions are meaningful and demonstrate tactical intent. Jones *et al*. (2004) stated that possessions which lasted less than 3 seconds, were not analysed as they were not related to strategic performance (Jones *et al*., 2004). However, this rule would have discounted some possessions that did demonstrate strategic intent e.g. quick attacking plays immediately following a

turnover. Tenga *et al*. (2009) excluded 0 pass possessions and Aranda *et al*. (2019) removed very short duration attacks which were not able to be categorised by the observer. Merlin *et al*. (2020) also excluded ball possession sequences less than 4 seconds that did not contain at least one successful pass. Whilst these four papers recognised the need for excluding certain possessions from the analysis they did not provide any discourse on the reasoning for decisions to exclude possessions or not. The vast majority of research papers, however, included all possessions and did not consider whether short possessions such as those coded as 0 passes or between 0-3 passes were in fact possessions where the team did not manage to develop a meaningful attack (e.g. Reep & Benjamin, 1968; Bate, 1988; Hughes & Franks, 2005; Tenga *et al*., 2009; Tenga *et al*., 2010a; Tenga *et al*., 2010c; Tenga & Sigmundstad, 2011; Lago-Ballsteros *et al*., 2012; Paixão *et al*., 2015; González-Ródenas *et al*., 2016; Garratt *et al*., 2017).

Successful teams score more goals than unsuccessful ones, otherwise they wouldn't be more successful. However, the reasons for scoring more goals is less clear. Jones *et al*. (2004) assessed whether successful EPL teams had longer possessions than unsuccessful teams and whether match status (whether a team was winning, drawing or losing at the time of the possession) had an effect. They found that successful teams had longer possessions than unsuccessful teams irrespective of match status and both sets of teams had longer ball possessions when losing compared to winning. This novel research identified that teams changed the way that they played due to match status and hence identified an important variable for consideration when analysing football matches. The authors contended that longer possession durations by successful teams were likely to be skill oriented as these players were able to keep possession of the ball better than less skilful players. This tends to support Hughes and

Franks (2005) suggestion that more skilful teams would favour elaborate play and seems prophetic given this style of play has more recently become the style of choice for the best teams in the world e.g. Barcelona, Manchester City etc. These findings have been replicated many times. For example, longer possessions where found when teams were losing compared to winning or drawing and ball possessions decreased in frequency and duration when playing against stronger opponents (for 27 Spanish professional football matches during the 2005 – 2006 season (Lago-Peñas, 2009), 380 matches in 2008 – 2009 Spanish La-Liga (Lago-Peñas & Dellal, 2010) and 380 matches in 2015 – 2016 English Premier League (Fernandez-Navarro *et al*., 2018). Bradley *et al*. (2014) found that high percentage ball possession teams (HPBPT) and low percentage ball possession teams (LPBPT) had similar duration ball possessions when drawing and losing, possessions decreased when winning but to a greater extent for LPBPT compared to HPBPT. Casal *et al*. (2017) found that successful teams had longer ball possessions than unsuccessful teams, moreover, their possessions occupied the middle offensive area longer than unsuccessful teams (see also Konefał *et al*., 2018 and Bilek & Ulas, 2019).

2.4 Defining playing styles

Tenga *et al*. (2011) assessed the characteristics of goal scoring possessions that originated from open play situations in Norwegian elite football i.e. no set pieces. They used the type and number of passes, duration of ball possessions and the pitch area where the possession originated as variables to characterise a playing style. The authors also provided descriptive definitions for the playing styles counter attack (direct play) and elaborate attack (build-up play) which related to the intent of playing directly towards goal or not. These did not correlate perfectly with length of possession as, for example, an elaborate attack could involve only a few passes. Since only goal

scoring possessions were evaluated the efficacy of possessions with different characteristics could not be ascertained. Similarly, individual team playing styles could not be differentiated beyond the fact that all teams seemingly employed each type of attack at some point in time.

Tenga *et al*. (2010a) analysed Norwegian elite football to take into consideration the effect of the opposition. They compared 203 team possessions leading to goals with 1688 randomly selected team possessions (control group). They found that the proportion of goals scored during counterattacks (52%) was higher than during elaborate attacks (48%) whereas in the control group elaborate attacks (59%) were used more than counterattacks (41%). This suggests that counterattacks were more effective than elaborate attacks, backed up by a logistic regression analysis that showed this to be true only when playing against an imbalanced defence (which tends to happen more often for counterattacks compared to elaborate play). Tenga *et al*. (2010c) measured the effectiveness of counterattacks compared to elaborate attacks in a further (very similar) study where, again, 163 Norwegian elite football matches were analysed. The same 203 goals, as Tenga *et al*. (2010a), were compared with 1688 randomly selected team possessions (control group). In this study, the effectiveness of the two attacking styles were compared for different possession outcomes i.e. goals, scoring opportunities and score box entries. They found that counterattacks were more effective than elaborate attacks for all three outcome measures suggesting that scoring opportunities and score box entries could be used as proxy measures for goal scoring in future studies. The advantage being the greater frequency of scoring opportunities and score box entries compared to goals scored resulting in a large reduction in sample size needed for inference testing.

Whilst the majority of academic literature has defined attacking play into two distinct styles of play, some research has provided further distinctions. For example, Sarmento *et al*. (2010) described three styles of play, counter attack, fast attack and positional attack. Unfortunately, this paper failed to provide operational definitions for these classifications, as well as for the other methodological categories used, but it is likely that positional attack equated to elaborate attacks (build up play). The authors therefore differentiated fast attacks (direct play) from counter attacks (also direct play) with the most likely reason being the position on the pitch where the possession originated. This is logical, as previously mentioned, possessions originating in the defensive third can, by using a long pass, quickly transition into a situation where an attacking player has possession of the ball in the attacking third of the pitch. This would not be deemed as a counter attack if the pass took place after the opponents had had sufficient time to organise their defensive formation. Using this definition, it therefore suggests that a counter attack takes place in situations when the opponents had not had sufficient time to organise their defensive formation. This view, therefore, considers, that an offensive playing style needs to consider the speed of the attack, the starting position of the possession and the organisation of the defending team (see also Tenga *et al*., 2010a).

The concept of defensive organisation was defined by Tenga *et al*., (2010b) as involving three features. First defensive pressure was deemed to be tight (required for a balanced defence) if the first defender was estimated to be within 1.5m of the ball carrier at each moment of attempting to win or receive the ball. If the first defender (excluding goalkeeper) was outside 1.5m then the pressure was considered loose (required for an imbalanced defence) although during a possession both of these situations could be present and hence the pressure defined as mixed. The second

measure, defensive backup, referred to a second defender being within 5m of the first defender. For this measure, a balanced defence needed this player, whereas imbalanced defences did not meet this criterion. Finally, defensive cover referred to the presence (balanced) or absence (imbalanced) of a third defender. The authors acknowledged the difficulty in accurately assigning these categories to match footage recorded for broadcast, mainly due to difficulties in measuring distances. However, the study purportedly found that long possessions (5 or more passes) were more effective than shorter ones (less than 3) for producing score-box possessions against balanced defences although two separate analyses (univariate and multivariate) were used with conflicting results. Irrespective of this, the recognition that offensive and defensive styles are irrevocably linked is a strength of this research and needs to be considered for future studies.

Lago-Ballesteros *et al*. (2012) analysed playing styles in a similar way to Tenga and colleagues (Tenga *et al*., 2010a; 2010b; 2010c; 2011). 908 team possessions that started from open play i.e. no set pieces, by one first division Spanish team were categorised as either producing a score-box possession (33.4%), achieved progression (52.5%) or failed to reach any sort of progression (14.1%). Their analyses found that direct attacks and counter attacks produced three times as many score-box possessions than elaborate attacks. Further analyses also revealed that when the team was drawing or winning the probability of achieving a score-box possession reduced (43% and 53% respectively) compared to when losing. This work supported the contention that match status is an important variable related to how teams play (e.g. Jones *et al*., 2004) but also provided their definitions of the different playing styles. In their view elaborate attacks involved "meticulous build-up" play, counterattacks were "fast direct attacks with few players" and "one or two players should rush forward to

support the player who receives the ball in plenty of space" and direct attacks involve the team "consciously plays the ball forward to the opposition's third and tries to finish quickly". These definitions did not include pitch location information with the distinction between the two fast attacks being that counterattacks only involve two or three attacking players detached from the rest of the team because of a pass into space, presumably close to the opponent's defensive line. However, the authors did assess the impact that the initial pitch location of a possession had on the outcome with possessions starting in the pre-defensive and pre-offensive areas (these two areas combined forms the middle third of the pitch), 2.5 and 19 times respectively, more likely to produce a score-box possession that possessions originating in the defensive third. This paper also found that possessions were 4.4 times less likely to produce a score-box possession when against a balanced defence in comparison to an unbalanced one.

Sarmento *et al*. (2014) categorised four offensive processes for the counter attack; 1) start of the Offensive process, 2) end of the offensive process, 3) area where the most important action was performed, and 4) interaction contexts. They also categorised the opponent status in relation to the number of defenders and the area of the pitch; 1) relative numeric inferiority, 2) absolute numeric inferiority, 3) absolute numeric superiority, 4) relative numeric superiority, 5) equal numeric under pressure and 6) equal numeric unpressured. Using these categorisations, they performed sequential analyses to find correlations between the probability of prior and subsequent events. Results suggested that Barcelona had a tendency towards offensive sequences which developed mainly from the central areas, particularly the right side of the defensive midfield zone, through the execution of dribbles and passes. Whilst this study adopted a different type of analysis and considered more variables than

previous studies the analysis of only one offensive situation limited the impact of the results.

The studies described above have provided relevant information regarding team playing styles albeit with significant questions unanswered. For example, no study seemingly provided the detail required by football teams to be practically useful, the so called "theory-practice gap" (Mackenzie & Cushion, 2013). Details regarding which players perform the important actions, in which areas of the pitch and with what frequency tend to be overlooked. Instead, the studies explained team playing styles using a few isolated performance indicators, such as goals and shots, usually focussing on specific situations without regard for contextualising the findings. In addition, it is not obvious in the literature that a single possession could involve more than one playing style. For example, if a team gained ball possession and tried a long forward pass (direct play) but the opponents managed to contain the threat resulting in the attacker in possession of the ball passing back to a defender who then initiated build up play. This situation might be defined as direct play in some studies as this was the original playing style even though the possession involved build up play. The literature tends to be unclear about this type of situation and hence the suggestion here for the need for more clarity regarding ball possessions. It is thus recommended that studies consider the possibility of multiple playing styles within a single team possession but also that a possession may not involve any playing style at all.

2.5 Statistic approaches of recent studies

Given the obvious limitations of studying complex behaviour using just one, or a few variables, recent studies have analysed team performance using multiple variables, using more complex statistical procedures. Rein and Memmert (2016) described this situation, stating:

> "The main limitation of the traditional notational approach is that almost all contextual information is discarded, these measures have shown weak explanatory power with limited adoption by practitioners. To circumvent this problem increasingly multi-variate approaches are being used to retain contextual information"

Rein and Memmert (2016) suggested adopting big data sets, which in their view is player movement data in conjunction with event data, to solve the limitations of traditional, event data only, approaches. This is undoubtedly true, but at this point in time, this type of research remains theoretically driven and the models used have limited ability to understand the complexity of tactical team performance in elite football (Rein & Memmert, 2016). However, using more variables whilst retaining contextual information is both a sensible suggestion and one that has gained popularity recently.

Various regression models have been utilised to determine the association between different variables and different outcome measures of performance in football (e.g. Taylor *et al*., 2008; Tenga *et al*., 2010a; Tenga *et al*., 2010b; Tenga *et al*., 2010c; Wright *et al*., 2011; Gómez *et al*., 2012; Gómez *et al*., 2016; González-Ródenas *et al*., 2016; Lago-Peñas *et al*., 2016; Lago-Peñas *et al*., 2017a; Fernandez-Navarro *et al*., 2018; Aquino, 2019; González-Ródenas *et al*., 2020). For example, González-Ródenas *et al*. (2016) analysed 452 counter attack possessions from 30 randomly selected matches in the 2014 Major League Soccer season using 14 categorical variables. Binary logistic regression showed that effective counterattacks were related to possession in the offensive zones when the initial penetration took less than 3 seconds. However, possessions involving four or more passes tended to be more effective than possessions involving less passes regardless of opponent defensive pressure. González-Ródenas *et al*. (2020) used multilevel logistic regression models to analyse the effect of contextual factors on shooting effectiveness in 80 random

European professional football league matches. Nine dimensional categories (initial penetration, initial opponent pressure, initial pitch zone, type of attack, penultimate action, finishing process, type of finishing, final opponent pressure and final pitch zone) and four contextual variables (match location, quality of team, match half and match status) impacted shooting effectiveness. This research found that shooting from the ultra-offensive zone was more effective than any other area and the probability of scoring decreased when against defensive pressure than without. Furthermore, high ranked teams had higher odds ratio for scoring goals than low-ranked team regardless of opponent defensive pressure. Whilst the results presented here were not particularly new or surprising, the use of multi-factorial statistical models such as regression start to answer the criticisms levelled at simplistic models which fail to consider the complexity inherent in a team sport such as football.

Factor analysis has also been used to assess the cumulative and interactive effects of multiple variables on football performance (e.g. Gómez *et al*., 2012; Fernandez-Navarro *et al*., 2016; Lago-Peñas *et al*., 2017b; Gómez *et al*., 2018) with z-scores used to present different scalar variables on the same scale such as to present team profiles (Greenham *et al*., 2017 and Lane *et al*., 2020). Fernandez-Navarro *et al*. (2016) found 6 attacking and defensive play factors using 19 performance indicators (possession time, pitch area information, direction of passes, crosses and shots were used for attacking indicators and ball regain areas for defensive indicators). Factor 1 (possession directness) loaded on a direct (D) style of play and possession style (P). Factor 2 (width of ball possession) was related to regain pressure and regaining the ball in both wide and central areas. Factor 3 (use of crosses) loaded on crossing (C) and no crossing (NC) variables. Factor 4 (possession with) was related to possession play in wider (WP) and central areas (NP). Factor 5 (defensive ball pressure) related

to low pressure (LP) and high pressure (HP) styles. Finally, factor 6 (progression of the attack) loaded on fast progression (FP) and slow progression (SP) variables. This study presented different playing styles for 37 Spanish La Liga and English Premier League teams based on a sample of 97 matches. It also provided evidence for more specific playing styles and the obvious utility of such an approach is to allow the comparison of individual teams. This is also the approach that coaches and applied practitioners use to prepare their team for an upcoming opponent and thus this method could be deemed to contribute towards closing the so-called theory practice gap (Mackenzie and Cushion, 2013). Gómez *et al* (2018) used factor analysis to organise and classify 62 variables with eight factors being presented. The factors were labelled ball possession, ending actions, individual challenges, counterattack, set-piece, transitional play, fouling action and free-kicks. The factor scores for ball possession, ending actions and individual challenges were shown to be significantly different for different team rankings and match location. Greenham *et al*. (2017) utilised the five moments of a game concept (set pieces, established offence, offensive transition, defensive transition and established defence) from Hewitt *et al* (2016) for use in Australian Football league, a derivative form of football. They converted 12 variables into z-scores (standardised scores with a mean of 0 and standard deviation of 1) to compare team profiles in each of the 5 moments of the game. The average z-score in each moment by a team were shown to explain characteristic patterns for individual team profiles. Whilst this reinforced the utility of this approach the authors recognised that not accounting for opponent quality limited the applicability of the findings. Lane *et al.* (2020) also used z-scores for 9 offensive and 12 defensive variables in Australian Football League matches to compare each season's offensive and defensive performance (1999 to 2019). This enabled the authors to suggest that the least

offensive and most defensive style was used in the 2019 season and that offensive play had decreased from 1999 to 2019. This study presented a good methodology for tracking average performance over time and could be utilised at an individual team level. Thus, the impact of new players, the loss of key players and potentially opponent quality, although a valid measure of this would be needed, could be ascertained with an appropriate data set.

A decision tree approach was utilised to classify team performances by Lago *et al.* (2016). Home advantage was analysed to assess the effect of scoring first on match outcome with the role of different independent variables examined. Three significant variables emerged for determining match outcome, venue of the team that scored first (home or away) the first split, second was quality of position and the third split in the decision tree was match period. 76.1% of matches were won when the away teams scored first against poorer quality opponents (57.4% against better quality opponents). Bilek & Ulas (2019) also used a similar decision tree analysis and found similar results to Lago *et al.* (2016). Scoring first was the most influential variable on match outcome for each decision tree with winning probabilities of 0.48, 0.62 and 0.86 against stronger, balanced and weaker opponents. Other factors that impacted the outcomes were the number of clearances, shots, shots on target and percentage possession, although these didn't necessarily affect all decision trees. The decision tree approach has thus been shown to be a useful technique for assessing the individual influence of multiple performance indicators on match outcome.

2.6 Conclusions

The studies presented above reflect the attempts of researchers to understand the complexities of attacking styles of play in football. However, there remains a conflict between achieving statistically meaningful results, requiring large data sets where

general rules can be ascertained, and determining practically important findings where individual team analysis is required using only relevant data i.e. historical data would likely be irrelevant due to different players and managers being involved. This is the theory practice gap alluded to by Mackenzie and Cushion (2013). Some recent studies have also only analysed one or two isolated variables even though the limitations of such an approach are well documented. However, large data sets have become more prevalent recently due to the increasing use of automatic movement capture via computer vision methods. This has led to novel approaches such as network analysis to assess pass density i.e. pitch areas where frequent passes take place between certain players (Gama *et al*., 2014; Clemente *et al*., 2015a; Clemente *et al*., 2015b; Clemente, 2018; Kawasaki *et al*., 2019 and Young *et al*., 2019).

The identification of complex tactical team performance remains a challenge (Kempe *et al*., 2014), but the various multivariate statistical approaches have shown progress is being made. Typical research designs in sports science have aimed to confirm or reject a hypothesis, formulated on theory, but the robustness of the model tends to be evaluated on the empirical evidence derived from modelling techniques developed in computer science (Goes *et al.*, 2020). This development in analytical procedures is gaining popularity in sports science but particularly so in performance analysis. However, whilst these approaches are aiding the understanding of team performance, the more fine-grained analyses, such as what conditions allow an individual player to undertake a particular pass or run to break an opponent's defensive line are largely ignored in research papers.

The aim of this thesis was, therefore, to re-define possession for a better understanding of team performance when attacking in open play moments. This was necessary given the lack of clarity e.g., how indeterminate possessions were handled,

and precision e.g., recycling the ball within a team possession, in previous studies. To achieve this, new, complex, assessments were needed to understand the attacking process more holistically including the role contextual variables such as time and match status have on a team's attacking performance. This thesis will provide new methodologies to categorise playing styles in relation to passing types, ball possession and location for two, exemplar, elite teams in the English Premier League. The use of two teams will allow comparisons to be made, accounting for situational variables such as match status, match venue, time period and team and opposition quality. However, this comparison only enables the strength of the methodologies to be tested and cannot provide details of how other teams play. Similarly, the role of individual players is beyond the remit of this thesis although it would seem plausible that in the applied world the extension of these methods would include individual player information. Hence, this thesis aims to help address the practical concerns of the validity of findings from large data sets by providing a robust methodology with scientific rigour that can be adapted by coaching staff and analysts to provide useful evidence of team tactics.

Chapter 3: Discriminating possession types in football to reflect influencing factors

3.1 Abstract

Possession in football has been classified in terms of the number of passes, duration or using descriptive labels such as build up and direct play. However, little consideration has been given to the fact that teams probably utilise all possession types at various times. This study considered factors likely to affect how a possession progresses. Possessions were classified according to the conditions at the initiation of possession (set piece or open play and which third of the pitch), possession type (build up, fast break, long ball, direct play or indeterminate) and where and how possession ended. All matches involving the first and second placed teams (Manchester City and Liverpool) in the 2018-19 English Premier League season were analysed. Manchester City had, on average, 10 less possessions per match than Liverpool but had more build up play, whereas Liverpool played more long balls and had more indeterminate possessions. Indeterminate possessions (25%) were excluded from the cluster analysis as these were possessions that did not allow the attacking intent (possession type) to be determined. Significant differences between Manchester City and Liverpool were found in a two-step cluster analysis for 6 of the 16 derived clusters suggesting that Manchester City tended to build up play from the back and had more possessions that started in the final third of the pitch compared to Liverpool. The cluster analysis was shown to discriminate the proportion of possessions types between teams. Future studies should consider that individual team possessions may include more than one attacking type and hence a methodology to cope with this needs to be employed.

3.2 Introduction

Football is an invasion game where both teams try to gain possession of the ball to facilitate the opportunity to score a goal. Since scoring more goals than the opponent results in a winning performance a lot of research has focussed on describing different attacking methods and their success in producing goal scoring opportunities. For example, attacking play has been classified as counterattack (Yiannakos & Armatas, 2006; Tenga *et al*., 2010a; Tenga *et al*., 2010b; Tenga *et al*., 2010c; Sarmento *et al*., 2010; Tenga & Sigmundstad, 2011; Lago-Ballesteros *et al*., 2012; Sarmento *et al*., 2014; González-Ródenas *et al*., 2016; Gómez *et al*., 2018; Zhao & Zhang, 2019; González-Ródenas *et al*., 2020), elaborate or possession play (Tenga *et al*., 2010a; Tenga *et al*., 2010b; Tenga *et al*., 2010c; Tenga & Sigmundstad, 2011, Lago-Ballesteros *et al*., 2012; Kempe *et al*., 2014), organised offense (Yiannakos & Armatas, 2006), combinative attack (González-Ródenas *et al*., 2015b; Mitrotasios *et al*., 2019) or positional attack (Sarmento *et al*., 2010; González-Ródenas *et al*., 2020). These studies have sometimes used different names for similar types of possession and provided differing levels of explanation for them. In conjunction with other possession types e.g. set piece, some researchers have described game styles based on the frequency of the different possession types exhibited by a team during a match or matches. For example, Fernandez-Navarro *et al*. (2018) used direct play, counterattack, maintenance and build-up possessions with operational definitions provided by Stats Perform (Stats LLC, Chicago, IL., USA). Similarly, Mitrotasios *et al*. (2019) used definitions for counterattack, combinative attack, fast attack and direct attack using the REOFUT observational framework developed by González-Ródenas *et al*. (2015a).

Possessions have also been classified in terms of the number of passes (Reep & Benjamin, 1968; Bate, 1988; Hughes & Frank, 2005; Redwood-Brown, 2008; Tenga *et al*., 2010a; Tenga *et al*., 2010b; Tenga *et al*., 2010c; Lago-Ballesteros *et al*.,

2012; González-Ródenas., 2016 % Garratt *et al*., 2017), duration (Jones *et al*., 2004; Bloomfield *et al*., 2005; Lago-Peñas, 2009; Lago-Peñas & Dellal, 2010; Fernandez-Navarro *et al*., 2016; Gómez *et al*., 2018; Mitrotasios *et al*., 2019) or both passes and duration (Tenga & Sigmundstad, 2011; Lago-Ballesteros *et al*., 2012; Kempe *et al*., 2014; Paixão *et al*., 2015) to analyse playing style. Limited research has explicitly considered possessions that potentially didn't contribute to playing style. Jones *et al*. (2004) suggested that possessions that lasted less than 3 seconds could be excluded from analyses because these possessions were not related to strategic attacking performance. For example, they were quick losses of possession before a meaningful attack could be undertaken or were specific moments in the game with strategies unique to the type of possession i.e. corners, penalties and free kicks. Tenga *et al*. (2009) excluded zero pass possessions and Aranda *et al*. (2019) removed very short duration attacks which were not able to be categorised by the observer. These papers considered that some team possessions, usually of short duration, did not portray any playing style that fit into the classification scheme being used in their study. This does not appear to have been considered by many researchers, and the non-removal of these possessions may have had a significant impact of research findings. For example, Tenga *et al*. (2010c) found that 884 out of 1688 random controlled possessions involved short possessions, defined as a team possession involving one or two passes. Similarly, Hughes and Franks (2005) found that zero and one pass possessions occurred more than the sum of all other possessions. These findings suggest that some thought should be given to this large proportion of possessions where attacking intent, from a strategic (decision-making) point of view, is not always evident. In this paper these possessions will be called indeterminate and not used to assess a team's attacking style of play.

The analysis of possessions in terms of passes, duration or descriptive classifications e.g. build up play, may be considered too simplistic if contextual variables are not included (McLean *et al*., 2017). For example, Kempe *et al*. (2014) suggested that analysing one or two isolated indicators was insufficient to investigate complex tactical performance whilst Rein and Memmert (2016) reiterated that the main limitation of traditional notational approaches was that almost all contextual variables were not considered. Hence, some previous studies have included indicators such as possession location (Lago-Peñas, 2009; Tenga *et al*., 2010a; Tenga *et al*., 2010b; Tenga *et al*., 2010c; Saramento *et al*., 2010; Tenga & Sigmundstad, 2011; Lago-Ballesteros *et al*., 2012; Saramento *et al*., 2014; Fernandez-Navarro *et al*., 2016; González-Ródenas *et al*., 2016), transition type (Kempe *et al*., 2014; Lago-Ballesteros *et al*., 2012; Saramento *et al*., 2010; Saramento *et al*. 2014; Fernandez-Navarro *et al*., 2018; González-Ródenas *et al*., 2020) and opponent defensive status (Tenga *et al*., 2010a; Tenga *et al*., 2010b and Lago-Ballesteros *et al*., 2012; González-Ródenas *et al*., 2016; González-Ródenas *et al*., 2020). The addition of these contextual variables required more complex analyses e.g. regression analysis (Tenga *et al*., 2010a; Tenga *et al*., 2010b; Tenga *et al*., 2010c; González-Ródenas *et al*., 2016; Fernandez-Navarro *et al*., 2018; González-Ródenas *et al*., 2020). This type of analysis has helped researchers provide more detailed information related to the attacking process but questions still remain, particularly regarding the simultaneous influence of multiple contextual variables. For example, how possession starts (set piece or open play), in which area of the pitch and how well organised the opponent's defence is, are all likely to affect how an attack progresses. However, other contextual variables such as the concurrent score-line (match status), team and opponent quality and the time period of the match may also be factors of potential influence.

Hewitt *et al*. (2016) presented a framework to analyse football performance which recognised five different moments of play (established offence/defence, transition to offence/defence and set pieces). This methodology recognised the importance of splitting the match into disparate groupings based on the fact that defences in these situations are usually in very different formations, which in turn, affects the attacking team's modus operandi i.e. game style. Effectively, this approach recognised that these different moments should be considered a contextual variable and hence should be delineated if game style is to be properly analysed. This concurs with the view of Jones *et al*. (2004) who suggested that possessions that lasted less than 3 seconds (potentially set piece and transitions) should be treated differently to longer possessions (established play). Whilst the methodologies between these two papers were very different, the recognition that all possessions should not be treated the same, is an important distinction that has been neglected in too many research papers.

In summary, performance in football is complex due to the relatively large number of players directly involved with the play at any one time and the different methods of attack and defence work in tandem. McLean *et al*. (2017) found that performance analysis research typically failed to account for this complexity, often presenting variables that lacked context e.g. no pitch area information, and the information provided was "interesting but not useable" (p. 8). In this study, exemplar data from Liverpool and Manchester City during the 2018-19 EPL was analysed. Indeterminate possessions were defined and results were presented that both included and excluded these possessions for comparison A two-step cluster analysis was utilised to detect patterns, which were then analysed for the contextual influence of multiple combinations of contextual variables for each possession for the two teams.

The quantity of different possession types exhibited over a full season were compared for the two teams to demonstrate between team differences and similarities.

3.3 Methods

3.3.1 Sample

All matches involving the first and second placed teams (Manchester City and Liverpool) in the 2018-19 English Premier League (EPL) season were downloaded from commercial broadcasts. The teams were selected as both team managers had remained unchanged for over two years (Liverpool since October 2015 and Manchester City since February 2016) suggesting that playing styles had remained relatively unchanged at each club. Both teams made relatively small changes to their playing squads from the previous season. Liverpool had signed 3 new outfield players for the 2018-19 season, they started 21, 16 and 11 of the 38 matches. They also signed a new goalkeeper who played every match. Manchester City signed one new player, who started 14 matches.

3.3.2 Operational definitions

This study describes each team possession by recording where on the pitch the possession started and ended (defensive, middle or attacking third), how possession started (set piece or open play) and a descriptor of the possession type (n=5, Table 3.1).

Table 3.1. Operational definitions for possession classifiers

Initiation of possession	Open play	Possession regained directly from the opponent's possession.
	Set piece	Possession starts from the ball out of play.
Possession type	Build-up	At least 4 non-attacking passes preceded an attempt to play the ball into the opponent's defensive area.

Fast break	The opponents had possession of the ball with an unestablished defensive structure, a turnover of possession then involved a quick (within 2 seconds or 3 passes) attempt to play the ball (dribbling or passing the ball to a teammate) into the opponent's defensive area. This is often referred to as a counter attack.
Direct play	The opponents had possession of the ball with an established defensive structure, a turnover of possession then involved a quick (within 2 seconds or 3 passes) attempt to play the ball (dribbling or passing the ball to a teammate) into the opponent's defensive area. This possession type involves a fast tempo but does not include a long pass from defence to attack.
Long ball	This is a direct play attack but using a long pass to initiate the attack. Hence, in possession of the ball, the opponents had an established defensive structure, a direct long pass into the attacking third of the pitch occurred within 2 seconds or 3 passes from the initiation of possession.
Indeterminate	Short ball possessions where little control was exerted on the ball or no obvious attacking process was present. The possession could not be classified as any other possession type.

3.3.3 Procedure

Each match was viewed full screen and coded in Dartfish Pro S v10.0 (build 20812) with a specifically designed template. This enabled each possession to have codes that registered start and finish times, the pitch area in which possession was gained and lost (defensive, middle or attacking 3^{rds}), initiation of possession (open play or set piece) and possession type (build up, fast break, direct play, long ball and indeterminate).

3.3.4 Statistical Analysis

The data were exported into Microsoft Excel (v2010, Microsoft, Redmond, USA) and IBM SPSS Statistics (v25, IBM Corp, Armonk, NY, USA) for analysis. Assessments of the frequency of different types of possession per match (total, build up, fast break, direct play, long ball and indeterminate) revealed some skewed distributions and the presence of outliers. Non-parametric tests (Mann-Whitney) were thus used to assess between team differences and appropriate descriptive statistics (median and interquartile range) and effect sizes were calculated (r = standardised test statistic z divided by the square root of the sample size). Cluster analysis is a data mining technique that enables the formation of groups within a data set based on maximising the homogeneity of cases within a group and the heterogeneity between clusters (Hair, Anderson, Tatham, & Black, 1995). Cluster analysis begins with all cases as separate groups and the two ‘‘most alike’’ cases are combined in the first step using the most appropriate distance measure (Log-likelihood). If all variables are continuous, the Euclidian distance measure is used, but with categorical variables and a continuous variable (in this study) the log-likelihood is more suitable since categorical variables are translated into dummy variables when the Euclidian distance measure is applied. The algorithm for two-step cluster analysis using log-likelihood has also been shown to be more accurate than the Euclidian distance measure (Chiu *et al*., 2001). This procedure results in the two cases with the smallest distance measure clustering together and a group mean (cluster centroid) calculated and used in the next step. The next two most alike cases (or groups once cases have been clustered) are then combined. This process continues until an optimal cluster solution is obtained, although this may be determined from a practical standpoint as there are no objective methods for determining the optimal number of clusters (Hair, *et al*., 1995). Hence, a two-step cluster analysis using a probability-based log-likelihood distance measure

(SPSS) enabled both continuous (duration of possession) and categorical (start pitch area, initiation of possession, possession type and finish pitch area) variables to be assessed. To compare Manchester City and Liverpool's use of each possession cluster a Pearson chi square test was used with post hoc multiple comparisons undertaken using the adjusted standardised residuals and a Bonferroni adjustment to the p value to account for the number of comparisons (Beasley, 1995). The effect size was calculated as Cramer's V which equals the square root of the chi-squared statistic divided by (the sample size multiplied by the number of teams minus 1).

3.3.5 Reliability

Intra- and inter-operator reliability tests on two randomly selected matches determined coding accuracy. Matches were re-coded by the experimenter (intra-test) 4 weeks after the initial coding to minimise memory effects. An experienced football coach and analyst was trained on the operational definitions before coding (inter-test). The original analysis suggested there were 221 possessions but the inter- (n=218, 95% agreement) and intra-observer tests (n=223, 96% agreement) found differently. This was a consequence of how a possession was defined since a small touch on the ball can be considered a possession in some studies but in this study a touch had to have a consequential effect on the ball for the touch to be considered a possession. These anomalies affected possessions classified as indeterminate in this study. Further inter- (n=152) and intra- observer reliability tests (n=163) on non-indeterminate possessions (Appendix 3) found high kappa values for starting and finishing pitch areas (inter- 0.89 and 0.95; intra-observer tests 0.94 and 0.92). Similarly, initiation of possession had high kappa values (inter = 0.99 and intra = 0.90) whereas possession type indicated slightly lower rates of reliability (inter = 0.87 and intra = 0.87).

3.4 Results

3.4.1 Team possessions

Manchester City had less (U = 381.50, z = -3.54, p < 0.001; r = 0.40) possessions per match (Median = 117, IQR = 13) than Liverpool (Median = 127.50, IQR = 14; Figure 3.1).

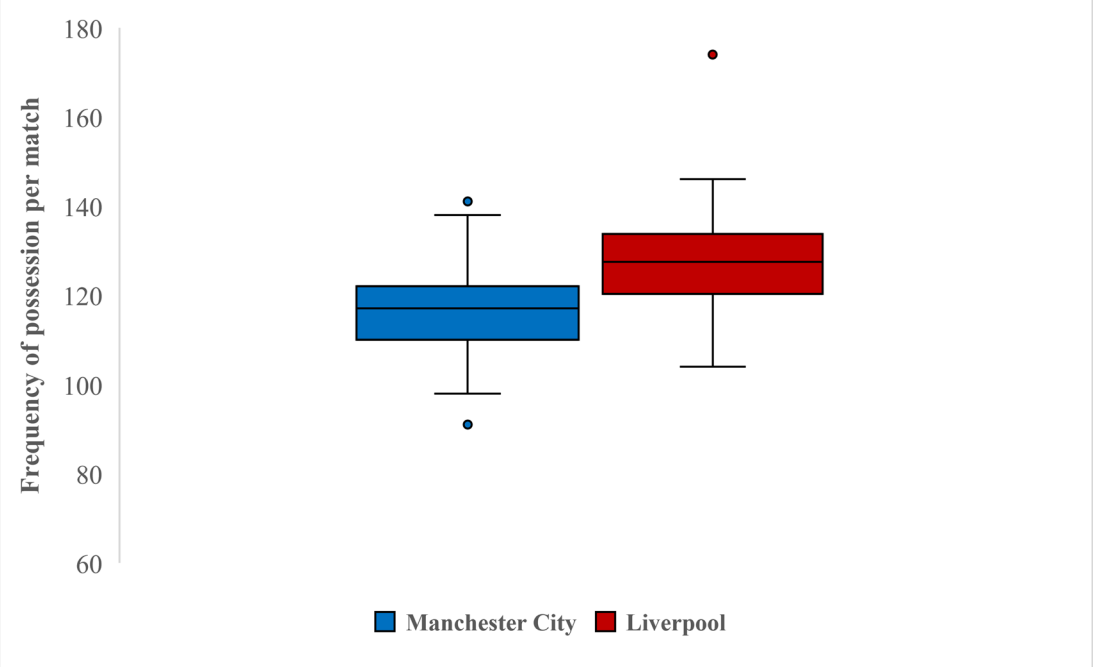


Figure 3.1. Frequency of possessions per match by Manchester City and Liverpool in the 2018-19 EPL season

This was due to Liverpool having more long ball possessions per match (Median = 14, IQR = 10; Figure 3.2) than Manchester City (Median = 8, IQR = 5; r = 0.58) and more indeterminate possessions per match (Median = 33, IQR = 10) than Manchester City (Median = 27.5, IQR = 9; r = 0.42). Conversely, Manchester City had more build up possessions (Median = 40, IQR = 9) than Liverpool (Median = 35.5, IQR = 13; r = 0.25) and both teams had a similar number of direct play (Manchester City: Median = 22.5; IQR = 10 and Liverpool: Median = 23.5; IQR = 12; r = 0.18) and fast breaks (Manchester City: Median = 18; IQR = 6 and Liverpool: Median = 19; IQR = 8; r = 0.14).

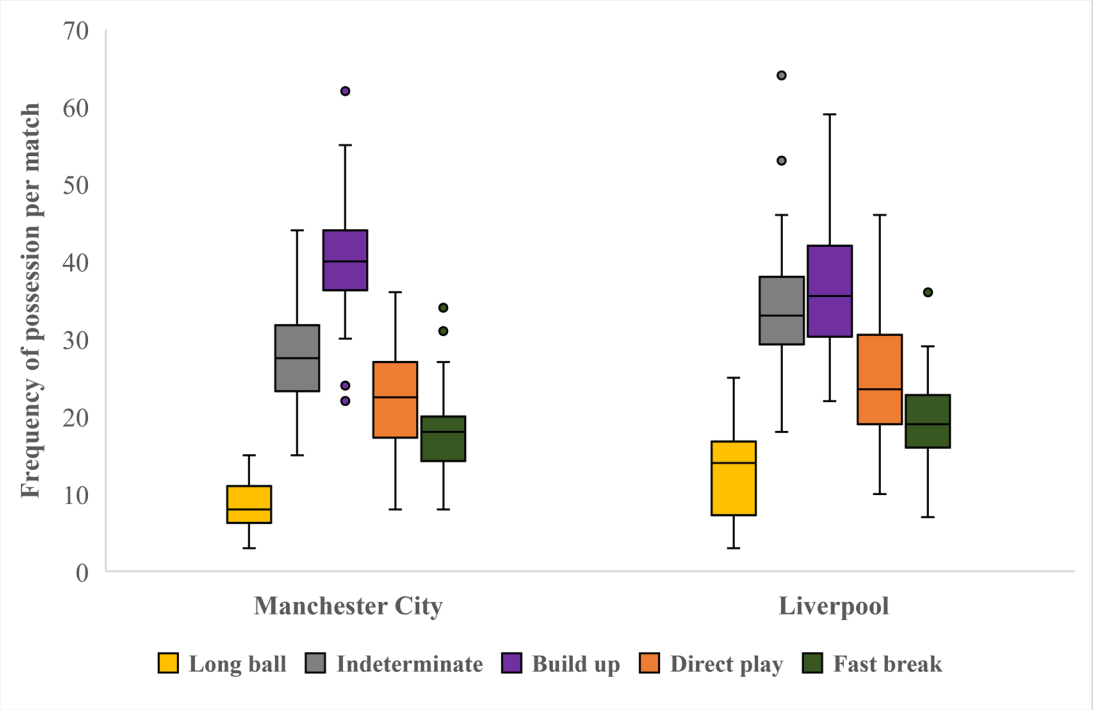


Figure 3.2. Frequency of possession types per match by Manchester City and Liverpool in the 2018-19 EPL season

Manchester City's direct play possessions tended (U = 327, z = -4.1, p < 0.01; r = 0.32) to last longer (Median = 14.12s; IQR = 3.78s; Figure 3.3) than Liverpool's (Median = 11.38s; IQR = 3.55s) as did their build up play (Median = 33.22s; IQR = 4.72s) compared to Liverpool's (Median = 31.41s; IQR = 6.74s; r = 0.32). In contrast, Liverpool's fast breaks (Median = 11.03s; IQR = 4.49s) were shorter than Manchester City's (Median = 14.83s; IQR = 4.81s; ES = 0.44) but the long ball (Manchester City : Median = 10.12s; IQR = 5.35s and Liverpool : Median = 9.58s; IQR = 2.86s; r = 0.16) and indeterminate plays (Manchester City: Median = 4.53s; IQR = 1;32s and Liverpool: Median = 4.1s; IQR = 0.78s; r = 0.21) were similar for both teams.

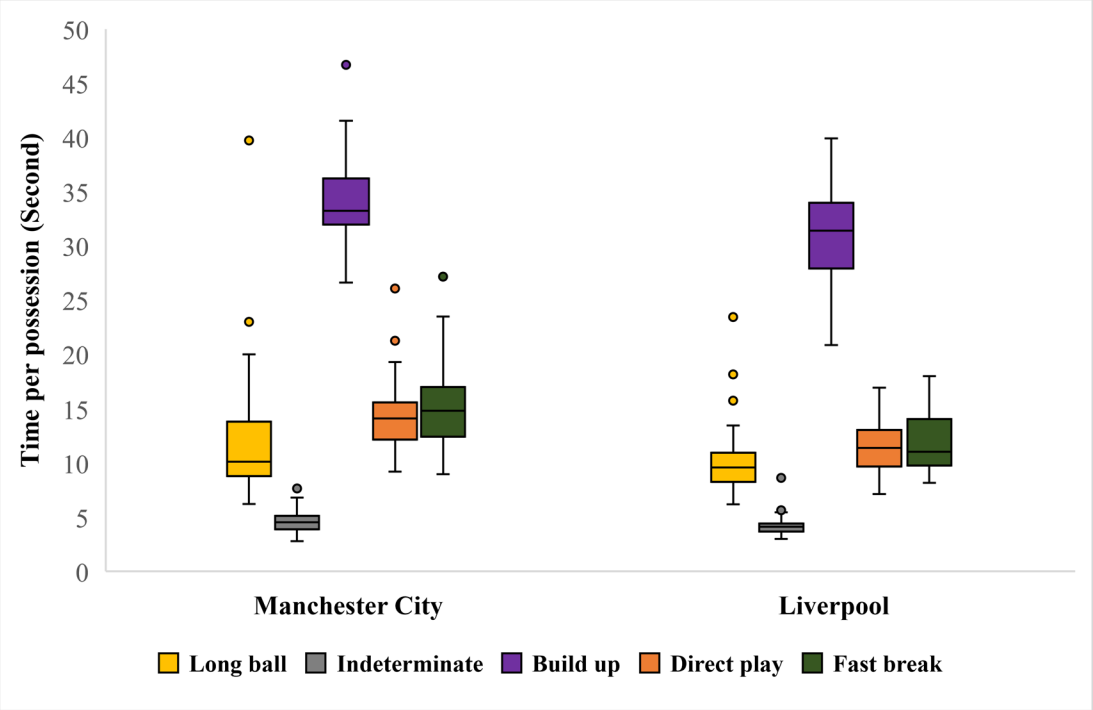


Figure 3.3. Duration of different possession types for Manchester City and Liverpool in the 2018-19 EPL season

3.4.2 Two-step cluster analysis

The indeterminate possessions (25.43%) were excluded from a two-step cluster analysis since these possessions could not be classified in terms of an attacking process. Sixteen clusters (plus an outlier cluster) were then formed from the classified possessions as represented by three categorical variables, the pitch area where possession initiated, the initiation type and possession classification as well as one continuous variable, the duration of the possession (average silhouette = 0.72 (good quality), valid cases = 6931; Figure 3.4).

The area of the pitch where the possession ended was not used in the cluster analysis as this variable was found to adversely affect the clustering algorithm resulting in clusters that were poor. However, the proportion of cases in each cluster that finished in the attacking third was presented for comparison purposes as this could be deemed as a measure of the success of a possession.

Fourteen of the clusters consisted of unique levels of the three categorical classification variables with two clusters (labelled 15 and 16, Figure 3.4) combining some cases involving direct play (n=182) and long balls (n=70) with fast breaks (n= 17) and cluster 16 combining long balls (n=56) with fast breaks (n=52). An outlier cluster had a mixture of all levels of all categorical variables and accounted for 2.23% (n= 155) of cases but was not represented in Figure 3.4.

Manchester City and Liverpool were compared for the frequency of their possessions in each cluster. The pitch area where possession initiated for each cluster is represented by a circle (Figure 3.4) but since possessions didn't always progress in a forward manner these circles didn't always represent the most defensive area a possession took place in. Lines for each cluster present the proportion of possessions that ended in the attacking third of the pitch (thicker line and percentage given) as well as the average duration of the cluster. Finally, since the different clusters represented different types of attack, the proportion of possession outcomes that resulted in shots or goals was also presented as a percentage (Figure 3.4).

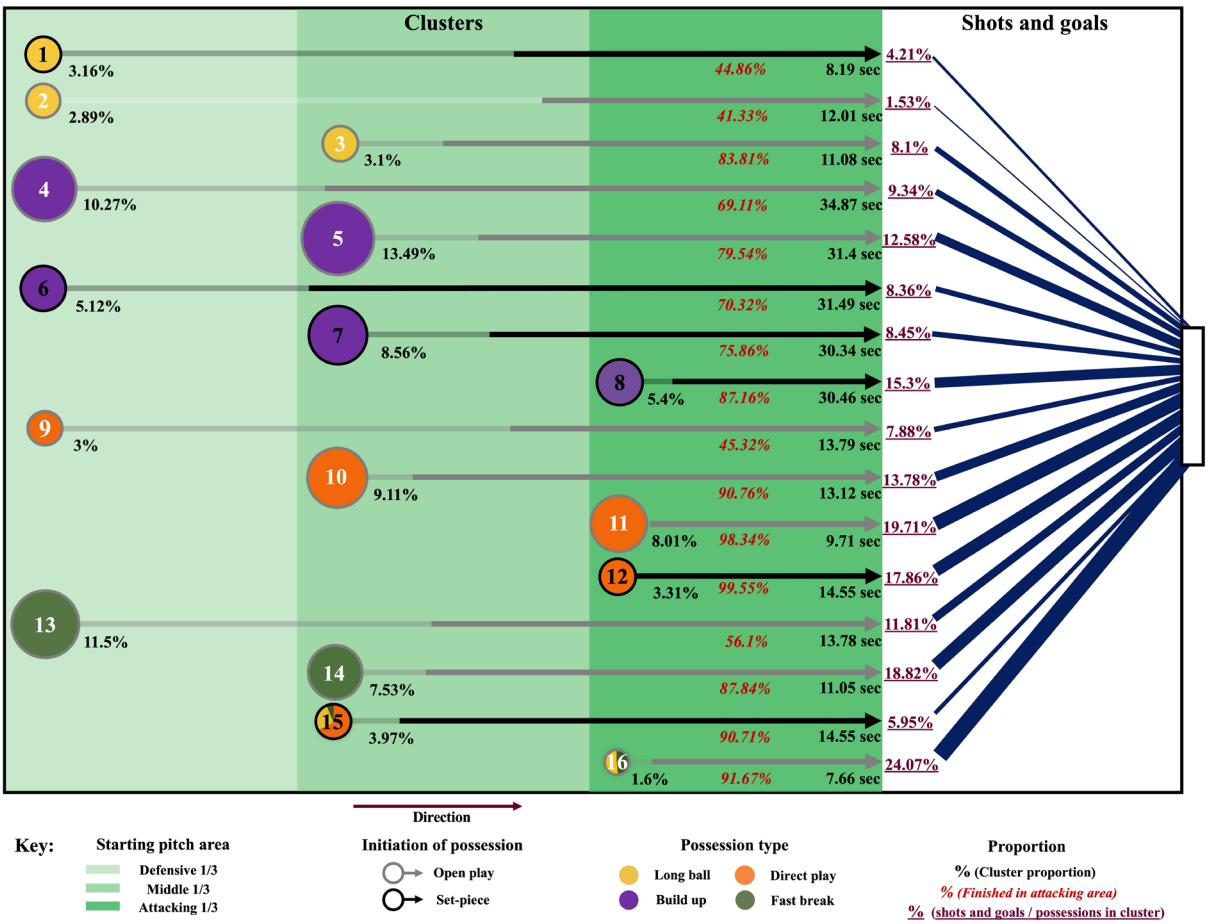


Figure 3.4. Sixteen possession types when indeterminate possessions were excluded

The proportion of possessions in each of the 16 clusters for Manchester City and Liverpool differed (Chi square = 126.04, df = 15, $p < 0.001$; V = 0.14, Figure 3.5). Liverpool played more long balls in open play possessions (cluster 2, $p < 0.001$ and cluster 3, $p < 0.01$) starting in the defensive third (4.17%, cluster 2) and middle third (3.85%, cluster 3) compared to Manchester City (1.55%, cluster 2 and 2.30%, cluster 3) and more long balls from set pieces ($p < 0.05$) starting in the defensive third (cluster 1, 3.82%) compared to Manchester City (2.46%). In contrast Manchester City utilised more ($p < 0.001$) build up play in open play originating in middle (cluster 5, 15.28%) thirds than Liverpool (11.79%). No significant differences were found between the teams for direct plays and fast breaks. Significant differences were found for one mixed cluster (open play possessions that started in the attacking third and involved

either long balls or fast breaks; cluster 16) where Manchester city had more possessions (2.33%) than Liverpool (0.89%).

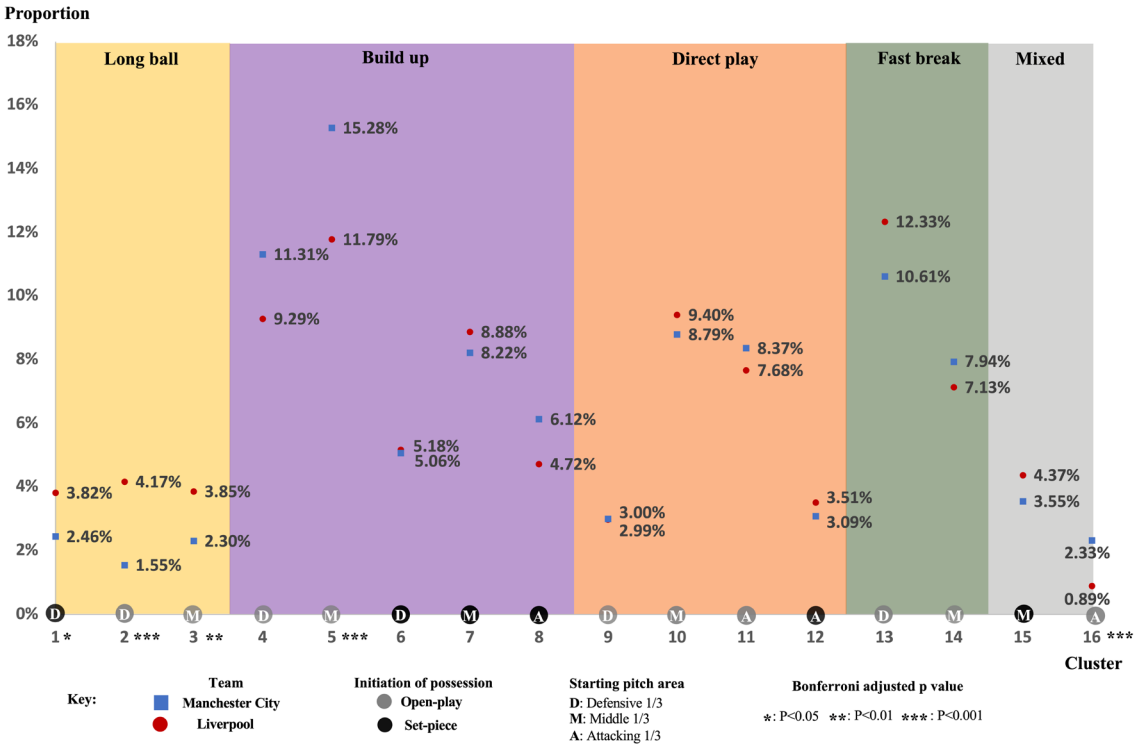


Figure 3.5. The proportion of possessions in each of the 16 clusters for Manchester City and Liverpool

The outliers were not presented in Figure 3.5 because these possessions could have belonged in the other clusters but were excluded for one of two reasons. The most prevalent reason being the time duration of the possession being extreme (very long) and unusual. Less prevalent was that there were some very small clusters that were rarely evident. For example, set piece play possessions that involved direct plays and started in the defensive third were observed 14 times in all matches. Similarly, set pieces that started and ended in the attacking third and involved a long ball were observed 39 times.

3.5 Discussion

This study has presented the case for removing indeterminate possessions from analysis procedures as, in this case, they could not be classified in terms of an attacking

process. This is similar to Jones *et al*. (2004), although here, unlike Jones *et al*., the removal of these possessions was when a possession failed to match any of the operational definitions. This is, in our opinion, a better methodology than Jones *et al*. where they simply removed possessions that lasted less than 3 seconds. Hence, this paper included set piece possessions, unlike Jones *et al*., which better reflects strategic attacking performance in its entirety. The indeterminate possessions occurred about 30 times per match out of a team's 120 odd possessions. This is less than the approximate 50% of possessions that involved very few passes (e.g. Tenga *et al*., 2010c; Hughes & Franks, 2005) because some possessions containing only a low number of passes could easily be classified in terms of attacking intent e.g. direct play. However, 25.43% of possessions is still a significant proportion and if these possessions were unreliably categorised into possession types their influence on study findings could be dramatic. For this reason, these possessions were excluded from the analysis of possession types and the suggestion to other researchers is to be alert to the possibility of falsely categorising events and the potential impact that this would have on results.

The second outcome of this research was to categorise possessions into attacking types using previously used typologies; long ball, build up, direct and fast break, but also grouping according to how (set piece or open play) and where (pitch thirds) the possession started. A two-step clustering algorithm initially also included where the possession ended (pitch thirds) but this did not result in clear divergence in the data and hence the outcome of possessions (pitch third and whether a shot took place) was used descriptively. The subsequent clustering presented the case for 16 different attacking possession types with an outlier cluster containing all of the very unusual possessions. Since about 25.43% of possessions had been removed from the

analysis because they had been deemed indeterminate, this research suggests that about 71% of all possessions (2.23% were outliers) could be classified as one of 16 possession types.

Possessions that started in the attacking third were more likely to end in the attacking third (mean = 94.18%) than possessions that started in the middle third (mean = 84.75%) or defensive third (mean = 54.51%). This is logical and presents a form of the success rate for team possessions in football. A more often used statistic for this is whether possessions resulted in a shot or goal and again the data revealed the logically plausible result that possessions that started in the attacking third were more likely to result in a shot or goal (mean = 19.23%) than possessions that started in the middle third (mean = 11.28%) or defensive third (mean = 7.19%). This concurs with González-Ródenas, *et al*. (2016) who found that the percentage of counter attacks leading to goal scoring opportunities decreased the further away from goal the possession started. Other studies e.g. Tenga, *et al*. (2010c) presented the overall proportion of goals scored from the different positions where possessions started on the pitch. These results suggested that the middle and defensive thirds produced more goals than possessions starting in the attacking third but these were not proportionate success rates i.e. the likelihood of a possession resulting in a shot, but were reflective of the fact that very few possessions started in the attacking third of the pitch and hence few shots resulted. It should be noted, however, that the percentages presented here excluded infrequently occurring possessions (deemed outliers), such as quickly taken set pieces in the attacking third which are likely to present the team with a high goal scoring potential. Other outliers, for example possessions that ended in the defensive third, clearly would not have resulted in goal scoring opportunities, at least not for the team that lost the ball! Given that football is a low scoring game and hence, a single

team possession could be decisive in terms of the outcome, it is recommended that future studies should examine these outliers in more detail as they may contain important information that ultimately changes the outcome of matches.

This study selected whole season data for two teams, Liverpool and Manchester City, based on the fact that they were the top two teams in the English Premier League at the time, their managers had been in position for a relatively long period and they used fairly similar tactics (high possession rates and excellent close control of the ball). This provided a good test of the efficacy of the clustering algorithm for discerning individual team differences but also the analysis provided evidence for the similarity of the teams with 11 of the 16 clusters not being significantly different between the two teams. These results showed that Liverpool played more long balls per match than Manchester City, particularly in open play possessions starting in the defensive and middle thirds of the pitch (clusters 2 and 3) and from set pieces starting in the defensive third (cluster 1). This is not too surprising given that Manchester City are well known for playing the ball out from the back i.e. not playing long balls from their own half of the pitch. Indeed, Manchester City, had more build up plays than Liverpool, particularly in open play possessions starting in the middle third (cluster 5). Both teams had a similar number of directs play and fast breaks which reflects their ability to regain possession and attack quickly. These findings show that small differences in performance can be determined but the findings here cannot be compared to other teams as the data for these comparisons was not available. However, the fact that the results are broadly what football pundits would expect suggests that the findings are valid and if future analyses used other teams then meaningful between team comparisons could be made e.g. to determine whether the proportion of fast breaks made by these two teams is league leading, average or low for the league. This

paper has been novel in its approach to classifying possessions at this detail although it may well be argued that more detail is required e.g. coaches would be interested in which players made key passes and whether dummy runs created the space for these passes. Hirano and Tsumoto (2005) presented a complex comparison of pass patterns for teams in the 2002 FIFA World Cup using a multi-scale structural matching technique in addition to rough clustering to eliminate dissimilarity. This demonstrated the complexity of football, in particular how no one possession was exactly the same as another, but also highlights the need, and the limitation of, clustering techniques.

Significant differences were found for the mixed cluster 16. Manchester City had more possessions in this cluster (2.33%) than Liverpool (0.89%) which were open play possessions that started in the attacking third and involved either long balls or fast breaks. This is probably a reflection of Manchester City's greater ability to regain possession in the final third and have more possessions that started in the final third of the pitch relative to Liverpool. Both teams are well known for tending to have a lot of possession in comparison to their opponents, but this study has highlighted where their strategies have diverged.

This study utilised the concept that football can be considered to contain five distinct moments (Hewitt *et al*., 2016) although the analyses were limited to the three attacking moments (see also Lago-Peñas *et al*., 2017b; Gómez *et al*, 2018; Gollan *et al*., 2018; Aranda *et al*., 2019). Hence, information related to where possession started and finished and how quickly the ball moved forwards was analysed to better understand game style. However, this approach has limitations in that one possession may not have only one defining pattern of play. For example, if a team intercepts the ball from their opponent's attacking moment and quickly transitions forwards, the possession would be categorised as a fast break. However, the opponents may be able

to reorganise themselves into a good defensive shape and prevent the fast break from being successful. This may result in the possession taking a second, different, form, namely build up play. This is particularly evident for teams like Manchester City and Liverpool who “recycle” the ball when an attack doesn’t succeed in creating a shot opportunity. In this situation the teams are happy to pass the ball backwards to restart the attacking process, albeit in one team possession. This means that one possession can involve more than one playing style. Furthermore, this study only considered how the ball was played forwards but there was no consideration of how the team tried to get the ball past the opponent’s defensive line. This will, therefore, be addressed in the next study where a methodology for identifying the specific tactics for penetrating the opponent’s defensive line will be investigated to better reflect how teams attack in open play moments.

3.6 Conclusion

This analysis of individual team possessions has generated a number of issues for researchers to contemplate. The recognition of indeterminate possessions has implications for any classification scheme of team possessions, to minimise errors and hence improve understanding. Sixteen different attacking possession types were shown to best describe the attacking methods for two of the best teams in the English Premier League but about 30% of possessions were not classified, predominately because an attacking intent was not evident. Other teams may have different attacking possession types and the proportion of indeterminate possessions unknown. No account was taken of an opponent’s defensive structure and this may impact the proportionate occurrence of the different attacking possession types. Finally, future analyses should consider that individual team possessions may include more than one attacking type and hence a methodology to cope with this needs to be employed.

Chapter 4: Separating a single team possession in football into phases to allow for several different attacking strategies

4.1 Abstract

Football teams are often described according to the way they tend to play e.g. build up play or counter attacks in possession and high press or low block in defence. These descriptions ignore facets of play that occur less frequently but nevertheless may help describe a team's playing style. Also, not all possessions will necessarily aid the description of playing style e.g. low duration ones and set pieces, and some may be more complex than to be categorised into one style. All matches involving the first and second placed teams in the 2018-19 English Premier League season were analysed. All possessions, except those classified as indeterminate i.e. short duration team possessions with little control on the ball or no obvious attacking process; were analysed. Possession was initially coded as long ball, build up, direct play or fast break. A secondary classifier was used when a team tried to break the opponent's defensive line e.g. using a dribble. If this was successful, but the team subsequently passed the ball back behind the opponent's defensive line, this was coded as a new phase of the same possession. Data mining was used to produce Association rules for the sequential possession data and decision trees then used to determine which sequential patterns produced the most shots at goal. Manchester City had 34 different attacking patterns (Liverpool 33) after patterns with rule support of less than 1% were removed. Possessional phases increased Manchester City's total attacking actions by about 16% (Liverpool 9%) suggesting that possessions were discriminated at a more detailed level than previously. Manchester City's most used possession type was build-up play culminating in a pass to the wide area to try to invade the opponent's defensive line. Liverpool utilised both fast transitions from defence to attack and build-up play with

the significantly greater use of long passes from behind than Manchester City. This approach was able to discriminate two teams of similarly high standard, and of similar playing style but future analyses should also consider the range of attacking strategies used in different specific situations based on the quality of the opponent, match status and venue.

4.2 Introduction

Possession in football has usually been analysed to better understand game style and overall team performance. Early studies tended to consider team possessions in relation to a few isolated indicators e.g. shots, goals or pitch area, where possessions were classified simply according to the number of passes undertaken (Reep & Benjamin, 1968; Bate, 1988; Hughes & Franks, 2005, Redwood-Brown, 2008, Ridgewell, 2011, Paixão *et al*., 2015) or duration (James *et al*., 2002; Jones *et al*., 2004, Paixão *et al*., 2015). However, these simplistic approaches could not fully describe performance (Hewitt, 2016) particularly since football is complex, multifaceted, and unpredictable (Mackenzie & Cushion, 2013).

More recently, possession has been considered with respect to contextual variables such as match status, venue and opposition quality (Gómez *et al*., 2012; Bradley *et al*., 2014; Moura *et al*., 2014; Clemente, 2015a; Liu *et al*., 2015; Pratas *et al*., 2016; Casal *et al*., 2017; Lago-Peñas *et al*., 2017a; Bilek and Ulas, 2019; Yi *et al*, 2019; Lepschy *et al*., 2020; Yi *et al*, 2020). Whilst these, more in-depth, analyses provided contextualised findings, they still tended to generalise the results over multiple, whole matches and teams, such that potentially interesting patterns of play remained hidden. In this regard, Hewitt *et al*. (2016) stated the importance of analysing the different moments of a football match separately. These moments were suggested to be transitions from defence to attack, established attacks, transitions from attack to defence, established defences, and set pieces. Essentially, this philosophy recognises that football teams operate in fundamentally different ways for three mutually exclusive events. Established play (one team's established attack, the other's established defence) involves relatively slow possessions where both teams have time to organise themselves generally in relation to their preconceived playing formation e.g. 4-4-2. Transitions are relatively quick possessions where a turnover of the ball

leads to an immediate attack. These are again assessed from both team's perspectives (attacking team and defending team) but due to the speed of the attack both teams are unable to organise their players in a recognised formation but rather try to take advantage, or mitigate against a weakness, in a relatively disorganised manner. Finally, set pieces involve both attacking and defending teams organising their players in very specific formations dependent on the type of set piece e.g. corners and free kicks near the opponent's penalty box. Recognising these very different moments of play, and analysing them separately, is very likely to facilitate the characterisation of game style in football (Hewitt *et al*., 2016) as erroneously summating the different moments of the game would clearly detract from describing the complex tactical nature of football.

Tenga and colleagues (2010a, 2010b) analysed team performance in the Norwegian men's professional league in the 2004 season, using multiple classifiers for each possession (starting zone, possession type, number of passes, pass penetration and team possession outcome, all related to opponent defensive status). These comprehensive analyses, albeit on Norwegian professional football, presented the odds ratio of scoring or producing a score box possession using one method compared to another e.g. an elaborate attack compared to a counter attack. Their findings tended to agree with Olsen and Larsen (1997) that counter attacks were more effective than elaborate attacks although they did also add an extra dimension to their analyses concluding that this result only held true against an imbalanced opponent defence. Whilst this is interesting, it should also be noted that counter attacks tend to occur against an imbalanced defence whereas this is not the case for elaborate attacks. Furthermore, counter attacks tend to start closer to the opponent's goal and therefore are more likely to result in a shot or goal (Gonzalez-Rodenas et al., 2016; Sarmento et al, 2018). Similarly, Lago *et al.* (2012) analysed 908 team possessions from 12

matches in La Liga 2009-10 seasons with multiple variables related to possession (starting zone, team possession type, duration, number of passes, players in possession, passing options and possession outcome related to opponent status). They found that 33.4% of possessions produced scorebox entries with direct and counter attacks more effective than elaborate attacks in this regard. Similar to Tenga *et al*. (2010b), they found that possessions starting in the pre-defensive and pre-offensive areas were more effective at producing scorebox entries than starting in the defensive area and attacking against a balanced defence was less effective than against an imbalanced one.

Gonzalez-Rodenas *et al*. (2015a) analysed all 857 possessions of the seven matches played by Spain in the 2010 World Cup. They used type of start-up possession, field starting zone, initial penetration, initial penetration zone, initial opponent number, initial opponent position, type of progression, pass number, duration and percentage of penetrative passes to find associations with creating scoring opportunities. Set plays, from advantageous situations i.e. near the opponent's penalty box, were the most effective possessions for creating goal scoring opportunities, with no difference between coroner kicks and free kicks. Initial conditions when the team recovered the ball were associated with the propensity to produce scoring opportunities i.e. position on pitch, penetration mode and the number of effective opponents and their positions. The area where the team penetrated the defence was also a significant factor.

Gonzalez-Rodenas *et al*. (2015b) analysed 3523 team possessions during selected matches (n=30) from the 2014 U.S. Major League Soccer season. They used slightly different variables to their 2015a study with initial invasive and defensive zones, initial defensive pressure, type of attack, passes per possession, percentage of penetrative passes and possession outcomes assessed. They also assessed the impact of contextual variables (match status, match half and match location) on creating

scoring opportunities. Similar to previous studies (Gonzalez-Rodenas *et al*. 2015a; Tenga *et al*. 2010a, 2010b; Lago, 2012) contextual and defensive variables were important factors for the creation of scoring opportunities. Set pieces were more effective than recoveries and restarts. Counter attacks were more effective than combinative attacks and direct attacks when the possession started in the invasive zone with no initial defensive pressure. Long possessions were more effective than short possessions for recoveries and restarts.

Gonzalez-Rodenas *et al.* (2016) analysed the counterattack possessions (n=452) in the selected matches (n=30) from the 2014 U.S. Major League Soccer season. Five offensive variables (ball recovery type, field starting zone, initial penetration, passes per possession and percentage of penetrative passes) and four defensive variables (initial invasive zone, initial opponent number, initial opponent position and initial defensive pressure) were used. Counter attacks that started in the pre-offensive zone and had initial penetration i.e. passes or dribbles at the opponent's defensive line during the first three seconds of ball possession, created more scoring chances than possessions started in the defensive zone without initial penetration. Perhaps counter intuitively, four or more pass possessions in counter attack were more effective than shorter possessions, regardless of the initial defensive pressure. Whilst speed tends to be the defining feature of the counter attack, hence shorter possessions may be considered advantageous, longer passing sequences also means that possession was retained, a positive factor, particularly since this may be in attacking areas given that the possession was a counter attack.

The studies described above have described the features used to classify possessions, with many utilising the initial conditions e.g. pitch area where the possession originated. However, no paper has discussed whether a team possession

has subsequently involved more than one level of any of the factors investigated. For example, a counter attack that starts with no initial defensive pressure could easily change to a build-up play situation with defensive pressure. If this was the case, did the researchers solely classify the attack as for the initial situation, and hence their possession outcomes would reflect this decision. Or did they also categorise the second situation and their possession outcomes would be different. This is not clear in the extant literature, but logically, if the effectiveness of a particular attacking strategy is to be derived, the second alternative would be more accurate.

Whilst many variables have been used to describe the attacking process, the method used to pass the ball through the opponent's defensive line has not been extensively considered. Pulling *et al*. (2014) analysed 1332 crosses during open play situations during the FIFA 2014 World Cup, with player actions, pitch area, delivery type, defensive press and cross outcomes recorded. These crosses resulted in 42 goals (3.2% success rate) with the pitch area the cross ended in and the defensive pressure significantly related to creating a scoring chance, but the side of the pitch and match time were not deemed relevant. The penultimate actions of possession (assists) have been analysed (Aranda *et al*., 2019; Mitrotasios *et al*., 2019; Gonzalez-Rodenas *et al*., 2020) but these actions e.g. a cross, pass or individual action, did not necessarily explain an invasion attempt, rather they described the last action before the scoring opportunity. Kim *et al*. (2019) described the attacking process in terms of situations of varying pressure on the opponents with three levels (stable – no pressure; advantage – imminent threat of goal scoring opportunity; unstable – goal scoring opportunity). Whilst this study did not consider the method of progress between these situations it did present the case that goal threat is related to pitch area, situation e.g. number of defenders, and action e.g. success of the cross.

Therefore, the aims of this study were; 1) to assess team possessions from the perspective that they may involve more than one level of the classification variables used; and 2) to describe the method used to pass the ball through the opponent's defensive line to see whether this affected performance outcomes. This methodology is expected to provide more detailed information regarding the attacking process and may help coaches and performance analysts to evaluate opponent invasion patterns and inform training procedures prior to a match.

4.3 Methods

4.3.1 Sample

The possessions for both teams in a match for all 2018-19 English Premier League matches for Manchester City (n=36) and Liverpool (n=36) were analysed. However, indeterminate possessions, defined as short ball possessions where there was little control of the ball or no obvious attacking process was evident (Study 1) were excluded.

4.3.2 Procedure

All matches were viewed and coded in Dartfish Pro S v10.0 (build 20812). Each team possession was coded for initiation of possession, attempts to break the opponent's defensive line (invasive action) and possession outcome (Figure 4.1). This new method of data entry recorded the invasive action and, if this was successful, the potential for the team to pass the ball back behind the defensive line, defined as a new phase of possession. New phases of possession could occur multiple times within a single team possession.

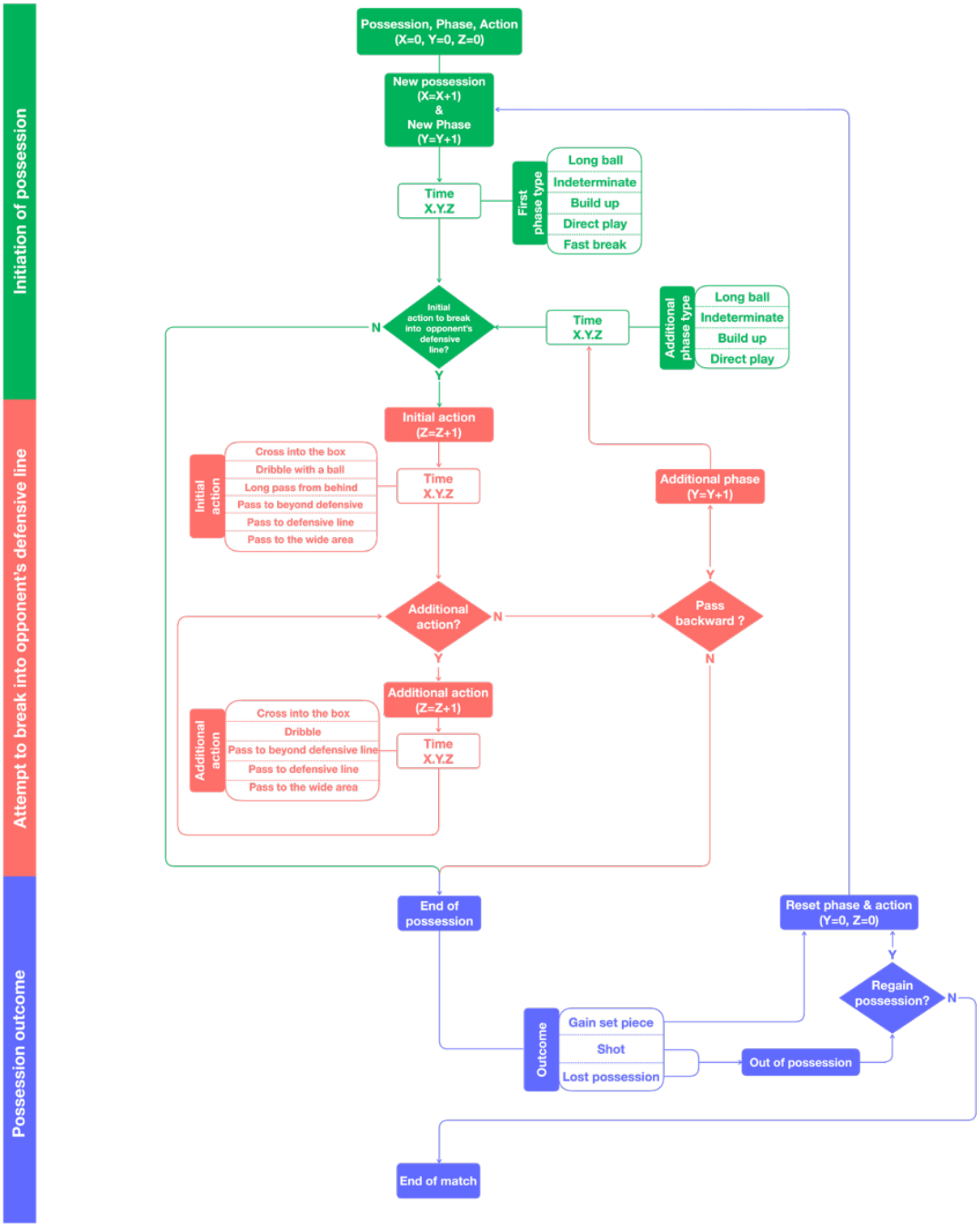
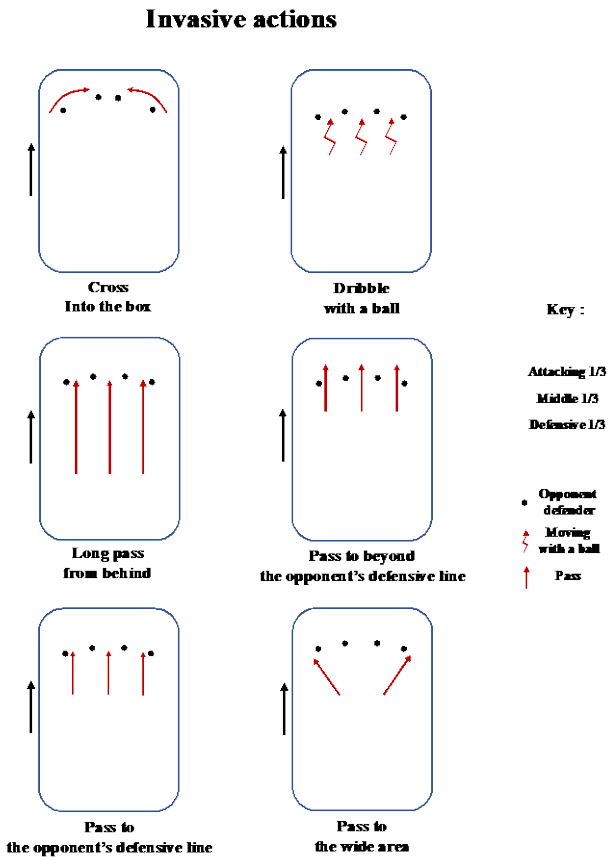


Figure 4.1. Flowchart of data entry for team possessions to allow multiple attempts to break an opponent's defensive line (invasive action) and potential for a corresponding new phase of possession

4.3.3 Operational definitions

Possession types were classified as for Study1 and the invasive actions defined in Table 4.1.

Table 4.2. Operational definitions for attempts to break the opponent's defensive line



Cross into the box	A cross from wide into the penalty area.
Dribble with a ball	Dribble with a ball close to the opponent's defensive line (always in the attacking third area).
Long pass from behind	A long-distance pass from the defensive half into the opponent's defensive line.
Pass to beyond the opponent's defensive line	The attacking team passes to a teammate beyond the opponent's defensive line.
Pass to the opponent's defensive line	The attacking team passes to a teammate near or on the opponent's defensive line.
Pass to the wide area	The attacking team passes from centre to side player who is located near the opponent defensive line.
None	No invasive action

4.3.4 Reliability

Intra- and inter-operator reliability tests (Appendix 4.1 and 4.2) were undertaken using the same procedures as Study 1. Two randomly selected matches determined coding accuracy for the new measure invasive action (inter = 0.85 and intra = 0.85) as well as starting pitch area (inter = 0.90 and intra = 0.95) and phase type (inter = 0.87 and intra = 0.87).

4.3.5 Statistical Analysis

All data were exported into Microsoft Excel version 2010 (Microsoft, Redmond, USA) and the data analysed in IBM SPSS modeler version 18.2 (IBM Corp, Armonk, NY, USA).

Descriptive statistics (mean and standard deviation) were used to describe the frequency of possessions and phases per match with differences between teams assessed by independent t-tests whose effect sizes were calculated as Cohen's d where the mean difference between the two groups was divided by the pooled standard deviation).

Association rule learning is a data mining technique (rule-based machine learning) first shown to show relationships between variables in a large customer transactions database (Agrawal *et al*., 1993). The technique enables finding items that frequently occur together and if-then patterns in large datasets (Hornik *et al*., 2005). All variables used to categorise the attacking process were assessed using association rules.

Agrawal *et al*. (1993) illustrated mining association rules as follows:

Let $I = \{i1, i2, \ldots, in\}$ be a set of *n* binary items; Let $T = \{t1, t2, \ldots, tm\}$ be a data set of transactions. Each transaction has its own ID which contained a subset of the items in *I*. Let X and Y be a set of some items in *I*, and an association rule is the implication of the form X→Y ($X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$). In this rule, X is called the antecedent or LHS (left hand side) and Y the consequent or RHS (right hand side). The data set included all variables (starting pitch area, phase type and invasive action) with association rules used to find the probability and extent of rule use for Manchester city and Liverpool. Two association rules were measured, starting pitch area (the antecedent) → phase type (the consequent) and secondly starting pitch area and phase

type (antecedent) → invasive action (consequent). Support and confidence were calculated as measures of the strength of each association rule.

Support is a measure of how often each association rule (itemset) occurs within the dataset. Hence, it is the percentage of the data set that contained an itemset (X) and the transaction (X→Y). If the support is low, the rule may have occurred by chance (Liu, 2011).

$$\text{Support (X)} = P(\text{X}) = \frac{n(X)}{n}$$

$$\text{Support (X} \rightarrow \text{Y)} = P(X \rightarrow Y) = \frac{n(X \rightarrow Y)}{n}$$

For example, the itemset starting pitch area middle 1/3 was observed in 1526 phases in the dataset of 4032 phases, hence Support was 1526/4032 = 37.85%. The transaction, middle 1/3 to build up play, occurred 781 times, hence Support was 781/4032 = 19.37%.

Confidence of an X→Y rule is the number of times the transaction (X→Y) occurred in all occurrences of the itemset (X), it is also referred to as the estimate of the conditional probability that the RHS occurs as well as the LHS i.e. P($X \rightarrow Y/X$). Hence, the confidence explains the relationship of X and Y and if the confidence is too low then the rule Y from X cannot be reliably inferred (Liu, 2011).

$$\text{Confidence (X} \rightarrow \text{Y)} = P(X \rightarrow Y/X) = \frac{n(X \rightarrow Y)}{n(X)}$$

The confidence for the transaction "Build up play from Middle 1/3" was 51.18% (781/1526) as there were 781 occurrences of "Build up play from Middle 1/3" out of the 1526 occurrences of possessions that started in the Middle 1/3.

Brin *et al*. (1997) found a key weakness of the confidence estimate to be that it was incorrect when the antecedent and consequent were independent of each other. Therefore, Lift (confidence divided by support) has become an additional measure of

interest which can be used to filter out association rules which don't satisfy both support and confidence thresholds (Hornik *et al*., 2005). The lift is the ratio of the probability of the antecedent and the consequent being observed (equals the Support) by the combined individual possibility of the antecedent and consequent (Support *X* multiplied by Support *Y*). This can be also calculated as the Confidence ($X \rightarrow Y$) divided by the Support (Y) since Support ($X \rightarrow Y$) divided by Support (X) equals Confidence ($X \rightarrow Y$).

$$\text{Lift (X} \rightarrow \text{Y)} = \frac{P(X \rightarrow Y)}{P(X) * P(Y)} = \frac{Support(X \rightarrow Y)}{Support(X) * Support(Y)}$$

$$= \frac{Confidence\ (X \rightarrow Y)}{Support\ (Y)}$$

In the "Build up from Middle 1/3" example, the support for the transaction "Build up play from Middle 1/3" was 19.37% (as calculated above); support for Middle 1/3 was 37.85% (as above) and support for Build up play was 52.18% (observed in 2104 phases in the dataset of 4032 phases). Therefore, the Lift of "Build up from Middle 1/3" was 19.37/(37.85*52.18) = 0.98. The Lift value is interpreted where 1 means the antecedent and consequent were independent; <1 signifies a negative relationship and >1 a positive relationship.

The final stage of analysis used decision tree to determine which association rules had the most influence on shot outcome. The accuracy of these models were 88.44% for Manchester city (90.4% Liverpool) with the CHAID algorithm and a p value less than 0.05 used to discriminate variables.

4.4 Results

Manchester City had a similar number of possessions per match (mean = 89.34, SD = 8.68) compared to Liverpool's 93.05 (SD = 8.99) when indeterminate possessions had been removed (Figure 4.2). When the phase methodology was applied, the number of phases per match was similar between Manchester City's 106.10 (SD = 11.60) and Liverpool 101.73 (SD = 10.55). However, Manchester City had more ($t = 6.99$, $df = 74$, $p < 0.001$, $d = 1.60$) additional phases per match (mean = 16.76, SD = 5.79) than Liverpool (mean = 8.68, SD = 4.16).

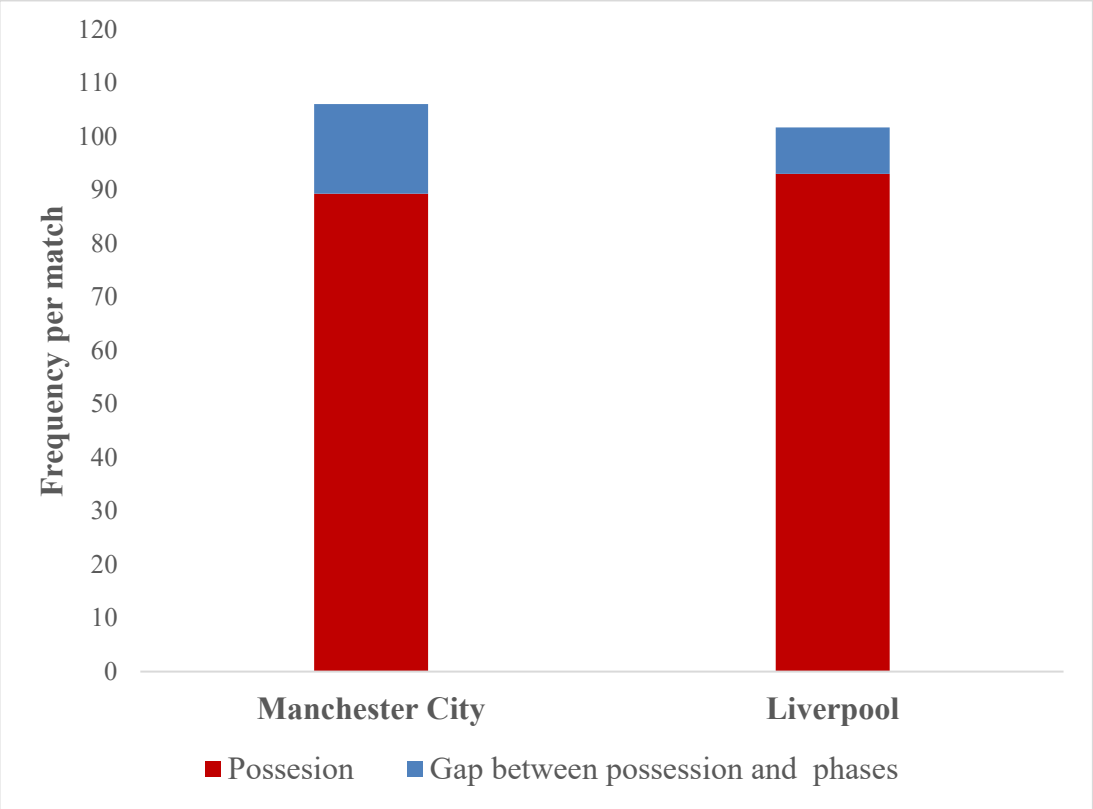


Figure 4.1. The frequency of possessions and possessional phases per match for Manchester City and Liverpool

Fifty-six association rules were found in the segmental patterns of invasive actions for each team using all 4032 phases (indeterminate possessions removed) for Manchester City (Liverpool n = 3866). The number of rules reduced to 34 for

Manchester City (33 Liverpool) when rules, where rule support was less than 1% of the total, were removed. These rules occurred less than 1.06 times per match for Manchester city (1 per match for Liverpool).

Figure 4.3 presents the association rules for the attacking process for Manchester City and Liverpool. These are layered with the initial feature (antecedent) being the starting pitch area followed by the phase type and finally the invasive action. The figure is scaled such that circle sizes signify the proportionate rule occurrence (rule support) out of the total number of phases. The thickness of the line joining two circles is proportionate to the confidence of the rule with the opacity of the line representative of the lift.

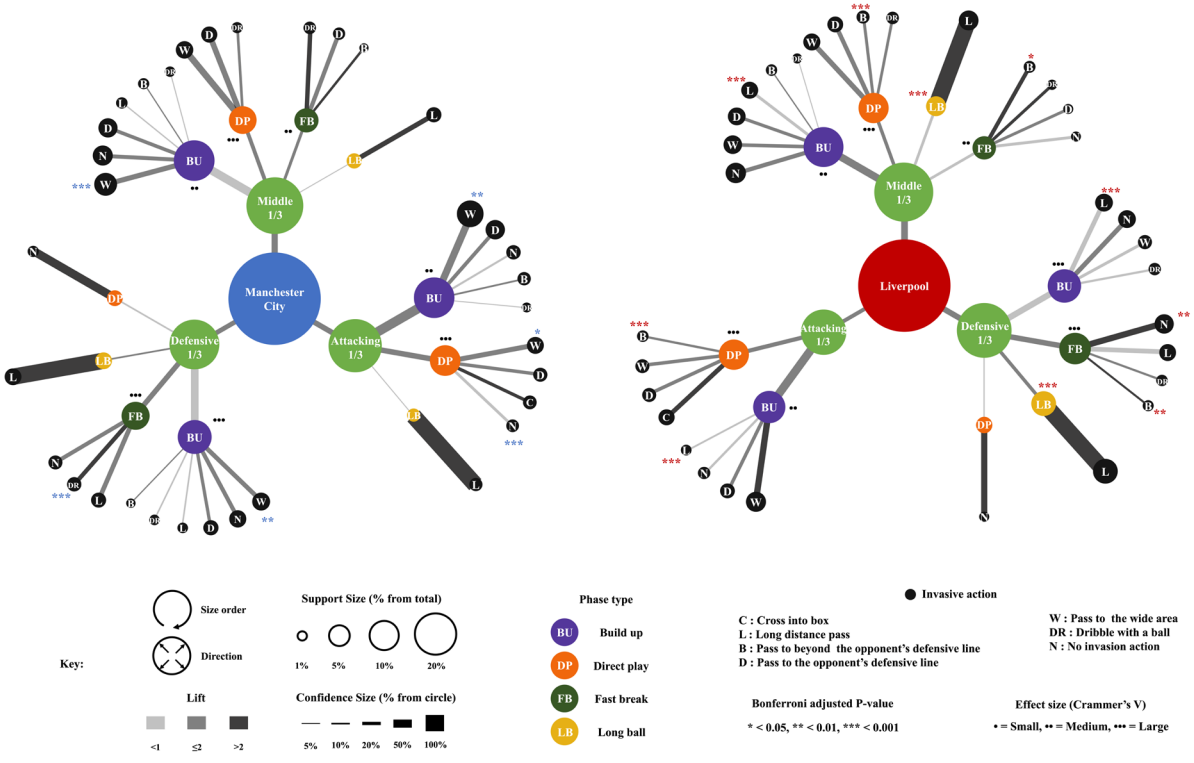


Figure 4.3. Association rules for the attacking process for Manchester City and Liverpool

Whilst the middle third was the most frequent starting pitch area for Manchester City (37.8%) and Liverpool (40.8%), all areas were commonly evident.

Build up play with passing to the wide area was the most frequent invasive action for Manchester City, significantly different to Liverpool when possession started in the middle third (Bonferroni adjusted p-value < 0.001) and the defensive and attacking thirds ($p < 0.01$). Liverpool tried significantly more long distance passes from behind during build up play, than Manchester City, in all areas ($P < 0.001$). Similarly, long ball play in the middle and defensive thirds were more prevalent for Liverpool than Manchester City ($p < 0.001$, Figure 4.3).

Shot efficiency (percentage of possessions that resulted in a shot) for each association rule was analysed by decision trees (Appendix 4.3) with the most efficient possessions presented in Figure 4.4. Overall, 11.56% of Manchester City's possessions ended up with a shot (Appendix 4.3.1). Their most efficient possessions involved passes to the wide areas or beyond the opponent's defensive line (node 1, Appendix 4.3.1). These could occur during fast breaks (21.58% resulted in shots), direct play (14.72%) or build up play (9.90%) and possessions could start in any area of the pitch (Figure 4.4). Possessions that involved a dribble as the invasive action were also very efficient (18.70%).

Figure 4.4. Shot efficiency by decision tree for Manchester City and Liverpool

Overall, 9.60% of Liverpool's possessions ended up with a shot (Appendix 4.3.2). Liverpool's most efficient attacks used passes beyond the opponent's defensive line, crosses or dribbles (Figure 4.4) with the fast break from any pitch area the most efficient (23.61%, Appendix 4.3.2).

Possessions that used a long-distance pass as the invasive action, or had no invasive action, tended to have low shot efficiencies for both Manchester City and Liverpool particularly if the possession started in the defensive or middle third of the pitch (Figure 4.4). However, Manchester City had a high shot efficiency (15.21%) when this type of possession started in the attacking area.

4.5 Discussion

The use of phases of possession has been presented here to enhance the understanding of the attacking process in football. Both Manchester City and Liverpool are well known for having a large proportion of possession in matches and this study perhaps helps to explain why. The concept of recycling the ball, passing the ball back from an attacking position (near or beyond the opponent's defensive line) to reset the attack, was presented as a new phase of the team possession. This allowed more than one game style to be described within one team possession, something not explicitly dealt with in previous studies. However, a very recent study did present the case for recycling ball out of the 18-yard box following a cross (Mitrotasios *et al*., 2022). Kim *et al*. (unpublished dissertation) also showed how one attacking process could be attempted but failed and a new attacking process was then started without losing ball possession.

The use of possessional phases resulted in Manchester City's possessions increasing by about 16% (Liverpool 9%) where these additional possessions being invisible to an analysis which simply defined team possessions according to the initial conditions. Each phase was thus akin to a new possession as the attacking process was restarted. This justification for treating individual phases as new possessions could be argued with since the second phase of a possession could be deemed to have been influenced by the first phase e.g. players positions on the pitch. This may be a beneficial consideration for future analyses but this would, potentially unnecessarily, complicate the analysis dramatically.

This study sought to develop a more comprehensive account of the attacking process in football. Hence, where the possession started and how a possession was traditionally categorised (build up, direct play, fast break and long ball play) was supplemented with how the team tried to invade the opponent's defensive area. The individual team possessions were then analysed using association rules and decision trees to determine the effectiveness of the different possession types for enabling a shot at goal. This analysis was only undertaken on the two best teams in the EPL for the season's data used. Thus, the results were exemplar in nature but demonstrative of the techniques ability to provide fine detail of both similarities and differences between two very similarly strong teams who played in relatively similar ways e.g. consistently high possession rates.

Manchester City used thirty-four different team possession types (Liverpool 33) once very infrequently occurring patterns were removed. Their most frequently used possession type was build up play culminating in a pass to the wide area to try to invade the opponent's defensive line. They utilised this attacking pattern more often than Liverpool irrespective of where the possession started on the pitch. Anecdotally,

City are well known for their very high possession rates meaning than opponents sit deep (all players in their own half) and try to crowd the critical areas in and around the penalty box (low block). This means that Manchester City have to create spaces in the opponent's defensive structure by moving the ball around, causing the defenders to move to defend different areas. Logically utilising the wide areas stretches the defensive line creating spaces for the attack to break through. These results suggest that City's coach (Pep Guardiola) supports this perceived logic and uses width as a mechanism to defeat the opponent's defence.

Liverpool tried significantly more long distance passes from deeper positions (in or near their own half) during build up play than Manchester City. This is not to describe Liverpool as a long ball team, which is generally considered the most prevalent attacking option for the less skilled teams. Rather Liverpool utilised fast transitions directly from defence to attack which is likely to be a consequence of their fast running forward players. Unlike Tottenham Hotspur, who are suggested to be a counter attack team, Liverpool used a lot of build up play but even during these attacking moments they used more long distance passes from behind than Manchester City. It would seem that Liverpool's coach Jurgen Klopp has devised an attacking strategy that fully embraces the superior speed of his players.

Whilst this study has presented the multiple methods of attack employed by the two best teams in the EPL during the season of data capture, there was no analysis of middle or low ranked teams. Hence, the information provided cannot provide a comparison of the differing tactics between these two teams with any other teams. Future analyses should consider the range of attacking strategies across and within different leagues to better understand the attacking process, particularly in respect to

individual player contributions and whether this is the most critical constraint on team tactics.

The analysis of the propensity for different attacking strategies to produce shots (referred to a shot efficiency) revealed that both teams were most successful when employing a fast break culminating with a pass beyond the defensive line. These attacks initiated with possession being gained in the defensive and middle thirds of the pitch. Whilst Gonzalez-Rodenas *et al*. (2016) also found high scoring opportunities when counterattacking from these pitch areas they did not investigate the process involved in the attack. Here, the pass beyond the defensive line was shown to be most effective, probably as a consequence of an imbalanced opponent's defence, most prevalent during a fast break (Tenga *et al*., 2010a; 2010c; Lago- Ballesteros *et al*., 2012; Lago-Peñas *et al*., 2017a). Similarly, penetrative passes used as the initial penetration of a defence produced 20.9% of scoring opportunities (Gonzalez-Rodenas *et al*., 2016); likewise, Tenga and colleagues (2010a; 2010c) found 25.4% of goals and 15.5% of scoring opportunities resulted from a penetrative pass. The actual percentages are likely to vary considerably between and within teams as a consequence of the standard of the opponent, hence from a practical perspective, the most important feature of these findings is the recognition that turnovers present teams, with fast skilful players, great opportunities to pass through the defence, particularly if the defence is low on numbers due to more players being committed to attack.

This study provided a methodology for better understanding how attacking performance develops from gaining possession to attempting to score. Kim *et al*. (2019) stated the importance of considering how different attacking procedures initiate, develop and progress attacks. On a similar theme, Hewitt *et al*. (2016) presented the importance of analysing the five moments of play i.e. differentiating between build up

play, fast attacks and set pieces. Whilst these themes were accounted for in this study, situation variables that could affect performance i.e. quality of the opponent, venue and match status were not. The focus of this study was to determine a methodological approach for finding specific strategy differences during the attacking process in open play moments between similar teams. If situation variables had been considered at this stage, there would have been too many variables for a robust examination of team performance to have taken place. However, future studies need to identify new methodologies that can enable a robust examination of how different situation variables, in isolation or in combination, affect attacking strategies in open play moments.

4.6 Conclusion

This new approach of analysing team possessions, where the ball can be recycled to restart the attacking process, aids the characterisation of possession since more than one label can be applied. More detailed process information, related to breaking the defensive line, further illuminated the attacking process, albeit in a case study of just two teams. The use of association rules provided practically useful information such that clear patterns of play were evident. This approach can be expanded by assessing performance at a more microscopic level by including specific situations based on the quality of the opponents, match status or venue.

Chapter 5: The impact of contextual variables on the attacking strategies during open play moments in football

5.1 Abstract

The way a football team attacks has been shown to be affected by contextual variables match status (e.g. Jones *et al*., 2004), match venue (e.g. Taylor *et al*., 2008), opposition quality and time period. Usually these have been assessed in isolation, but here, to present practically valid results, the combined influence of all levels of these four contextual variables were assessed. All team possessions, except those classified as indeterminate i.e. short duration and no obvious attacking process, were analysed from all matches involving the first and second placed teams in the 2018-19 English Premier League season. Possession was coded as long ball, build up, direct play or fast break with a secondary variable to classify to how the team tried to break the opponent's defensive line. If this was successful, but the team subsequently passed the ball back behind the opponent's defensive line, this was coded as a new phase of the same team possession. Decision tree models for each team, confirmed that match status was the most important contextual variable, followed by match venue. Post hoc Node diagrams illustrated how the possessional phases changed in these decision tree models. Both Manchester City and Liverpool typically maintained similar possessional phases throughout matches, suggested to reflect their ability to control matches. Manchester City's did increase their use of direct play at the expense of fast breaks during losing situations in the second half of matches, albeit a relatively unusual occurrence. Liverpool tended to have more fast breaks in winning situations and during the last 30 minutes of matches there was an increased use of direct play. The methods used in this study allowed multiple levels of 4 contextual variables to be assessed although data limitations restricted the analysis to the frequency of

possessional phases. Future studies should consider the most appropriate methods to examine multiple individual teams within an analysis that maintains practical significance.

5.2 Introduction

The analysis of football, be it technical, tactical, physical, or any combination of them, have been assessed by various variables collected during a match. The way a team plays has been shown to be affected by various contextual (also referred to as situational) variables such as match status (e.g. Jones *et al*., 2004), match venue (e.g. Taylor *et al*., 2008) and opposition quality (Hewitt *et al*., 2016). For example, match status has consistently shown differing performance levels dependent on the score in the match at the time of the event taking place. (e.g. Jones *et al*., 2004; Redwood-brown, 2008; Taylor *et al*., 2008; Lago, 2009; Lago & Dellal, 2010; Bradley *et al*., 2014; Gonzalez-Rodenas *et al*., 2015a; Gonzalez-Rodenas *et al*., 2015b; Gonzalez-Rodenas *et al*., 2016; Casal *et al*., 2017; Sarmento *et al*., 2018; Wunderlich *et al*., 2021). Similarly, contextual variables have consistently been shown to be related to performance with match venue frequently shown to influence superior performance of the home team over the away (e.g. Taylor *et al*., 2008; Lago, 2009; Lago & Dellal, 2010; Gómez *et al*., 2012; Gonzalez-Rodenas *et al*., 2015a; Gonzalez-Rodenas *et al*., 2015b; Gonzalez-Rodenas *et al*., 2016; Lago *et al*., 2016; Lago *et al*., 2017a; Sarmento *et al*., 2018; Redwood-brown *et al*., 2019; Lepschy *et al*., 2020; Wunderlich *et al*., 2021, Zhou *et al*., 2021). Other variables have been used less extensively in studies but nevertheless have consistent findings such as the quality of opposition, to some extent, determines the way a team plays (e.g. Taylor *et al*., 2008; Lago, 2009; Lago & Dellal, 2010; Bradley *et al*., 2014; Liu *et al*., 2015; Lago *et al*., 2016; Bilek & Ulas, 2019; Wunderlich *et al*., 2021; Zhou *et al*., 2021) and the time within a match utilised to assess differences in performance at different time periods, such as halves of a match (Gonzalez-Rodenas *et al*., 2015a; Gonzalez-Rodenas *et al*., 2015b; Gonzalez-Rodenas *et al*., 2016; Casal *et al*., 2017; Sarmento et al., 2018; Gonzalez-Rodenas *et*

al., 2020) or 15 minute intervals (Gómez *et al*., 2016; Zhao & Zhang, 2019; Fernandez-Navaro *et al.*, 2020; Wunderlich *et al*., 2021).

Whilst consistent findings are generally found when studies have assessed the impact of contextual variables, determining causality has rarely been the objective. For example, Gomez *et al*. (2012) assessed the frequency of goals, shots, fouls, turnovers, ball recoveries and crosses with pitch area information from 1900 matches from the 2003-04 to 2007-08 Spanish Professional Football League seasons. Factor analysis showed that the value of all variables was higher when teams were playing at home and match venue and outcome were the main determinants for all variables. Similarly, Lepschy *et al*. (2020) analysed 918 matches from the 2014-15 to 2016-17 German Bundesliga seasons. Home teams made significantly more total shots (mean = 2.08), shots on target (mean = 0.66), shots from inside the six-yard box (mean = 0.27) and shots from inside the penalty area (mean = 1.21) than away teams. Home teams also performed more crosses (mean = 2.33) and higher duel success rates (0.76%) than away teams. Whilst all of these positive attributes of performance favoured the home team the negative aspects also backed up a home advantage hypothesis. Away teams conceded more fouls (mean = 0.89), made more defensive errors (mean = 0.07) and played more clearances (mean = 3.18). Playing at home was therefore deemed a significant factor that translated to an average of 0.36 more goals scored for the home team.

Whilst large samples of matches clearly show the positive impact that contextual variables like match venue have on performance generally, it is also true that individual matches do not necessarily follow this pattern. The explanation for why a contextual variable, such as match venue, has an impact on performance is less well understood. Indeed, from an applied perspective, research does not adequately explain

whether teams deliberately play differently according to the match situation, be it playing at home or losing to an inferiorly rated opponents with ten minutes left to play. More recently, studies have tried to assess team performance, not only focusing on the outcome of performance within a contextual variable framework, but also analysing the style of play of a team. Feranadez-Navaro *et al*. (2018) assessed the influence of contextual variables on 8 game styles from all possessions of 380 matches in the 2015-2016 English Premier League. Match status, venue and opposition quality all had significant effects on playing styles. For example, direct play decreased, with teams increasing the amount of build-up play and produced a more sustained threat, when in losing situations compared to drawing. Direct play and counterattack increased, whereas maintenance, build up and sustained threat decreased when winning. These results tend to support the hypothesis that teams change the way they play in certain situations, although the truism that "you can only play as well as your opponent lets you" also has a bearing on performance. For example, if the opponents sit back in a low block (all players position themselves in their own half) as a consequence of them being in a winning situation, then build-up play and a more sustained threat (by the team who is losing) are almost inevitable.

Differing results have also been found using the same contextual variables on different data sets. For example, teams were consistently shown to have more possessions when losing than drawing and winning (Jones *et al*., 2004; Lago & Martin, 2007; Lago, 2009; Lago & Dellal, 2010; Bradley *et al.,* 2014; Fernandez-Navaro *et al.*, 2018). However, when Bloomfield *et al.* (2005) analysed the possession time for the three best teams of the 2003-04 English Premiere League they found that Manchester United had the most ball possession time when winning whereas for Chelsea this was when drawing. This was suggested to be a consequence of different

teams employing different strategies when winning, drawing or losing and thus a reflection of individual playing styles. Casal *et al.* (2017) found teams had more ball possession in winning and drawing situations compared to losing, suggesting that the increase in possession when losing, found in many studies, isn't a universal truth. The assumption that a losing team's only aim will be to score a goal, and the winning team's main objective to prevent this happening, is likely more prevalent at the end of matches. Hence, strategy changes, by both teams in this situation, will potentially only be evident at the end of matches. Machado *et al*. (2014) found that the Spanish National team kept playing their own style, including maintenance of ball possession, irrespective of match status. This may, however, be due to the Spanish team being very successful in this period and therefore dictated matches and did not have many moments where they needed to change this playing style.

Taylor *et al*. (2008) suggested the independent examination of contextual variables could only provide limited insight into the complex nature of football performance. This is because the different contextual variables could interact and any analysis not considering any particular contextual variable(s) therefore prone to simplistic and erroneous conclusions. As a consequence of this realization, studies have considered the effects of contextual variables in combination, for example, match location and opposition quality (Lago, 2009); opposition quality, match venue and match status (Taylor *et al.*, 2008; Fernandez-Navaro *et al.*, 2018).

Some researchers have postulated that performance can simply be related to the time period i.e. a potential contextual variable. For example, Gonzalez-Rodenas and colleagues (2015a, 2015b, 2016) found a higher proportion of scoring opportunities in the second half and Sarmento *et al*. (2018) showed teams had 1.29 times higher successful offensive sequence chances in the second half than the first

half. However, this was not assessed in relation to match status which could very easily be a confounding variable since teams will always try to retrieve a losing situation, if there is any possibility, at the end of the match (second half). Fifteen-minute time periods have also been analysed, for example to determine the influence of substitutions (Gomez *et al*., 2016), defensive performance (Fernandez-Navaro *et al*., 2020) and technical performance in extra time (Harper *et al.*, 2014). Again, the limitation of these studies tends to be that they didn't consider other contextual variables which could have interactive effects. The general issue regarding many of these studies is that explanations for differences, or lack of them, are not well elucidated and consideration for other influential, potentially confounding, factors also largely remiss. One study that did consider interactive effects of different contextual variables was Lago *et al*. (2016) who assessed home advantage and the effect of scoring first. Home teams won 78.8% of matches against stronger teams when they scored first during the last three time periods (45-90 mins; 67.5% if scored in 1-45 mins). In contrast, away teams won 39.8% of matches against stronger teams when they scored first in the first time period (1-15mins) and 65.8% if scored during 16-90 mins. These results show the importance of scoring first, mainly because football is a low scoring game, but also the differing effect of home advantage in relation to holding on to a lead.

The assessment of the combined effect of contextual variables on team performance has tended to use regression models (e.g. Taylor *et al*., 2008; Lago, 2009; Lago & Dellal, 2010; Bradley *et al*., 2014; Lago *et al*., 2017a; Fernandez-Navaro *et al*., 2018; Gonzalez-Rodenas *et al*., 2020). This approach uses one level of the contextual variable as the reference value with the other levels compared against it. The main findings tend to support previous research although the selection of the

reference value may have had a significant bearing and causality between variables can only be inferred rather than proven.

This study will assess the combined effect of contextual variables (match status, opposition quality, time period and venue) on team performance. Using the same methodological approach as the previous study, this study will provide detailed information about team playing styles in very specific moments of matches using decision trees to determine the use of different types of possession at all levels of each contextual variable.

5.3 Methods

5.3.1 Sample

All matches of Manchester City and Liverpool during the 2018-19 English Premier League (EPL) season were analysed. Indeterminate possessions (see Study 1) were removed as no obvious attacking process was present. Attacking phases were used to describe the attacking process rather than a single categorisation of possession (see Study 2). Hence, 3866 phases of possession by Liverpool, 4032 by Manchester City, were analysed. Ethical approval was granted by a University Ethics Committee.

5.3.2 Variables and Procedure

All matches were viewed and coded, using a purposely designed template, in Dartfish Pro S v10.0 (build 20812). Four independent contextual variables, match status (winning, drawing, losing); opponent team quality (top, middle, bottom); time period (0-15 mins, 16-30 mins, 31-end of half, 45-60 mins, 61-75 mins, 76-end of match); and venue (home, away) were coded for all data to assess their impact on team performance. Each attacking phase was categorised for starting pitch area, phase type and invasive action (as in Study 2).

End of season points, for determining team quality has been criticised (Carling *et al*. 2014; Kim *et al*., unpublished) because differing form throughout the season is unaccounted for. Hence, points gained in the previous 5 matches (Kim, unpublished) can increase the accuracy of the team quality rating. Kim (unpublished) also used points gained during the previous season, as a measure of team quality, but this was not used due to the difficulty in assigning appropriate points to promoted clubs. The opponent team quality was hence calculated by using three different measures. End of season points divided by 38, current total match points divided by number of matches already played and points gained from previous five matches divided by five. The average of these three measures was then used for each match where ≥1.5 points equated to top quality, between >1.05 and <1.5 middle and ≤1.05 bottom quality.

5.3.3 Statistical analysis

Data were exported into Microsoft Excel for Mac (v16, Microsoft, Redmond, USA) and later analysed in IBM SPSS Modeler for Mac (v18, IBM Corp, Armonk, NY, USA). Descriptive statistics determined the frequency of all variables per phase per match for both teams.

The frequency of phase types per match for all levels of all contextual variables were shown to be normally distributed for opposition quality and venue whereas skewed distributions were present for match status and time period. Therefore, the frequency of phase types per match were compared using independent t-tests for venue (home and away) and effect sizes (Cohen's d) calculated as the mean difference between the two groups divided by the pooled standard deviation. One-way ANOVA's assessed opposition team quality differences (top, middle and bottom) with effect sizes calculated as (Eta squared η^2) where the sum of squares for one variable was divided

by the total sum of squares for all variables in the ANOVA model. Non-parametric Kruskal-Wallis tests compared the frequency of phase types for match status (winning, drawing and losing) and time period (1^{st}, 2^{nd}, 3^{rd}, 4^{th}, 5^{th} and 6^{th}) with effect sizes (Epsilon squared ε^2) calculated as the Kruskal-Wallis *H* statistic multiplied by ((the number of observations plus 1) divided by (the number of observations squared) minus 1)). Decision trees were used to classify the use of phases types for different levels of the contextual variables, assigned a relative weight for their importance in relation to the target variable (phase type). The C 5.0 algorithm was applied to create multi-split and pre-pruning to avoid overfitting issues (Kohavi & Quinlan, 2002). The best fit decision trees presenting all terminated nodes, were used to determine differences in the proportion of phase types using Chi-square tests with post hoc multiple comparisons using adjusted standardised residuals, Bonferroni corrections for p values and effect size (Cramér's V) calculated as the square root of the chi-squared statistic divided by ((the sample size) multiplied by (the minimum number of levels of the two variables) minus 1)).

5.4 Results

5.4.1 The frequency of possessional phases for different contextual variables

Liverpool had more phases of possession at home (Mean = 105.74, SD = 11.51) compared to away (Mean = 97.74, SD = 7.91; t = 2.50, df= 36, $p < 0.05$; d = 0.81; Figure 5.1). This differential was similar for Manchester City but not significant (d = 0.36).

Figure 5.1. Home and away phases of possession for Manchester City and Liverpool

Manchester City and Liverpool tended to have more phases of possession when playing against middle ranked opposition, but this was only significant for Manchester City (Mean = 113.62, SD = 10.62) in comparison to their phases against top quality opponents (Mean = 98.67, SD = 9.12; $F = 6.86$, $df = 2$, $p < 0.05$; $\eta^2 = 0.88$, Figure 5.2).

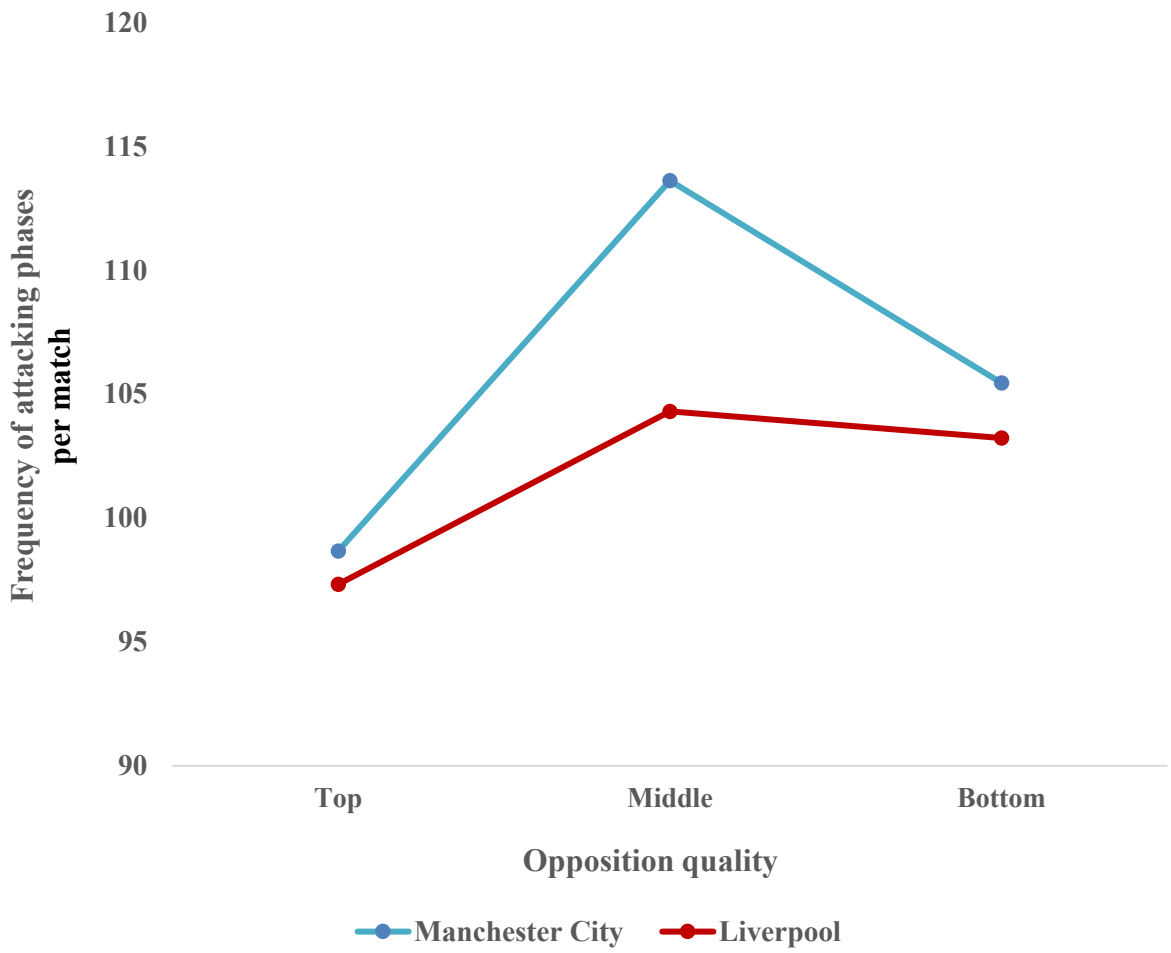


Figure 5.2. Phases of possession for Manchester City and Liverpool playing against top, middle and bottom quality opposition

The frequency of phases of possession differed according to match status for both Manchester City ($H = 62.31$, $df = 2$, $p < 0.001$; $\varepsilon^2 = 0.55$) and Liverpool ($H = 52.05$, $df = 2$, $p < 0.001$; $\varepsilon^2 = 0.46$, Figure 5.3). Both teams played more phases when winning compared to drawing with the losing status very rare for both teams resulting in zero median values and outliers.

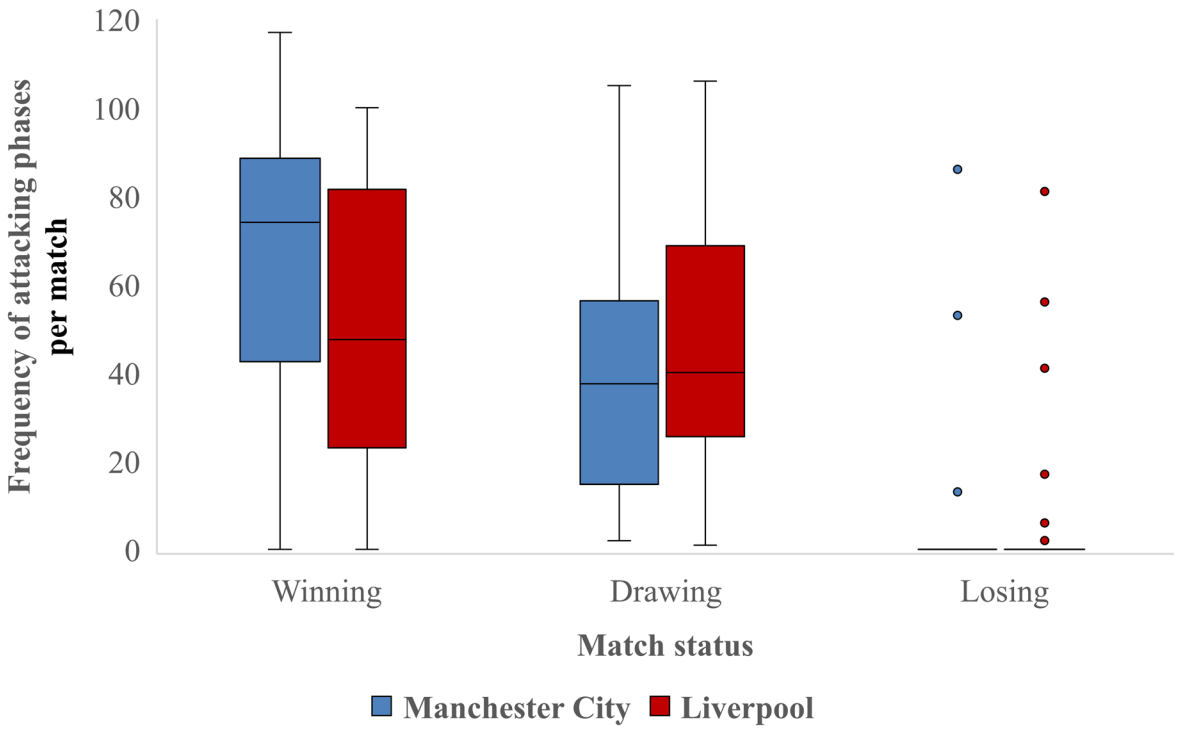


Figure 5.3. Phases of possession for Manchester City and Liverpool whilst winning, drawing and losing

Manchester City and Liverpool primarily used build up play as their preferred attacking style (Figure 5.4). However, when Manchester City were losing they tended to increase the amount of direct play (36.11%) at the expense of build-up play and fast breaks (Chi square = 36.58, df = 6, $p < 0.01$, V = 0.67). Liverpool's attacking play was similar regardless of match status (Chi square = 21.33, df = 6, $p < 0.05$, V = 0.53).

Figure 5.4. Attacking phase types for Manchester City and Liverpool whilst winning, drawing and losing

Manchester City ($H = 14.35$, $df = 5$, $p < 0.05$, $\varepsilon^2 = 0.63$) and Liverpool ($H = 23.16$, $df = 5$, $p < 0.001$, $\varepsilon^2 = 0.10$) had similar profiles for the number of phases of possession during each 15-minute time period (Figure 5.5). Manchester City had less phases in the 5th period (60-75 mins; Median = 16, IQR = 6) compared to the 6th (75 mins-end of game; Median = 19, IQR = 4, Bonferroni adjusted $p < 0.05$). Similarly, Liverpool had less phases in the 5th period (Median = 15, IQR = 5) compared to the 1st (0-15 mins; Median = 18.50, IQR = 5; Bonferroni adjusted $p < 0.05$) and 6th periods (Median = 19, IQR = 7, Bonferroni adjusted $p < 0.001$).

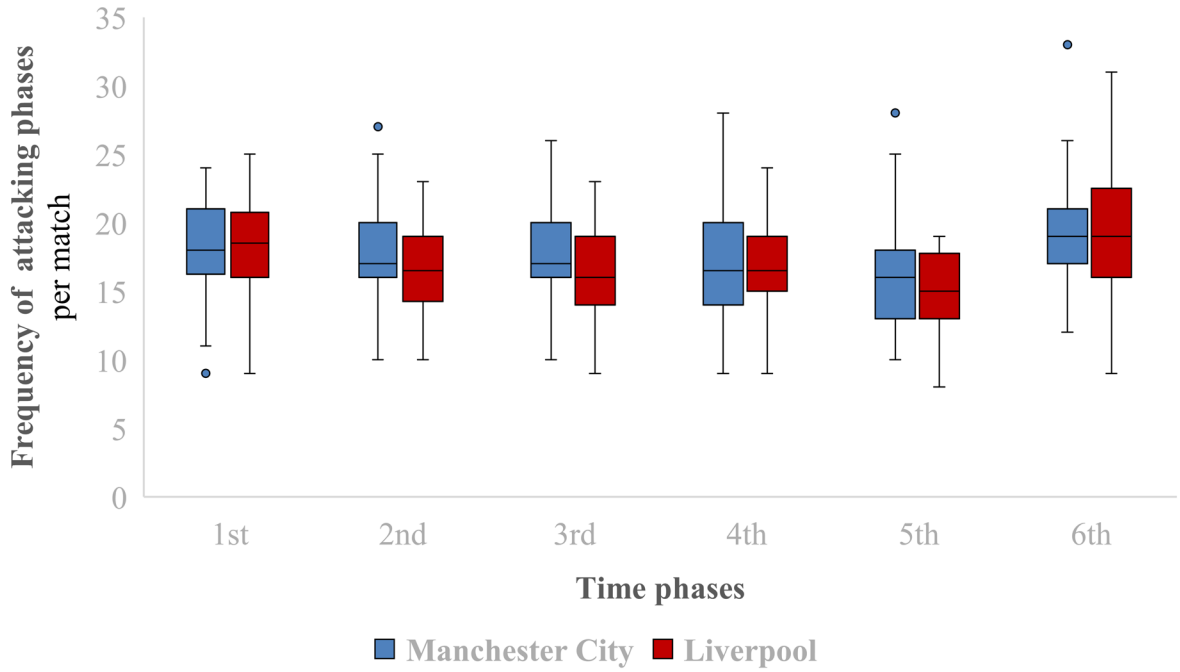


Figure 5.5. Phases of possession for Manchester City and Liverpool during 15 minutes time periods

5.4.2 The relative importance of contextual variables for determining the number of phases of possession

The contribution of four contextual variables (match status, opposition team quality, time period and venue) for determining the number of phases of possession was assessed using decision trees. When the data for both teams were assessed the most important discriminating feature between Liverpool and Manchester City was match status, accounting for 48% of difference in variation of possessional phase frequency between the two teams. In conjunction with the other variables, this gave an accuracy of 60.50% (Table 5.1). Given that the two teams differed in their performance at different levels of the contextual variables, Manchester City and Liverpool were then modelled independently. These independent models confirmed that match status was the most accurate model for both teams (Manchester City = 73.34%, Liverpool =

69.04%; Table 5.2) with time period the most important feature (Manchester City = 0.93, Liverpool = 0.72). Similarly, the second most accurate model for both teams was venue (Manchester City = 61.01%, Liverpool =63.66%; Table 5.1) with match status the most important feature (Manchester City = 0.61, Liverpool = 0.70). The models based on opposition team quality and match venue were less than 50% accurate and were therefore not analysed further.

Table 5.1. The importance and accuracy of contextual variable decision tree models for determining the frequency of possessional phases

Team	Target	Accuracy	Feature importance			
			Match status	Venue	Opposition quality	Time period
Both	Team	60.50%	**0.48**	0.14	0.10	0.28
Manchester City	Match status	73.34%	-	0.03	0.04	**0.93**
	Venue	61.01%	**0.61**	-	0.25	0.14
	Opposition team quality	46.01%	0.31	**0.35**	-	0.34
	Time phase	31.30%	1	0	0	-
Liverpool	Match status	69.04%	-	0.21	0.07	**0.72**
	Venue	63.66%	**0.70**	-	0.11	0.19
	Opposition team quality	46.15%	0.32	**0.51**	-	0.17
	Time phase	30.86%	**0.88**	0	0.12	-

Node diagrams were created to illustrate how the possessional phases changed in the decision tree models for match status and venue. These models split a node only

when different levels of a variable resulted in significantly different frequencies of possessional phases.

5.4.2.1 The match status decision tree model for possessional phases

Manchester City's decision tree classification of the number of possessional phases using match status as the primary determining variable resulted in 11 nodes i.e. where time period, opposition quality and venue significantly affected the proportion of phases (Figure 5.6). The initial classification determined the time phases (importance = 0.93) should be reclassified with the 4th, 5th and 6th time periods combined into one (node 11; Figure 5.6). Opposition quality (importance = 0.04) was only relevant for nodes 2 (2nd time period) and 5 (3rd time period). Finally, venue (importance = 0.03) only effected node 7 (3rd time period playing against middle quality opposition).

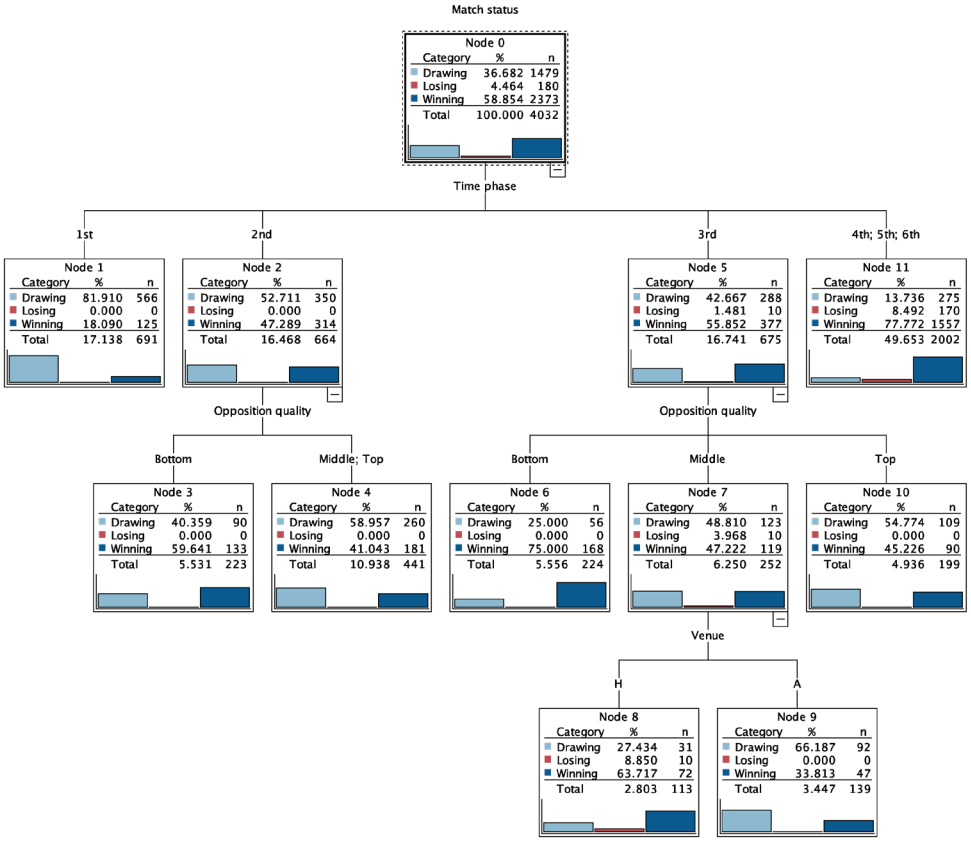


Figure 5.6. Decision tree classification of the number of possessional phases for Manchester City with match status the primary determining variable

Chi square tests compared the use of the different phase types (build up, direct play, fast break and long ball) between winning, drawing and losing situations for each of the 11 nodes (Figure 5.6). Significant differences were only found for node 11 (Manchester City's second half; Chi square = 34.94, df = 6, $p < 0.001$; V = 0.09; Figure 5.7). Manchester City used more direct play when losing (37.06%) but less direct play when winning (19.78%). They also had less fast breaks when losing (8.82%) than expected (19.13%).

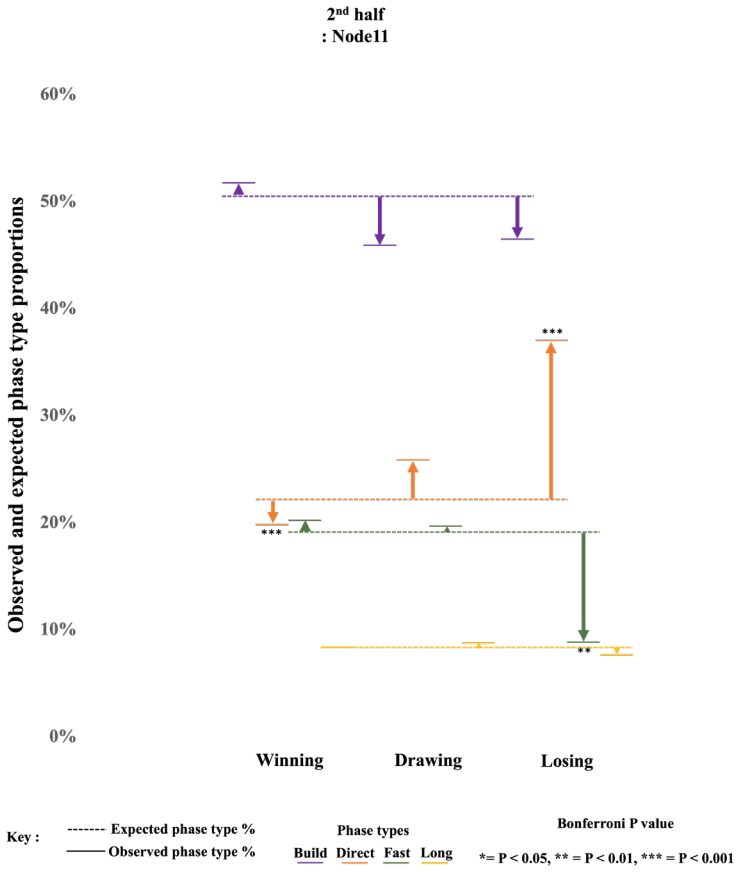


Figure 5.7. The periods of play when Manchester City played significantly different proportions of phase types between winning, drawing and losing periods of matches

Liverpool's decision tree classification of possessional phases with match status as the primary determining variable resulted in 18 nodes (Figure 5.8). As with Manchester City, time period was the first split (importance = 0.72) although only the

5th and 6th time periods were combined. Venue (importance = 0.21) was a more significant factor for Liverpool compared to Manchester City (importance = 0.03) and unlike Manchester City was more important than opposition quality (importance = 0.07).

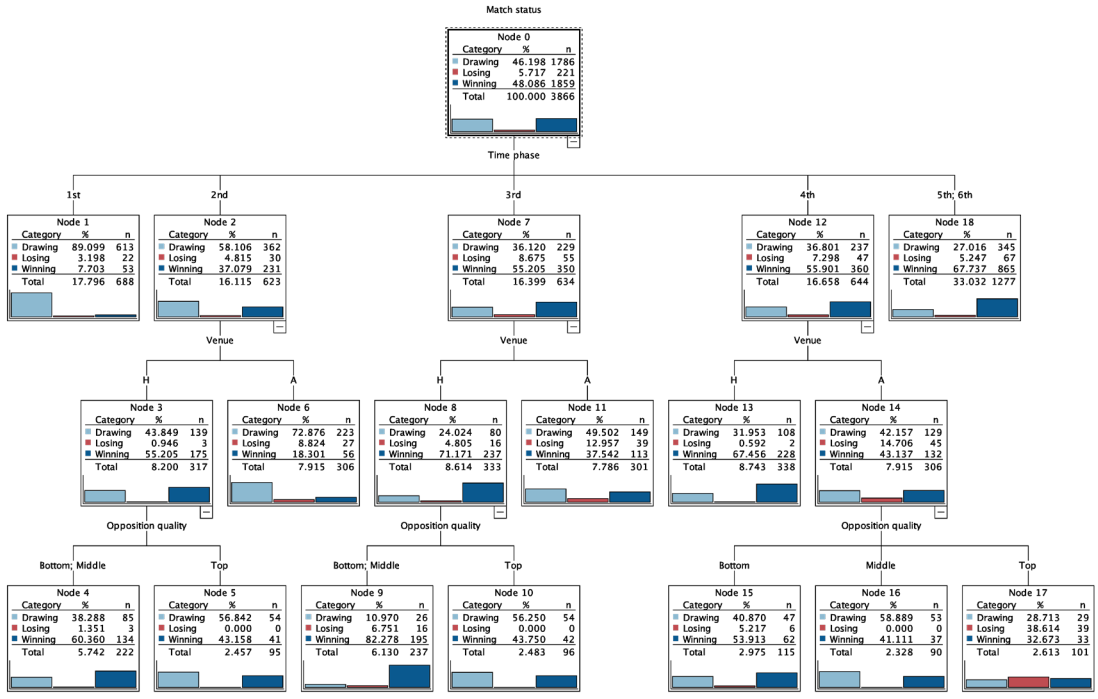


Figure 5.8. Decision tree classification of the number of possessional phases for Liverpool with match status the primary determining variable

Liverpool used different phase types between winning, drawing and losing situations for three (1, 16 and 18) of the 18 nodes (Figure 5.8). Liverpool had more fast breaks when winning (30.19%) in the first 15 minutes of games (node 1; Chi square = 14.47, df = 6, $p < 0.05$; V = 0.10; Figure 5.9). They also had more fast breaks when winning (48.65%) than drawing (16.98%) during the first 15 minutes of the second half, but only when playing away against middle quality teams (node 16; Chi square = 10.73, df = 3, $p < 0.05$; V = 0.35, Figure 5.9). Finally, they had more direct

plays when drawing (31.88%) in the last 30 minutes of games (node 18; Chi square = 16.98, df = 6, p < 0.01; V = 0.1; Figure 5.9).

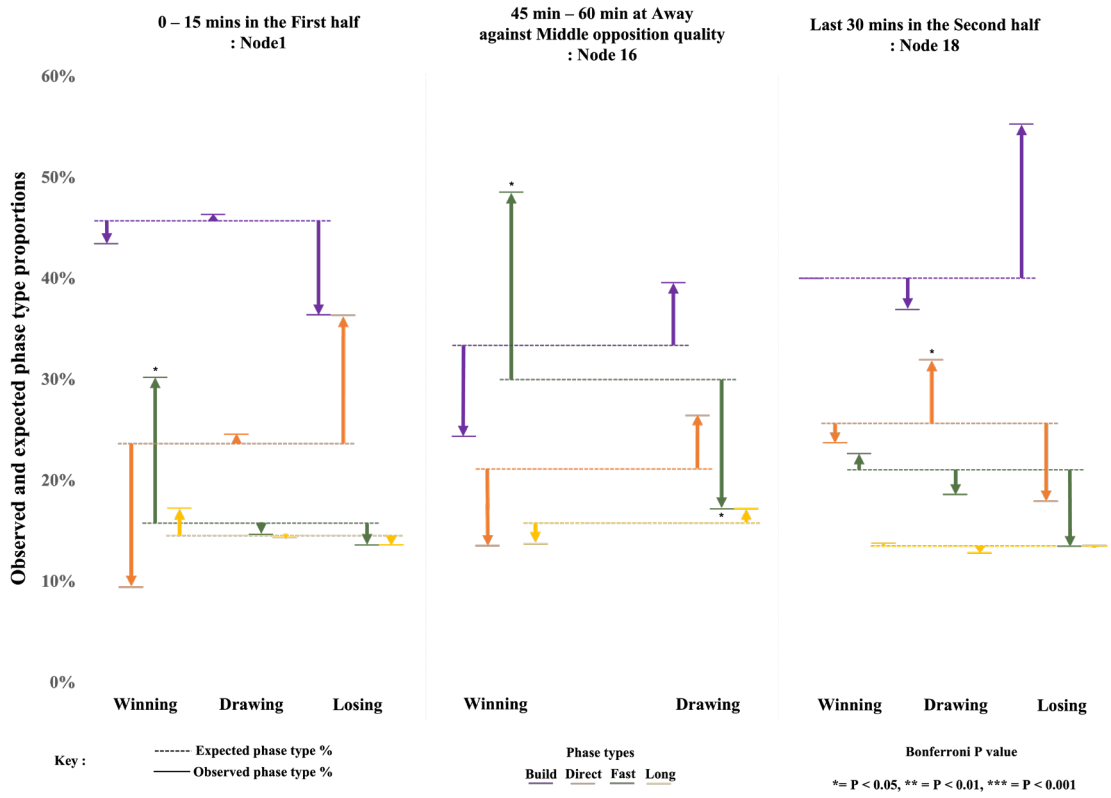


Figure 5.9. The periods of play when Liverpool played significantly different proportions of phase types between winning, drawing and losing periods of matches

5.4.2.2 The match venue decision tree model for possessional phases

The decision tree classification for Manchester City's possessional phases using venue as the target variable resulted in 11 nodes (Figure 5.10). Match status was most important classifier (importance = 0.61; Table 2) with winning, drawing and losing forming different nodes. The second split was affected by opposition quality (importance = 0.25) with top and middle ranked opponents consistently split apart. Time phase (importance = 0.14) only split one (9, winning against middle ranked opponents; Figure 5.10).

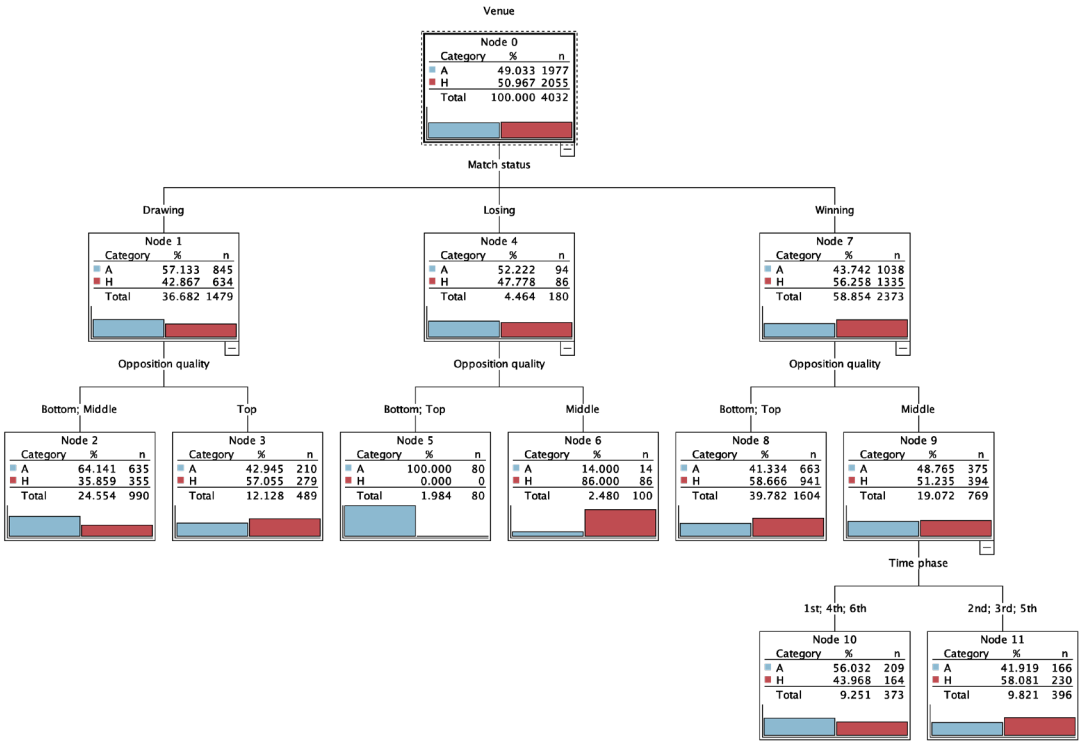


Figure 5.10. Decision tree classification of the number of possessional phases for Manchester City with match venue the primary determining variable

Manchester City used different phase types when playing home and away for two (3 and 10) of the 11 nodes (Figure 5.10). When drawing against top quality teams (node 3, Chi square = 8.95, df = 3, $p < 0.05$; V = 0.14; Figure 5.11) build up and direct play appeared to be favoured when playing at home in comparison to fast breaks and long balls away. When winning against middle quality teams in the first 15 minutes of both halves and the last 15 minutes of the match (node 10; Chi square = 9.05, df = 3, $p < 0.05$; V = 0.16, Figure 5.11) the had more build up play at home (59.76%) than away (44.98%).

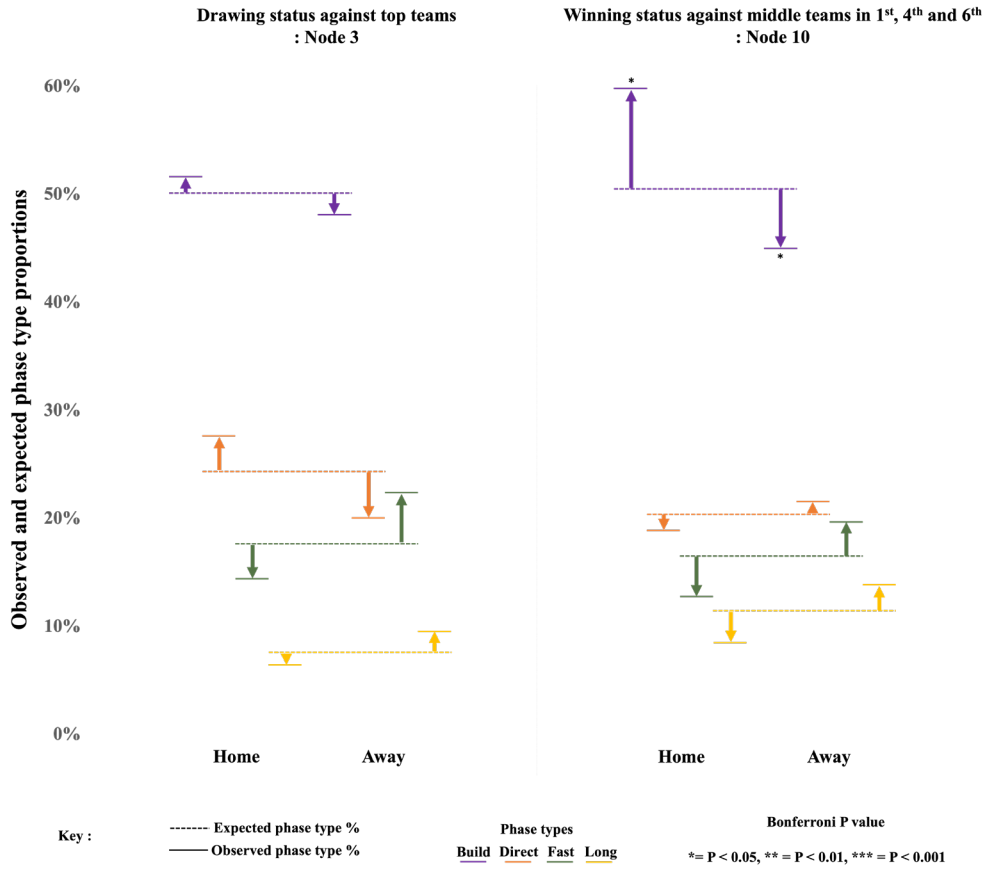


Figure 5.11. The periods of play when Manchester City played significantly different proportions of phase types between playing home and away

The decision tree classification for Liverpool's possessional phases using venue as the target variable resulted in 8 nodes (Figure 5.12). Similar to Manchester City, match status (importance = 0.70; Table 2) formed different nodes for winning, drawing and losing. However, opposition quality (importance = 0.11) only affected possessional phases when games were being drawn (node 1; Figure 5.12). Time phase (importance = 0.19) only split one node (3, drawing against middle ranked opponents; Figure 5.12).

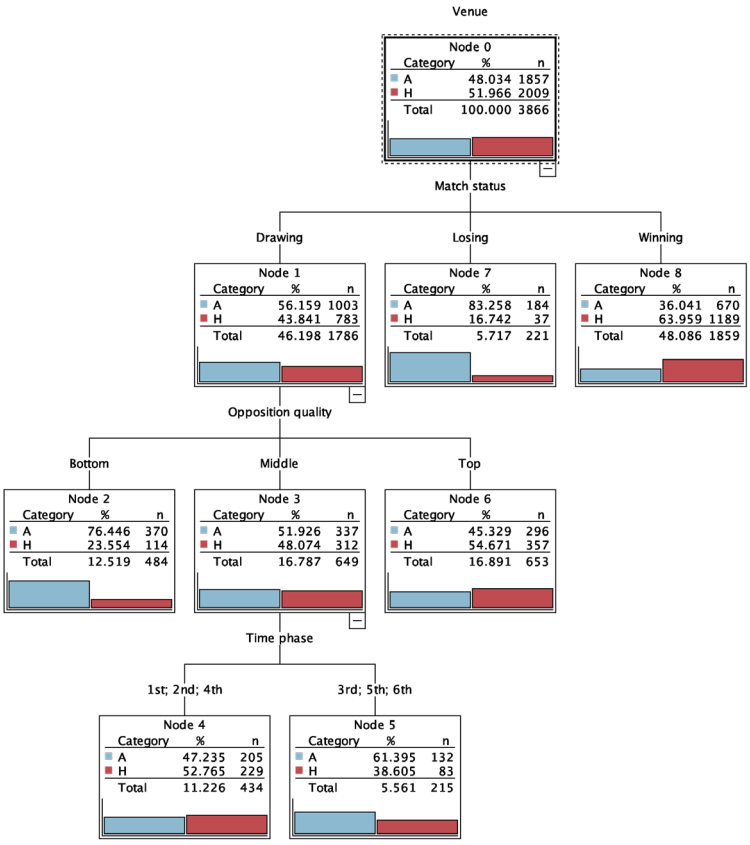


Figure 5.12. Decision tree classification of the number of possessional phases for Liverpool with match venue the primary determining variable

Liverpool used different phase types when playing home and away for five (2, 4, 5, 6 and 8) of the 8 nodes (Figure 5.12). Liverpool played more direct plays (32.75%) at home than away (19.02%) when drawing against middle quality teams in the first 30 minutes of the first half and the first 15 minutes of the second half (node 4; Chi square = 11.17, df = 3, $p < 0.05$; V = 0.16; Figure 5.13). Similarly, when they played the last 15 mins of the first half and the last 30 mins of the game when drawing against middle quality teams (node 5; Chi square = 13.92, df = 3, $p < 0.01$; V = 0.25; Figure 5.13) they played more direct play (42.17%) and more fast breaks (23.48%) at home. Finally, when they were winning (node 8; Chi square = 35.52, df = 3, $p < 0.001$ V =

0.14; Figure 5.13), they played more build up play at home (46.43%) than away (37.31%) and less long ball play at home (10.01%) than away (17.46%).

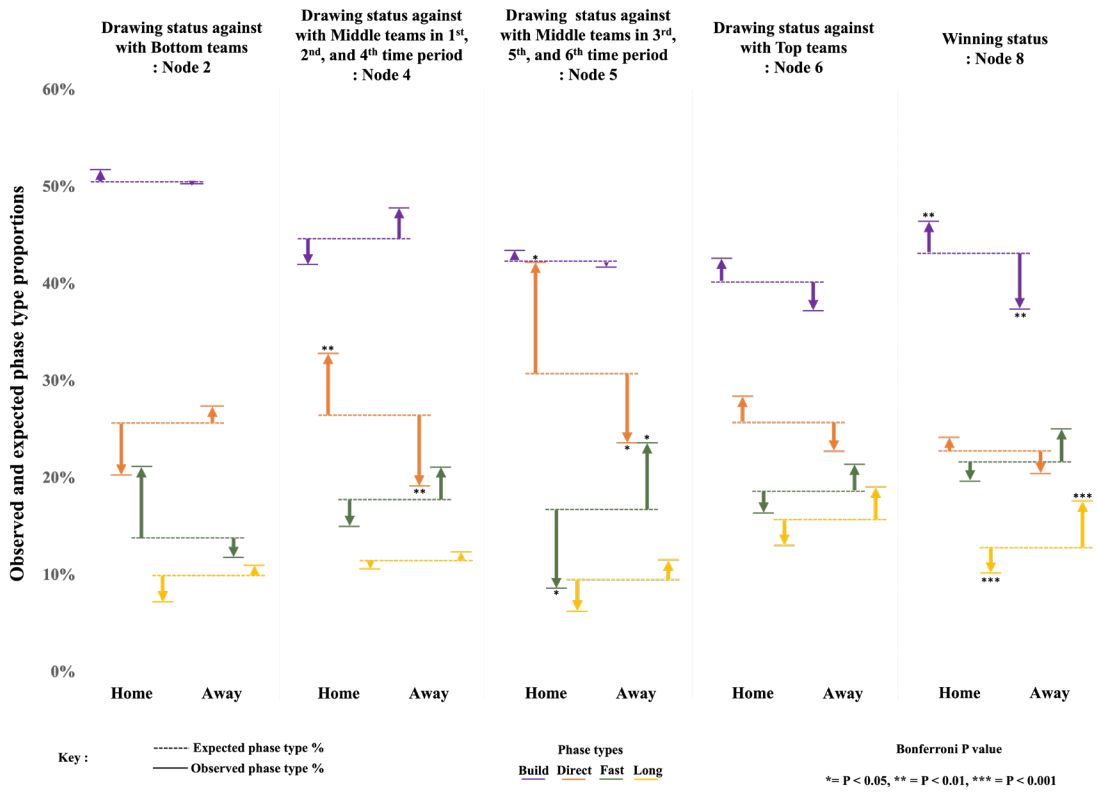


Figure 5.13. The periods of play when Liverpool played significantly different proportions of phase types between playing home and away

5.5 Discussion

In response to the paucity of studies considering the interactive effect of contextual variables this study considered four contextual variables and how they related to team performance, using Manchester City and Liverpool as exemplars. When the contextual variables were initially assessed in isolation the typical home advantage was significant for Liverpool's frequency of possessional phases but non-significant for Manchester City. This was, in all likelihood, due to the very large amount of possession they had irrespective of venue. Similarly, as would be expected, the highest quality opponents reduced the number of possessional phases that both Liverpool and

Manchester City had, albeit only significantly in comparison to middle ranked opponents against Manchester City. The explanation for why Manchester City had so many possessional phases against the middle-ranked opponents is unclear but may be due to differing tactical approaches or because they recycled the ball more often i.e. because the defence was well set up. Match status could not be fully assessed because neither team were in losing situations often, however, more possessional phases occurred when winning compared to drawing for both teams. Further analysis did reveal that when Manchester City were losing they tended to increase the amount of direct play at the expense of build-up play and fast breaks whereas Liverpool's attacking play was similar regardless of match status. This finding was a little surprising given the superiority of both teams and the general impression that Manchester City tended to remain patient and adhere to their attacking strategy. It was logical to suppose that more detailed explanations would only be forthcoming when the combined effects of the contextual variables were assessed together.

Surprisingly, both teams had fewer phases of possession in the 5th period compared to the 6th period. The fact that both teams were usually winning during these periods (Manchester City 77.8%, Liverpool 67.7%) suggests that both teams were usually in control of the matches and if they scored in the 5th period to put themselves out of reach of the opposition then potentially both teams may have "taken their foot of the gas" and allowed the opposition more possession. There were probably a few possible explanations for the lower frequency of attacking phases in this 15 minute period (relative to other periods of the match) but without more detailed analysis any conclusions would be speculative.

When the influence of the four contextual variables were assessed together, match status was found to be the variable that affected the performance of the two

teams the most (48% of the variation of possessional phase frequency). On this basis the two teams were assessed independently to determine how the combined effects of different levels of the four contextual variables influenced performance.

The decision tree models for each team confirmed that match status was the most important contextual variable, followed by match venue. Node diagrams were then used to illustrate how the possessional phases changed in these decision tree models.

For Manchester City it was evident that they were losing in the first half match for such a small period of time (less than 50% of the final 15 minutes). The change in game state, from drawing to winning, throughout the first half meant that the decision tree split the match periods into 4, the three first half periods and, because there was little change in the second half, it was treated as one. Further splits were trivial, meaning that it did not really matter what the opposition quality or match venue were, Manchester City had similar amounts of possessional phases that were only influenced by match status. The node diagrams did determine the only real change in Manchester City's attacking strategy whereby during losing situations in the second half of matches, a relatively unusual occurrence (8.5%), they increased their use of direct play at the expense of fast breaks. However, there may well have been a lack of opportunity to use a fast break in these situations on the basis that is very likely that Manchester City had even more possession of the ball than normal due to the opponents trying to protect the lead.

Similar to Manchester City, Liverpool's decision tree model for match status showed how the score-line changed significantly during the successive 15-minute periods with the exception of the final two i.e. the final 30 minutes of matches. The difference between the end of the first half and the beginning of the second was the

least obvious difference, the only noticeable change in possessional phases (8.6% to 7.3%) was when losing, but match venue had a dramatic impact from the start of the second period of the first half to the start of the second period of the second half, where much greater amounts of possessional phases were seen when winning at home. These findings showed that Liverpool's attacking strategy remained pretty consistent except they tended to have more fast breaks in winning situations (significant during the first 15 minutes of the first half in all situations but only against middles ranked team when playing away at the beginning of the second half). Like Manchester City, during the last period of matches (final 30 minutes for Liverpool) there was an increased use of direct play but the explanation for this is probably the same as for Manchester City.

When match venue was used as a target variable, match status was unsurprisingly the important discriminating factor, since it was the most important contextual variable. Possessional phases were more prevalent for Manchester City in drawing and losing situations when playing away, but at home, this occurred when they were winning. Attacking strategic changes were not very evident, except when drawing against top quality opposition, when they played more build up and direct play. Fernandez-Navarro *et al*. (2018) found a similar result where away teams decreased build up play and increased direct play compared to home. Whilst the two results are not exactly the same it can be ascertained that in certain circumstances teams are likely to adjust their attacking strategies to either overcome or maintain their opponents. The results here are more specific and hence more practically useful. This was also true for Liverpool when match venue was used as a target variable for the decision tree analysis. As for Manchester City, possessional phases were greater losing and drawing away but lower playing at home. Subtle differences in attacking strategy

were more prevalent for Liverpool than Manchester City when considering home and away performance.

The clear advantage of this study over previous ones was the degree of detail provided. Utilising four contextual variables, and examining the combination of all levels of them, allowed very specific differences to be ascertained. Attacking strategies in these different situations were able to be examined and significant departures from the usual patterns identified. However, this study had limitations that should be considered. The generalisability of the results is clearly very limited given that no analysis of other teams was undertaken. This was an accepted limitation but clearly there is scope for this to be rectified in future studies. The second limitation was a methodological one based on the sample used. The number of variables used to classify attacking strategies resulted in 56 patterns for Manchester City (55 for Liverpool), hence a maximum of 6048 possibilities when the four contextual variables were utilised. This meant that only possessional phase frequencies were considered within the contextual variable analysis to reduce the complexity to manageable levels. In future a dimension reduction technique such as factor analysis could be a solution to this problem.

5.6 Conclusion

This study applied a new method for analysing the combined effect of multiple levels of four contextual variables to better understand the attacking performance in football. This was at a level of detail useful practically for coaches and performance analysts to analyse their own and opponent's performance. Very specific changes in attacking strategy were found, at specific moments of matches, for the two best teams in the

English Premier League, who predominately dominated and won matches, and were hence able to play their own preferred attacking strategy most of the time.

Chapter 6: General discussion

This thesis has provided a number of novel techniques to assess attacking strategies in football. The main premise was to provide a more detailed analysis of the attacking process given the apparent lack of clarity surrounding three main issues. Firstly, the extant literature did not appear to address the fact that some team possessions are so short that it is impossible to describe them in any logical way that describes an attacking intent. This is a very important consideration, as if a research paper did not acknowledge these possessions, the possibility that they were falsely categorised remains a distinct possibility. Researchers may well have removed them from their analyses, but if so, explicit acknowledgement is necessary.

The second observation made at the beginning of this study was the ability of some teams, particularly Manchester City, to keep hold of the ball over relatively long periods of time. It was obvious when watching them play that they would pass the ball amongst teammates until they got to a position very close to the opponent's defensive line. They would then try to get the ball past the defensive line but if this was unsuccessful, or the opportunity failed to materialise, they would pass the ball back to a player in a deeper position. It was clear that the team possession was resetting and starting the attacking process again. In order to classify this scenario, the idea of phases of possession was formulated. This recognised, for the first time, the concept of recycling the ball and was described in this thesis as passing the ball back from an attacking position, near or beyond the opponent's defensive line, to reset the attack. This, therefore, allowed more than one attacking strategy to be described within a single team possession, something not explicit in previous studies.

The third contribution in this thesis was the classification of the ways in which a team tried to invade the opponent's defensive area. The addition of this concept

provided a more detailed account of the attacking process and enabled more clarity from an applied perspective. This was one of the main objectives of the thesis, to adopt scientifically robust methods to answer real world questions.

Clearly there are differences in methodological approaches between academic and applied analyses. Academic studies need large data sets to produce robust findings. Applied analyses only concern themselves with identifying patterns in the data that can be useful for them to gain an advantage. Hence, the decision to try to bridge this gap, referred to as the theory-practice gap by Mackenzie and Cushion (2013), was laudable but difficult. Ultimately, some compromises were likely, but the limitation of only using two teams, meaning that external validity was virtually non-existent, meant that a reasonably large data set could be robustly examined using various statistical methods. Given the novelty of the general approach, a number of relatively novel methods, at least for performance analysis of sport, were hence used.

In study 1 clustering algorithms were used to try to provide fine grained differences in attacking strategies. Given that the analysis involved the two highest ranked teams in the English Premier League, and who played with very similar styles, it was not surprising that 11 of the 16 derived attacking strategy clusters were not significantly different between the two teams. Indeed, it was considered a strength of the methodology that there were 5 significant differences. However, the actual differences found were relatively simple e.g. Manchester City, had more build up plays than Liverpool, particularly in open play possessions whereas Liverpool played more long balls particularly in open play possessions starting in the defensive and middle thirds of the pitch.

Study 2 developed the new methodology for categorising how a team tried to invade the opponent's defensive area whilst also introducing the concept of phases of

possessions. This, more detailed account of the attacking process enabled more fine-grained discriminations of, and between, Manchester City and Liverpool's attacking strategies. Indeed, thirty-four different attacking strategies were found for Manchester City (Liverpool 33), after very infrequently occurring patterns were removed. Of course, the infrequently occurring patterns may have provided some of the most interesting patterns from an applied perspective, but when utilizing robust statistical procedures they were removed for analysis. Even so, the results provided more detail in respect to the attacking strategies such that association rules and decision trees enabled it to be shown that Manchester City frequently build up play culminating in a pass to the wide area, likewise Liverpool used a lot of build up play but used more long distance passes from behind. This methodology had taken a big step towards achieving the loft goal of closing the so-called theory-practice gap but it was recognised that the importance of contextual variables had not yet been taken into consideration.

Study 3 considered the combined influence of all levels of four contextual variables; match status, match venue, opponent quality and time period within a match. These had been consistently considered in previous literature as factors influencing performance albeit often in isolation rather than the combinative effects. Interestingly, match status was found to be the most influential factor, with match venue, often regarded in the literature as the preeminent contextual variable, a relatively distant second. Decision tree models for each team and post hoc Node diagrams illustrated how Manchester City's possessional phases were only influenced by match status and the only real change in Manchester City's attacking strategy, when they were losing in the second half of matches, was they increased their use of direct play at the expense of fast breaks. Of course, as the best team in the league, it is likely that they dominated

most matches and were therefore able to stick to their game plan. Whilst this finding would not be too surprising to a football expert, it is the first time such a finding has been explicit in scientific research. The utility of this approach was further presented for Liverpool whose attacking strategy remained pretty consistent except they tended to have more fast breaks in winning situations and during the last 30 minutes of matches there was an increased use of direct play.

The progression of this thesis developed the concept of an attacking strategy from a relatively comprehensive starting point, at least from the perspective of the extant literature, to a more detailed and practically useful method. Conceptually challenging methods were applied to the data to provide a much richer account of the fine nuances of the different approaches to attacking in football by two of the most respected coaches in football.

6.1 Summary findings from the thesis

Study1

1. The case for removing **indeterminate possessions** was made. These were team possessions which could not be classified in terms of an attacking process. This is an important consideration when classifying team possessions, and not very evident in the literature. If a research paper does not acknowledge the presence of these short duration team possessions, the possibility of falsely categorising them is a distinct possibility and, the potential impact that this would have on results, present.
2. **Clustering algorithms** were used to provide fine grained differences in attacking strategies between the two highest ranked teams in the English Premier League who play with very similar styles. Eleven of the **16 derived**

attacking strategy clusters were not significantly different between the two teams. Manchester City, had more build up plays than Liverpool, particularly in open play possessions whereas Liverpool played more long balls particularly in open play possessions starting in the defensive and middle thirds of the pitch.

Study 2

1. **Phases of possession** were presented to recognise the concept of recycling the ball which was described as passing the ball back from an attacking position, near or beyond the opponent's defensive line, to reset the attack. This allowed more than one attacking strategy to be described within a single team possession, something not explicit in previous studies.
2. The way in which a team tried to **invade the opponent's defensive area** was presented. The addition of this concept provided more a detailed account of the attacking process enabling more clarity from an applied perspective. Team possessions were then analysed using **association rules** and **decision trees** to determine the effectiveness of the different possession types for enabling a shot at goal.
3. Manchester City used thirty-four different attacking strategies (Liverpool 33) once very infrequently occurring patterns were removed. Their most frequently used possession type was build up play culminating in a **pass to the wide area** to try to invade the opponent's defensive line. Liverpool used a lot of build up play but even during these attacking moments they used more **long distance passes from behind** than Manchester City.

Study 3

1. The **combined influence of all levels of four contextual variables** were assessed with **match status** found to be the most influential, it also contributed to the greatest variation in performance between the two highest ranked, and very similar teams in the English Premier League.
2. **Decision tree models** for each team, confirmed that match status was the most important contextual variable, followed by **match venue**. Post hoc **Node diagrams** illustrated how the possessional phases changed in these decision tree models.
3. Manchester City had similar amounts of possessional phases that were **only influenced by match status**. The node diagrams did determine the **only real change** in Manchester City's attacking strategy whereby during **losing situations in the second half of matches**, a relatively unusual occurrence (8.5%), they increased their use of **direct play** at the expense of **fast breaks**. Match venue had a dramatic impact from the start of the second period of the first half to the start of the second period of the second half for Liverpool, where much greater amounts of possessional phases were seen when **winning at home**. Liverpool's attacking strategy remained pretty consistent except they tended to have more **fast breaks** in winning situations and during the last 30 minutes of matches there was an increased use of **direct play**.

6.2 Key limitations identified in this thesis

1. The comparison of two teams meant the external validity of the findings were very low. No comparison with other teams was therefore made and should not

be on the basis that the specific findings only hold true for the two team analysed in the season played.

2. Individual player contributions were not assessed which meant that the overall usefulness of the approaches presented were limited. If teams adopt any of the methods presented in this thesis they should add player information, this would reduce the ability to present statistically plausible results but this is not a major concern in the applied world.
3. The number of variables used to classify attacking strategies resulted in 6048 possibile variations in the data when four contextual variables were utilised. This meant that only possessional phase frequencies were considered within the contextual variable analysis to reduce the complexity to manageable levels. In future a dimension reduction technique such as factor analysis could be a solution to this problem.

6.3 Future research directions

The attacking process in football is complex given the fact that 11 players try to overcome 11 opponents to get the ball into the opponent's goal. Over time new ways of playing have become prevalent such as the high press where players try to pressurise the opponents when they have possession in their own half. Interestingly, Reep and Benjamin (1968) suggested that regaining the ball high up the pitch created lots of goal scoring opportunities. Perhaps football strategy hasn't changed that much. In performance analysis of sport the main progression has been the development of computer science techniques to larger data sets. This transition is not without problems however, as utilising large data sets tends to amalgamate many matches involving different teams into one rather unspecific and practically useless whole. This has led to some researchers (e.g. Mackenzie and Cushion, 2013) describing the theory-

practice gap. Whether it is possible for these two ends of a continuum to merge such that scientifically robust methods can be used to answer practically relevant questions remains to be answered. This thesis has tried to address this issue with some success but clearly future studies can further address the issue with new methods that may be able to assess multiple teams within a single model to produce between team differences of interest to coaches and players alike.

6.4 Conclusions

This thesis has attempted to analyse team possessions in a more complete manner than presented in the literature. Using progressively more complex assessments the results ultimately, study 3, provided very precise periods of matches where two teams, both of whom were very successful and hence won most matches, changed their attacking strategies. This would seem to offer very practical information of use to opponents who could try to counter these attacking threats by changing their defensive systems at the appropriate times. This final outcome was achieved within the overall aim of the thesis which was to test scientifically valid methods at the level where practically useful findings could be obtained. To do this a model of the attacking process had to be formulated, based on the methods available in the literature. The main developments in this progression were to identify and remove indeterminate possessions and the classification of how teams invade the opponent's defensive area as this enabled the determination of different phases within team possessions. These methodological developments then allowed the final assessment of the attacking process at a more detailed level than previously seen in the literature. However, the amount of relevant data was deemed to be one full season as previous seasons data are less relevant due to team changes etc. This meant that the final data analysis procedures were limited to assessing the frequency of possessional phases, as more

detailed analyses within one general model would violate acceptable cases to variables ratios, due to the inclusion of four contextual variables. However, post hoc tests were able to overcome some of these limitations as the specific findings of the overall model were used to specify which data to test for significant differences. This approach has presented a novel approach to analysing a specific complex situation, namely the attacking process in football. Future studies should take this approach and further develop the methods to enable multiple teams to be differentiated.

Chapter 7: References

Agrawal, R., Imielinski, T., Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. In Proceedings of the 1993, *ACM SIGMOD International Conference on Management of Data*, 22(2), 207-216.

Alves, D. L., Osiecki, R., Palumbo, D. P., Moiano-Junior, J. V., Oneda, G., & Cruz, R. (2019). What variables can differentiate winning and losing teams in the group and final stages of the 2018 FIFA World Cup? *International Journal of Performance Analysis in Sport*, *19*(2), 248-257.

Aquino, R., Machado, J.C., Clemente, F.M., Praça, G.M., Gonçalves, L.G.C., Melli-Neto, B., Ferrari, J.V.S., Vieira, L.H.P., Puggina, E.F., & Carling, C. (2019). Comparisions of ball possession, match running performance, player prominence and team network properties according to match outcome and playing formation during the 2018 FIFA World Cup, *International Journal of Performance Analysis in Sport*, *19*(6), 1026-1037.

Aranda, R., González-Ródenas, J., López-Bondia, I., Aranda-Malavés, R., Tudela-Desantes, A., & Anguera, M. T. (2019). "REOFUT" as an Observation Tool for Tactical Analysis on Offensive Performance in Soccer: Mixed Method Perspective. *Frontiers in Psychology*, *10*, 1476.

Bate, R. (1988). Football chance: Tactics and strategy. In T. Reilly, A. Lees, K. Daivds, & W.J. Murphy (Eds.), *Science and football* (pp. 293-301). London: E & FN Spon.

Beasley, T.M., & Schumacker, R.E. (1995). Multiple regression approach to analyzing contingency tables: Post hoc and planned comparison procedures. *The Journal of Experimental Education*, *64*(1), 79-93.

Bilek, G., & Ulas, E. (2019). Predicting match outcome according to the quality of opponent in the English premier league using situational variables and team performance indicators. *International Journal of Performance Analysis in Sport*, *19*(6), 930-941.

Bloomfield, J., Polman, R., & Donoghue, P. G. (2005). Effects of score-line on team strategies in FA Premier League Soccer. *Journal of Sports Sciences*, *23*(2), 192-193.

Borrie, A., Jonsson, G. K., & Magnusson, M. S. (2002). Temporal pattern analysis and its applicability in sport: an explanation and exemplar data. *Journal of sports sciences*, *20*(10), 845-852.

Bradley, P. S., Lago-Peñas, C., Rey, E., & Sampaio, J. (2014). The influence of situational variables on ball possession in the English Premier League. *Journal of Sports Sciences*, *32*(20), 1867-1873.

Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data* (pp. 255-264).

Carling, C., Wright, C., Nelson, L. J., & Bradley, P. S. (2014). Comment on 'Performance analysis in football: A critical review and implications for future research'. *Journal of sports sciences*, *32*(1), 2-7.

Casal, C. A., Maneiro, R., Ardá, T., Marí, F. J., & Losada, J. L. (2017). Possession zone as a performance indicator in football. The game of the best teams. *Frontiers in psychology*, *8*, 1176.

Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment.

In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 263-268).

Clemente, F. M. (2018). Performance outcomes and their associations with network measures during FIFA World Cup 2018. *International Journal of Performance Analysis in Sport*, *18*(6), 1010-1023.

Clemente, F. M., Martins, F. M. L., Kalamaras, D., Wong, P. D., & Mendes, R. S. (2015a). General network analysis of national soccer teams in FIFA World Cup 2014. *International Journal of Performance Analysis in Sport*, *15*(1), 80-96.

Clemente, F. M., Martins, F. M. L., Wong, P. D., Kalamaras, D., & Mendes, R. S. (2015b). Midfielder as the prominent participant in the building attack: A network analysis of national teams in FIFA World Cup 2014. *International Journal of Performance Analysis in Sport*, *15*(2), 704-722.

Fernandez-Navarro, J., Fradua, L., Zubillaga, A., Ford, P. R., & McRobert, A. P. (2016). Attacking and defensive styles of play in soccer: analysis of Spanish and English elite teams. *Journal of sports sciences*, *34*(24), 2195-2204.

Fernandez-Navarro, J., Fradua, L., Zubillaga, A., & McRobert, A. P. (2018). Influence of contextual variables on styles of play in soccer. *International Journal of Performance Analysis in Sport*, *18*(3), 423-436.

Fernandez-Navarro, J., Ruiz-Ruiz, C., Zubillaga, A., & Fradua, L. (2020). Tactical variables related to gaining the ball in advanced zones of the soccer pitch: Analysis of differences among elite teams and the effect of contextual variables. *Frontiers in Psychology*, *10*, 3040.

Gama, J., Passos, P., Davids, K., Relvas, H., Ribeiro, J., Vaz, V., & Dias, G. (2014). Network analysis and intra-team activity in attacking phases of professional football. *International Journal of Performance Analysis in Sport*, *14*(3), 692-708.

Garratt, K., Murphy, A., & Bower, R. (2017). Passing and goal scoring characteristics in Australian A-League football. *International Journal of Performance Analysis in Sport*, *17*(1-2), 77-85.

Goes, F. R., Meerhoff, L. A., Bueno, M. J. O., Rodrigues, D. M., Moura, F. A., Brink, M. S., & Lemmink, K. A. P. M. (2021). Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *European Journal of Sport Science*, *21*(4), 481-496.

Gollan, S., Ferrar, K., & Norton, K. (2018). Characterising game styles in the English Premier League using the "moments of play" framework. *International Journal of Performance Analysis in Sport*, *18*(6), 998-1009.

Gómez, M. Á., Gómez-Lopez, M., Lago, C., & Sampaio, J. (2012). Effects of game location and final outcome on game-related statistics in each zone of the pitch in professional football. *European Journal of Sport Science*, *12*(5), 393-398.

Gómez, M. Á., Lago-Peñas, C., & Owen, L. A. (2016). The influence of substitutions on elite soccer teams' performance. *International Journal of Performance Analysis in Sport*, *16*(2), 553-568.

Gómez, M. Á., Mitrotasios, M., Armatas, V., & Lago-Peñas, C. (2018). Analysis of playing styles according to team quality and match location in Greek professional soccer. *International Journal of Performance Analysis in Sport*, *18*(6), 986-997.

Gonzalez-Rodenas, J., Calabuig, F., Lopez-Bondia, I., Aranda, R., & James, N. (2015a). Association between playing tactics and creating scoring opportunities in elite football. A case study in Spanish Football National Team. *Journal of Human Sport and Exercise*, *10*(1), 65-80.

Gonzalez-Rodenas, J., Lopez-Bondia, I., Calabuig, F., Pérez-Turpin, J. A., & Aranda, R. (2015b). The effects of playing tactics on creating scoring opportunities in random matches from US Major League Soccer. *International Journal of Performance Analysis in Sport*, *15*(3), 851-872.

Gonzalez-Rodenas, J., Lopez-Bondia, I., Calabuig, F., Pérez-Turpin, J. A., & Aranda, R. (2016). Association between playing tactics and creating scoring opportunities in counterattacks from United States Major League Soccer games. *International Journal of Performance Analysis in Sport*, *16*(2), 737-752.

Gonzalez-Rodenas, J., Mitrotasios, M., Aranda, R., & Armatas, V. (2020). Combined effects of tactical, technical and contextual factors on shooting effectiveness in European professional soccer. *International Journal of Performance Analysis in Sport*, *20*(2), 280-293.

Greenham, G., Hewitt, A., & Norton, K. (2017). A pilot study to measure game style within Australian football. *International Journal of Performance Analysis in Sport*, *17*(4), 576-585.

Hahsler, M., Grün, B., & Hornik, K. (2005). arules-A computational environment for mining association rules and frequent item sets. *Journal of statistical software*, *14*, 1-25.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). Multivariate data analysis (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Hewitt, A., Greenham, G., & Norton, K. (2016). Game style in soccer: what is it and can we quantify it? *International Journal of Performance Analysis in Sport*, *16*(1), 355-372.

Hirano, S., & Tsumoto, S. (2005). Grouping of soccer game records by multiscale comparison technique and rough clustering. In *Fifth International Conference on Hybrid Intelligent Systems (HIS'05)* (pp. 6-pp). IEEE.

Hughes, M. D., & Bartlett, R. M. (2002). The use of performance indicators in performance analysis. *Journal of sports sciences*, *20*(10), 739-754.

Hughes, M., & Franks, I. (2005). Analysis of passing sequences, shots and goals in soccer. *Journal of sports sciences*, *23*(5), 509-514.

Hughes, M., & Lovell, T. (2019). Transition to attack in elite soccer. *Journal of Human Sport & Exercise*, *14*(1), 236-253.

James, N., Mellalieu, S. D., & Hollely, C. (2002). Analysis of strategies in soccer as a function of European and domestic competition. *International Journal of Performance Analysis in Sport*, *2*(1), 85-103.

Jones, P. D., James, N., & Mellalieu, S. D. (2004). Possession as a performance indicator in soccer. *International Journal of Performance Analysis in Sport*, *4*(1), 98-102.

Kawasaki, T., Sakaue, K., Matsubara, R., & Ishizaki, S. (2019). Football pass network based on the measurement of player position by using network theory and clustering. *International Journal of Performance Analysis in Sport*, *19*(3), 381-392.

Kempe, M., Vogelbein, M., Memmert, D., & Nopp, S. (2014). Possession vs. direct play: evaluating tactical behavior in elite soccer. *International Journal of Sports Science*, *4*(6A), 35-41.

Kim, J., James, N., Parmar, N., Ali, B., & Vučković, G. (2019). The attacking process in football: a taxonomy for classifying how teams create goal scoring opportunities using a case study of crystal Palace FC. *Frontiers in psychology*, *10*, 2202.

Kim, J. (Unpublished PhD). The influence of situational variables on the attacking process in football. Middlesex University.

Kohavi R, Quinlan J.R. (2002). Data mining tasks and methods: Classification: decision-tree discovery. In Klösgen W, Zytkow JM (Eds.), *Handbook of data mining and knowledge discovery* (pp. 267–276). New York: Oxford University Press.

Konefał, M., Chmura, P., Zacharko, M., Chmura, J., Rokita, A., & Andrzejewski, M. (2018). Match outcome vs match status and frequency of selected technical activities of soccer players during UEFA Euro 2016. *International Journal of Performance Analysis in Sport*, *18*(4), 568-581.

Lago-Ballesteros, J., Lago-Peñas, C., & Rey, E. (2012). The effect of playing tactics and situational variables on achieving score-box possessions in a professional soccer team. *Journal of sports sciences*, *30*(14), 1455-1461.

Lago-Peñas, C., & Martín, R. (2007). Determinants of possession of the ball in soccer. *Journal of Sports Sciences*, *25*(9), 969–974.

Lago-Peñas, C. (2009). The influence of match location, quality of opposition, and match status on possession strategies in professional association football. *Journal of Sports Sciences*, *27*(13), 1463–1469.

Lago-Peñas, C., & Dellal, A. (2010). Ball possession strategies in elite soccer according to the evolution of the match-score: the influence of situational variables. *Journal of human kinetics*, *25*(2010), 93-100.

Lago-Peñas, C., Gómez-Ruano, M., Megías-Navarro, D., & Pollard, R. (2016). Home advantage in football: Examining the effect of scoring first on match outcome in the five major European leagues. *International Journal of Performance Analysis in Sport*, *16*(2), 411-421.

Lago-Peñas, C., Gomez, M. Á., & Pollard, R. (2017a). Home advantage in elite soccer matches. A transient effect? *International Journal of Performance Analysis in Sport*, *17*(1-2), 86-95.

Lago-Peñas, C., Gómez-Ruano, M., & Yang, G. (2017b). Styles of play in professional soccer: an approach of the Chinese Soccer Super League. *International Journal of Performance Analysis in Sport*, *17*(6), 1073-1084.

Lane, J. C., van der Ploeg, G., Greenham, G., & Norton, K. (2020). Characterisation of offensive and defensive game play trends in the Australian Football League (1999–2019). *International Journal of Performance Analysis in Sport*, *20*(4), 557-568.

Lepschy, H., Wäsche, H., & Woll, A. (2020). Success factors in football: an analysis of the German Bundesliga. *International Journal of Performance Analysis in Sport*, *20*(2), 150-164.

Liu, B. (2011). Web Data Mining: Exploring hyperlinks, contents, and usage data (2nd ed). Verlag Berlin Heidelberg: Springer.

Liu, H., Yi, Q., Giménez, J. V., Gómez, M. A., & Lago-Peñas, C. (2015). Performance profiles of football teams in the UEFA Champions League considering situational efficiency. *International Journal of Performance Analysis in Sport*, *15*(1), 371-390.

Machado, J. C., Barreira, D., & Garganta, J. (2014). The influence of match status on attacking patterns of play in elite soccer teams. *Revista Brasileira de Cineantropometria & Desempenho Humano*, *16*, 545-554.

Mackenzie, R., & Cushion, C. (2013). Performance analysis in football: A critical review and implications for future research. *Journal of sports sciences*, *31*(6), 639-676.

McLean, S., Salmon, P. M., Gorman, A. D., Read, G. J., & Solomon, C. (2017). What's in a game? A systems approach to enhancing performance analysis in football. *PloS one*, *12*(2), e0172565.

Merlin, M., Cunha, S. A., Moura, F. A., Torres, R. D. S., Gonçalves, B., & Sampaio, J. (2020). Exploring the determinants of success in different clusters of ball possession sequences in soccer. *Research in Sports Medicine*, *28*(3), 339-350.

Mitrotasios, M., Gonzalez-Rodenas, J., Armatas, V., & Aranda, R. (2019). The creation of goal scoring opportunities in professional soccer. tactical differences between spanish la liga, english premier league, german bundesliga and italian serie A. *International Journal of Performance Analysis in Sport*, *19*(3), 452-465.

Mitrotasios, M., Kubayi, A., Armatas, V., & Larkin, P. (2022). Analysis of Crossing Opportunities at the 2018 FIFA World Cup. *Montenegrin Journal of Sports Science and Medicine*, *11*(1), 43-52.

Moura, F. A., Martins, L. E. B., & Cunha, S. A. (2014). Analysis of football game-related statistics using multivariate techniques. *Journal of sports sciences*, *32*(20), 1881-1887.

Olsen, E., & Larsen, O. Y. V. I. N. D. (1997). Use of match analysis by coaches. *Science and football III*, *3*, 209-220.

Paixão, P., Sampaio, J., Almeida, C. H., & Duarte, R. (2015). How does match status affects the passing sequences of top-level European soccer teams? *International Journal of Performance Analysis in Sport*, *15*(1), 229-240.

Pratas, J. M., Volossovitch, A., & Carita, A. I. (2016). The effect of performance indicators on the time the first goal is scored in football matches. *International Journal of Performance Analysis in Sport*, *16*(1), 347-354.

Pulling, C., Eldridge, D., Ringshall, E., & Robins, M. T. (2018). Analysis of crossing at the 2014 FIFA World Cup. *International Journal of Performance Analysis in Sport*, *18*(4), 657-677.

Redwood-Brown, A. J. (2008). Passing patterns before and after goal scoring in FA Premier League Soccer. *International Journal of Performance Analysis in Sport*, *8*(3), 172-182.

Redwood-Brown, A. J., O'Donoghue, P. G., Nevill, A. M., Saward, C., & Sunderland, C. (2019). Effects of playing position, pitch location, opposition ability and team ability on the technical performance of elite soccer players in different score line states. *PloS one*, *14*(2), e0211707.

Reep, C., & Benjamin, B. (1968). Skill and chance in association football. *Journal of the Royal Statistical Society. Series A (General)*, *131*(4), 581-585.

Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*, *5*(1), 1-13.

Ridgewell, A. (2011). Passing patterns before and after scoring in the 2010 FIFA World Cup. *International Journal of Performance Analysis in Sport*, *11*(3), 562-574.

Russell, M., Rees, G., & Kingsley, M. I. (2013). Technical demands of soccer match play in the English championship. *The Journal of Strength & Conditioning Research*, *27*(10), 2869-2873.

Sarajärvi, J., Volossovitch, A., & Almeida, C. H. (2020). Analysis of headers in high-performance football: evidence from the English Premier League. *International Journal of Performance Analysis in Sport*, *20*(2), 189-205.

Sarmento, H., Anguera, T., Campaniço, J., & Leitão, J. (2010). Development and validation of a notational system to study the offensive process in football. *Medicina*, *46*(6), 401.

Sarmento, H., Anguera, M. T., Pereira, A., Marques, A., Campaniço, J., & Leitão, J. (2014). Patterns of play in the counterattack of elite football teams-A mixed method approach. *International Journal of Performance Analysis in Sport*, *14*(2), 411-427.

Sarmento, H., Figueiredo, A., Lago-Peñas, C., Milanovic, Z., Barbosa, A., Tadeu, P., & Bradley, P. S. (2018). Influence of tactical and situational variables on offensive sequences during elite football matches. *The Journal of Strength & Conditioning Research*, *32*(8), 2331-2339.

Taylor, J. B., Mellalieu, S. D., James, N., & Shearer, D. A. (2008). The influence of match location, quality of opposition, and match status on technical performance in professional association football. *Journal of sports sciences*, *26*(9), 885-895.

Tenga, A., Kanstad, D., Ronglan, L. T., & Bahr, R. (2009). Developing a new method for team match performance analysis in professional soccer and testing its reliability. *International Journal of Performance Analysis in Sport*, *9*(1), 8-25.

Tenga, A., Holme, I., Ronglan, L. T., & Bahr, R. (2010a). Effect of playing tactics on goal scoring in Norwegian professional soccer. *Journal of Sports Sciences*, *28*(3), 237-244.

Tenga, A., Holme, I., Ronglan, L. T., & Bahr, R. (2010b). Effect of playing tactics on achieving score-box possessions in a random series of team possessions from Norwegian professional soccer matches. *Journal of sports sciences*, *28*(3), 245-255.

Tenga, A., Ronglan, L. T., & Bahr, R. (2010). Measuring the effectiveness of offensive match-play in professional soccer. *European Journal of Sport Science*, *10*(4), 269-277.

Tenga, A., & Sigmundstad, E. (2011). Characteristics of goal-scoring possessions in open play: Comparing the top, in-between and bottom teams from professional soccer league. *International Journal of Performance Analysis in Sport*, *11*(3), 545-552.

Vogelbein, M., Nopp, S., & Hökelmann, A. (2014). Defensive transition in soccer–are prompt possession regains a measure of success? A quantitative analysis of

German Fußball-Bundesliga 2010/2011. *Journal of sports sciences*, *32*(11), 1076-1083.

Wade, A. (1996). Principles of Team Play, Reedswain Inc.

Wright, C., Atkins, S., Polman, R., Jones, B., & Sargeson, L. (2011). Factors associated with goals and goal scoring opportunities in professional soccer. *International Journal of Performance Analysis in Sport*, *11*(3), 438-449.

Wunderlich, F., Seck, A., & Memmert, D. (2021). The influence of randomness on goals in football decreases over time. An empirical analysis of randomness involved in goal scoring in the English Premier League. *Journal of Sports Sciences*, *39*(20), 2322-2337.

Yi, Q., Groom, R., Dai, C., Liu, H., & Gómez Ruano, M. Á. (2019). Differences in technical performance of players from 'the big five' European football leagues in the UEFA Champions League. *Frontiers in psychology*, *10*, 2738.

Yi, Q., Liu, H., Nassis, G. P., & Gómez, M. Á. (2020). Evolutionary trends of players' technical characteristics in the UEFA Champions League. *Frontiers in psychology*, *11*, 1032.

Yiannakos, A., & Armatas, V. (2006). Evaluation of the goal scoring patterns in European Championship in Portugal 2004. *International Journal of Performance Analysis in Sport*, *6*(1), 178-188.

Young, C. M., Luo, W., Gastin, P., Lai, J., & Dwyer, D. B. (2019). Understanding effective tactics in Australian football using network analysis. *International Journal of Performance Analysis in Sport*, *19*(3), 331-341.

Zhao, Y. Q., & Zhang, H. (2019). Analysis of goals in the English Premier League. *International Journal of Performance Analysis in Sport*, *19*(5), 820-831.

Zhou, C., Calvo, A. L., Robertson, S., & Gómez, M. Á. (2021). Long-term influence of technical, physical performance indicators and situational variables on match outcome in male professional Chinese soccer. *Journal of Sports Sciences*, *39*(6), 598-608.

APPENDICES

Appendix 1.1 Ethical approval Letter for all studies from London Sport Institute REC

Middlesex University London
London Sport Institute REC
The Burroughs
Hendon London
NW4 4BT
Main Switchboard: 0208 411 5000

08/05/2019

APPLICATION NUMBER: 7689

Dear Ungho Gwon

Re your application title: The categorisation of team styles in football

Supervisor: Nic James, Nimai Parmar

Co-investigators/collaborators:

Thank you for submitting your application. I can confirm that your application has been given approval from the date of this letter by the London Sport Institute REC.

Although your application has been approved, the reviewers of your application may have made some useful comments on your application. Please look at your online application again to check whether the reviewers have added any comments for you to look at.

Also, please note the following:

1. Please ensure that you contact your supervisor/research ethics committee (REC) if any changes are made to the research project which could affect your ethics approval. There is an Amendment sub-form on MORE that can be completed and submitted to your REC for further review.

2. You must notify your supervisor/REC if there is a breach in data protection management or any issues that arise that may lead to a health and safety concern or conflict of interests.

3. If you require more time to complete your research, i.e., beyond the date specified in your application, please complete the Extension sub-form on MORE and submit it your REC for review.

4. Please quote the application number in any correspondence.

5. It is important that you retain this document as evidence of research ethics approval, as it may be required for submission to external bodies (e.g., NHS, grant awarding bodies) or as part of your research report, dissemination (e.g., journal articles) and data management plan.

6. Also, please forward any other information that would be helpful in enhancing our application form and procedures - please contact MOREsupport@mdx.ac.uk to provide feedback.

Good luck with your research. Yours sincerely

Chair Dr Rhonda Cohen London Sport Institute REC

Appendix 3.1 Inter-observer reliability test results for Study 1

3.1.1 Inter-observer test for starting pitch areas

Crosstabulation

Starting pitch area		Observer 2			
		A	D	M	Total
Observer 1	A	28	1	3	32
	D	0	46	4	50
	M	1	2	67	70
	Total	29	49	74	152

Symmetric Measures

Starting pitch area		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	0.885	0.33	14.990	< .001

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

3.1.2 Inter-observer test for finishing pitch areas

Crosstabulation

Finishing pitch area		Observer 2			
		A	D	M	Total
Observer 1	A	126	0	0	126
	D	0	3	0	3
	M	2	0	21	23
	Total	128	3	21	152

Symmetric Measures

Starting pitch area		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	0.953	0.33	13.009	< .001

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

3.1.3 Inter-observer test for initiation of possession

Crosstabulation

Initiation of possession		Observer 2		
		Open	Setpiece	Total
Observer 1	Open	94	1	95
	Setpiece	0	57	57
	Total	94	58	152

Symmetric Measures

Starting pitch area		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	0.986	0.14	12.158	< .001

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

3.1.4 Inter-observer test for possession type

Crosstabulation

Possession type		Observer 2				
		Build	Directplay	Fast break	Longball	Total
Observer 1	Build	74	0	0	0	74
	Directplay	3	35	2	0	40
	Fast break	3	1	19	0	23
	Longball	2	1	1	11	15
	Total	82	37	22	11	152

Symmetric Measures

Starting pitch area		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	0.867	0.35	16.582	< .001

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

Appendix 3.2 Intra-observer reliability test results for Study 1

3.2.1 Intra-observer test for starting pitch areas

Crosstabulation

Starting pitch area		Observer 2			
		A	D	M	Total
Observer 1	A	32	1	2	35
	D	0	49	3	52
	M	0	0	76	76
	Total	32	50	81	163

Symmetric Measures

Starting pitch area		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	0.941	0.23	16.552	< .001

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

3.2.2 Intra-observer test for finishing pitch areas

Crosstabulation

Finishing pitch area		Observer 2			
		A	D	M	Total
Observer 1	A	129	0	2	131
	D	0	3	0	3
	M	2	0	27	29
	Total	131	3	29	163

Symmetric Measures

Starting pitch area		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	0.924	0.38	12.809	< .001

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

3.2.3 Intra-observer test for initiation of possession

Crosstabulation

Initiation of possession		Observer 2		
		Open	Setpiece	Total
Observer 1	Open	105	0	105
	Setpiece	7	51	58
	Total	112	51	163

Symmetric Measures

Starting pitch area		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	0.904	0.35	11.592	< .001

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

3.2.4 Intra-observer test for possession type

Crosstabulation

Possession type		Observer 2				
		Build	Directplay	Fast break	Longball	Total
Observer 1	Build	74	0	1	0	75
	Directplay	4	39	2	0	45
	Fast break	0	1	24	1	26
	Longball	2	2	1	12	17
	Total	80	42	28	13	163

Symmetric Measures

Starting pitch area		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	0.871	0.32	17.495	< .001

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

Appendix 4.1 Inter-observer reliability test results for Study 2 (also aplicable for Study 3)

4.1.1 Inter-observer test for starting pitch areas

Crosstabulation

Starting pitch area		Observer 2			
		A	D	M	Total
Observer 1	A	41	1	3	45
	D	0	46	4	50
	M	1	2	67	70
	Total	42	49	74	165

Symmetric Measures

Starting pitch area		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	0.897	0.30	16.157	< .001

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

4.1.2 Inter-observer test for possessional phase type

Crosstabulation

Phase type		Observer 2				
		Build	Directplay	Fast break	Long ball	Total
Observer 1	Build	87	0	0	0	87
	Directplay	3	35	2	0	40
	Fast break	3	1	19	0	23
	Long ball	2	1	1	11	15
	Total	95	37	22	11	165

Symmetric Measures

Starting pitch area		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	0.872	0.33	17.152	< .001

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

4.1.3 Inter-observer test for invasive actions

Crosstabulation

Invasive action		Observer 2							
		Beyond	Cross	Defense	Dribble	Long	None	Wide	Total
Observer 1	Beyond	21	0	0	0	1	0	1	23
	Cross	0	5	0	0	0	1	0	6
	Defense	1	0	24	3	0	0	0	28
	Dribble	2	0	1	14	0	0	0	17
	Long	1	0	1	1	24	2	0	29
	None	1	0	2	1	0	19	0	23
	Wide	2	0	0	0	0	0	37	39
	Total	28	5	28	19	25	22	38	165

Symmetric Measures

Starting pitch area		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	0.848	0.31	24.890	< .001

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

Appendix 4.2 Intra-observer reliability test results for Study 2 (also aplicable for Study 3)

4.2.1 Intra-observer test for starting pitch areas

Crosstabulation

Starting pitch area		Observer 2			
		A	D	M	Total
Observer 1	A	47	1	2	50
	D	0	49	3	52
	M	0	0	76	76
	Total	47	50	81	178

Symmetric Measures

Starting pitch area		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	0.948	0.21	17.735	< .001

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

4.2.2 Intra-observer test for possessional phase type

Crosstabulation

Phase type		Observer 2				
		Build	Directplay	Fast break	Longball	Total
Observer 1	Build	88	1	1	0	90
	Directplay	4	39	2	0	45
	Fast break	0	1	24	1	26
	Longball	2	2	1	12	17
	Total	94	43	28	13	178

Symmetric Measures

Starting pitch area		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	0.869	0.32	17.947	< .001

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

4.2.3 Intra-observer test for invasive actions

Crosstabulation

Invasive action		Observer 2							
		Beyond	Cross	Defense	Dribble	Long	None	Wide	Total
Observer 1	Beyond	20	0	3	0	0	0	0	23
	Cross	1	6	0	0	0	0	0	7
	Defense	2	0	26	2	0	0	2	32
	Dribble	0	0	1	17	0	0	0	18
	Long	1	0	2	1	28	0	1	33
	None	0	0	0	0	0	25	0	25
	Wide	1	0	3	2	0	0	34	40
	Total	25	6	35	22	28	25	37	178

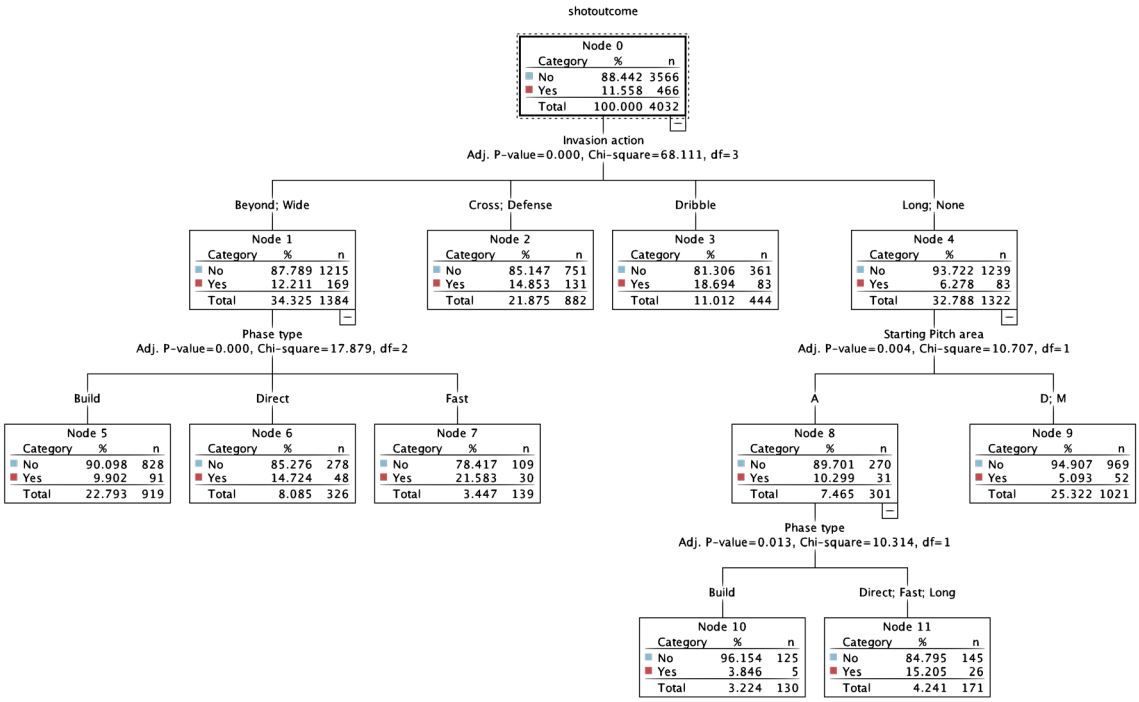
Symmetric Measures

Starting pitch area		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	0.852	0.30	26.139	< .001

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

Appendix 4.3 Decision tree for shot efficiency

4.3.1 Decision tree for the chance of a shot occuring depending on the invasive action, possession type and starting pitch area for Manchester City



4.3.2 Decision tree for the chance of a shot occuring depending on the invasive action, possession type and starting pitch area for Liverpool