City University London

MSc in Data Science

Project Report

2016

The use of Data Science Techniques to Help Understand a Football Player's Skillset and Find Their Optimum Playing Position

Ella Walters

Supervised by: Cagatay Turkay

Submitted: 09-12-2016

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed: *Ella Walters*

ABSTRACT

This study explores how data science techniques can be used to provide an in depth quantifiable measurement on where a football player performs best on the pitch. It proposes that standard positional footballing labels such as defender, midfielder and striker are too broad, given the complexities of the game. Using Principal Component Analysis and Random Forest models the study allows for much closer scrutiny of where exactly within these broad positions a player is most successful. For example, it compares a striker who performs best on the inside of the opposition's goal box to a striker who performs best outside the opposition's box. The research uses Opta's entire 2013-2014 Premier League dataset to create a 'Performance Indicator' for each on-the-ball event, modelling 100 different areas of the pitch, and a 'Performance Metric' which relates this back to the individual players. From what can be ascertained given the confidential nature of top-flight football clubs this is an innovative area of research.

Keywords: Principal Component Analysis, Random Forests, Mean Decrease Gini, Football Positions, Premier League.

Table of Contents

1 - Introduction and Objectives:

1.1. Formatting and Technical Terms

This paper frequently uses the term on-the-ball event to broadly define a footballing action such as a pass, tackle or shot. A full list of all on-the-ball events can be found in the F24: Feed appendices document.

The term 'event_qualifier_id' is used to refer to a specific on-the-ball event with a qualifier associated to it. So in the case of 'pass_long ball' pass is the 'event_id' and long ball the 'qualifier_id'.

Variable names from the dataset are written in 'single speechmarks'

Performance Indicator is the term used for the models created which gives values to all 100 segments of the pitch for each on-the-ball event irrespective of player.

Player Metric is the term used for the values created by transforming the frequency of player on-the-ball events for each segment of the pitch.

Performance Metric is the term used for the model created which combines the Performance Indicator and Player Metric to give the overall performance of a player on each of the 100 segments of the pitch.

Names of R packages are presented using *black italic*

Models are referred to in blue in the form XY Model where X is the number of seconds and Y is the number of previous events e.g. 3060 Model is every 30 seconds for the previous 60 events.

1.2. The Background, Reasons and Beneficiaries

On 18th March 1950, Charles Reep, an accountant from Cornwall, sat down to watch Swindon play Bristol Rovers. Using a pencil and piece of paper he documented the match, breaking down the continuous flow of the game into discrete actions of every on-the-ball event (Anderson & Sally, 2013). He created shorthand codes for each major action such as passes and shots and included not only their occurrence but also where the pass began and ended, its length, direction, height and outcome (Pollard & Reep, 1997). The match was the first of over 2000 matches he documented. This data was heavily analysed over many years and produced

some novel findings that would shape the game of football as we know it.

Historically, football has been a game of subjective decisions; where spectators watch and provide personal opinions. However, now, more than ever, statistics and basic analytics are used to help inform choices, with most top level clubs utilising the skills of in-house analysts to help player and team development. Previously, the issue had been how to *gather* the data; now it is how to *use* it effectively. Today, there are whole companies set up to collect sports data at the most intricate levels. Opta, the world's leading sports data provider, record between 1,600 and 2,000 events for every major football match (Opta, n.d.). They combine the use of analysts (three per game who sit and record all events) with modern day technology, sometimes also incorporating data collected from stadiums to attain fully timestamped x,y coordinate detail of player actions on the pitch (Opta, n.d.). With the growth of big data and data science techniques it seems only natural that Machine Learning and Artificial Intelligence algorithms will be used more frequently to provide insight into the game.

This seems particularly likely because of the large, and growing, sums of money involved. This was illustrated recently when Paul Pogba moved from Juventus to Manchester United Football Club for a record transfer fee of £89m (Lake, 2016). Given that there are hundreds of high quality midfielders available it was not known what made manager Joes Mourinho and his team decide that Pogba was worth this record-breaking amount. No detailed research has been publicly shared beyond that of watching Pogba play and analysing his basic match statistics (goals scored, possession etc.). It is not known whether data science was influential in making this decision.

Due to the competitive nature of the sport and therefore the privacy elements attached to the game it is difficult to know exactly what methods top flight clubs use today. Football clubs are certainly aware of the developments in data science. In 2012, Manchester City Football Club, in collaboration with Opta, opened the 2011-2012 on-the-ball dataset for public use for the first time. Recognising a gap in the availability of data for non-professionals it aimed to explore how students, analysts or just lovers of the game may help fill the knowledge gap in this area. Even if clubs do not publicly discuss their use of data analytics this exemplifies their awareness of the value it may hold.

At top levels the pressure to perform is unbounded; between 2012 and 2015 Chelsea football club hired and fired four different managers for not making the right choices or getting the required results, in some cases paying out large fees for the right to do so. In a time when big data is prevalent in all areas of the business world it seems only natural that it will have an input into such key business decisions within football and be the natural next step in helping clubs stay ahead of their rivals.

This paper focuses specifically on player positions. Wayne Rooney was recently the topic of much debate regarding his position both within Manchester United and the England team. Having always played a striking position Jose Mourinho, the Manchester United manager, publicly said he didn't believe Rooney had the pace and sharpness needed to continue at the top of the pitch. "*Maybe he [Rooney] is not a striker, not a No. 9 anymore but for me he will never be a No. 6…. To be there and put the ball in the net is the most difficult thing. For me he will be a 9, a 10, a nine-and-a-half but never a 6 or even an 8.*"[1] (Ducker, 2016) It is unclear what research has been done to support this statement but from the frequent use of the word 'me' it would seem that Mourinho's subjective opinion was a main factor.

As it stands there is no *known* quantifiable measure that decides where on the pitch a player is best suited to play. The formation of a team generally defines the positions available for the players. A common formation is the 4-4-2 shown in Figure 1, which has four defenders, four midfielders and two strikers. Many attributed Leicester City's unexpected win of the 2015-2016 Premiership season down to its use, suggesting it was the best fit for the players on the team (Bull, 2016). However, considering the size of the pitch, the variance of player skills and the complex nature of the game it seems naive to group players in such a limited way. This research discusses what specific skillsets define the positions of individual players and whether these standard team formations are too broad.

[1] No.6 position = Defensive Midfielder, No.8 position = Central Midfielder, No.9 position = Striker, No. 10 position = Striker.

Figure 1 4-4-2 formation

This study will be beneficial for a number of different footballing professionals.

Player Coaches – The Numbers Game (Anderson & Sally, 2013) describes how approximately 50% of match results come down to chance. The aim here is to make sure that the other 50% of results can be optimised with the help of efficiently using large spatio-temporal data sets. If coaches can understand the features of a player that are most important when picking certain positions, they can also use this as a training tool and focus on improving these skills.

Team Managers – With performance analytics becoming more important, coaches and managers are analysing statistics much more frequently. A run in between Wayne Rooney and Sam Allardyce, the England manager for one game, exposed that it is the manager rather than the player who ultimately decides what position a footballer plays. With Rooney quoted as saying: "*I'll play wherever the manager wants me. I don't pick myself, I haven't ever picked myself*" (Taylor, 2016). This is a clear example of how the subjective nature of decision-making in the game has an impact on player/manager relationships. It also highlights the pressures placed on the managers to make the right choices.

Club Scouts – A player's strengths could more easily be recognised when trying to purchase a player for a certain position. On the flip side it would also be an advantageous tool for scouting the opponent's players to understand the strengths and weaknesses of the opposition.

This paper, with the help of the 2013-2014 Premier League Opta dataset (Opta, n.d.), aims to answer the question:

Can Data Science Techniques be used to Help Understand a Player's Skillset and Find their Optimum Playing Position?

1.3. Objectives

In order to answer the question posed in this research paper there were a number of key objectives that needed to be completed.

- **Understand the characteristics of different footballers' positions on the football pitch**. Understand what on-the-ball events (e.g. pass, tackle, shot) define a player and whether similar positions have similar events related to them. The output of clustering analysis will provide insight into the similarities and differences in each known playing position.

- **Find the on-the-ball events that are most important in defining a player's position**. Create a sub-group of variables which are deemed the most important when defining a player's position. For example, is tackling more important for a defender than a striker?

- **Find the on-the-ball events that are most important relative to success.** Understand how to quantify success in a football match and from this extract the key variables which contribute to success to create a Performance Indicator. For example, are passes more important than tackles when scoring goals?

- **Find a player's optimum position on the pitch.** Create a visual and quantifiable Performance Metric which relates where a player has played relative to where the model would have played them. For example, does a striker who plays predominantly

on the left wing actually perform better on the right wing?

1.4. Methods for Completing Objectives

- **Understand the characteristics of different footballers' positions on the football pitch**. This was achieved using clustering techniques and Principal Component Analysis, with the aim of understanding whether there was any relationship between the number of times a player carried out a certain on-the-ball event and their position on the pitch. The output of the Principal Component Analysis also provided insight into the type of events that define each position by the value of their loadings in the top components.

- **Find the events that are most important in defining a player's position**. Running a Random Forest with the number of times each player completed an on-the-ball event as the input variables, and the position of the player as the target variable, produced a tool for finding the most important variables in order to reduce the size of the dataset to a more manageable amount. Combining this with the exploration of frequencies of each event on the pitch provided a succinct set of variables for modelling.

- **Find the events that are most important relative to success.** Converting the Opta dataset into paths was needed to understand what sequence of events leads to success. Using this dataset, the output of a number of Random Forest models for different areas of the pitch was used to create an on-the-ball event based Performance Indicator that could be related back to individual players. The Performance Indicator was tested by visualisation to ensure no biases occurred towards certain areas of the pitch.

- **Find a player's optimum position on the pitch.** The aim of this objective was to use the Performance Indicator combined with individual player attributes to create a Performance Metric for each player. To test the model's accuracy, the real-life results of a successful team in the season were compared with an unsuccessful team in the season and related back to the model's analysis of those teams. This was also done at an individual player level, comparing real-life facts about the players with the model's analysis of them for that season.

1.5. Work Plan

The work plan shown in Figure 2 varies slightly from the initial plan shown in the proposal (Appendix A). In general, the order of the tasks stayed the same, with only one or two revisions. Aware of the complex nature of the dataset, the parsing of data remained one of the first tasks given a significant amount of time to complete. Creating features by player was used to complete the first objective of understanding characteristics of different footballing positions. Dimensionality reduction was used for the second objective of finding the most important on-the-ball events when defining a player's positions. Creating the Performance Indicator answered the third objective of obtaining the most important on-the-ball events relative to success and creating the Performance Metric related to the fourth main objective of finding a player's optimum position. Finally, evaluation of results and writing up were given significant amounts of time to ensure the project and results were correctly presented.

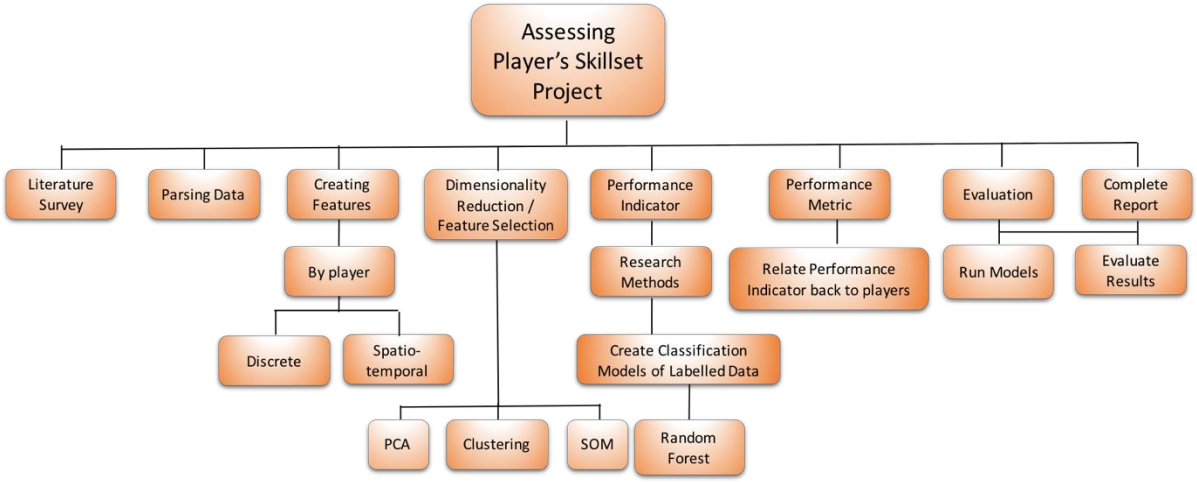


Figure 2 Work plan

1.5. Changes in Goals and Methods

The main goal of this project did not change substantially from the initial proposal but many tangential routes were explored before arriving at the final output. Self Organising Map's (SOMs) were used when initially exploring the dataset but it was soon evident that they did not provide any additional value to the research. Instead they merely presented the initial

dataset without performing any useful evaluations. They were therefore excluded from the results. Initially, it was proposed to compare and contrast a number of different models when creating the Performance Indicator. However, the time it took to run the Random Forests and test for robustness meant that it was deemed more suitable to stick to one method and research it fully.

1.6. Structure of the Report

Section 1 provides an introduction to the problem, highlighting the four main objectives and the key methods for answering them.

Section 2 provides context surrounding the use of data in sport and more specifically football today. It also describes the theory behind the key algorithms used in this paper.

Section 3 describes in detail the methods needed to complete the objectives. This includes how to understand the characteristics of different positions on the pitch, how to find the most important on-the-ball events relative to position and success for use in creating a Performance Indicator and finally how to relate this Performance Indicator back to each player to find their optimum areas of play on the pitch.

Section 4 explains the outputs produced by using the methods. It highlights some of the more interesting player properties.

Section 5 discusses these results in comparison to the objectives.

Section 6 evaluates the project as a whole and recommends areas for further work.

2 - Context:

2.1. The Use of Data in Sport Today

2.1.1 Evolution of Sport Driven by Data

The release of Moneyball (Lewis, 2003) in 2003 was the beginning of a new phase of interest in the positive effect statistics could have in sport. Lewis documents how the Oakland Athletics baseball team overcame financial disadvantages by using data and analytical measures to remain a top-flight team. It focuses on the weaknesses caused by the subjective nature of the managers and coaches when assigning value to players and highlights the benefits of statistical analysis.

In 2005, Dean Oliver released Basketball on Paper, a book that uses statistical tools to help explain the winning and losing ways of basketball teams (Oliver, 2005). Basketball is a more difficult game to interpret than baseball due to its continuous flow of play and interaction of players and so Oliver's research was extremely popular due to its effectiveness at the time (Lucey, et al., 2013). In the same year, volume 1 of the Journal of Quantitative Analysis in Sports was released, marking a turning point in the use of data in sport.

In recent years information has become more readily available and the ever-evolving world of data collection in sports science, nutrition and coaching has continued to grow (Lucey, et al., 2013). The world-wide success of research joined with this growth of information has transformed sport with the majority of all major clubs now employing in-house analytics teams as part of their continuing efforts to stay ahead of their rivals. The use of data in professional sports is becoming more common, not only for betting and gambling purposes but also to help understand individual players and improve the quality of the team (Lucey, et al., 2013). This fact is widely known but the results of such in-house research are generally kept confidential.

Discrete statistics in sport have been around for a long time e.g. number of passes, shots on target, total possession. However, with most sports involving multiple players who are constantly moving and interacting with each other, these simple statistics do not suffice to capture the complex nature of the game (Lucey, et al., 2013). To overcome this in recent years, tracking devices which can locate the position of the ball and players at any time have become

prevalent, increasing exponentially the size of datasets available. The availability of spatio-temporal data in sport is something that provides opportunity to not just understand what happened, but also to understand how and why (Lucey, et al., 2013). In 2015 FIFA and The Football League allowed the use of electronic performance and tracking system devices. Whole companies devoted to collecting sports data such as Opta (Opta, n.d.) and STATs (STATS, n.d.) now exist and are a main source of both data and analytics for top clubs.

2.1.2 Complexities of Analytics in Continuous Sports

Although there is now an abundance of data available there are still many issues to be overcome in the analysis of that data. The major problems in continuous sports are the uninterrupted flow of play, the complexity of interactions between players and the low scoring events (Duch, et al., 2010). The uninterrupted flow of play and complexity of interactions between players means it is not only on-the-ball events that are important but also what is happening off the ball. This dimensional complexity makes modelling the game a real challenge in today's sport analytics. One reason for this is the issue of rare goals as discussed in The Number's Game (Anderson & Sally, 2013) *"Goals really are rare and precious events: more than 30 per cent of matches end with one goal or none."* Without the inclusion of target labels to represent value which are needed for classification algorithms, it is very hard to relate which properties of the team such as formation and tactics relate to success (Lucey, et al., 2013).

One method of trying to simplify the game is by gridding the pitch into discrete areas. There has been some research done into the best way of splitting a pitch for analysis (Gudmundsson & Horton, n.d.). Some methods involve an equal splitting (Lucey, et al., 2013) whereas another common approach is to look at the player's dominant regions. "*The dominant region of a player p is the region of the playing area where p can arrive before any other player*." (Taki, et al., 1996). A method of gridding will need to be used to successfully include a spatial element to this research.

2.1.3 The use of Data in Premiership Football

Premier league football is a 90-minute game where two teams of 11 players compete to score goals, with the team scoring the highest number at the end of the game winning. The premier league gives three points for a win, 1 point for a draw and 0 points for a loss.

Charles Reep and Bernard Benjamin (Reep & Benjamin, 1968) were some of the first football analysts who began to shape the way football tactics were used to try and win games. (Hughes & Franks, 2005). Reep and Benjamin's statistical analysis of football matches between 1958 and 1967 looked at distributions of passing movements and concluded that although chance dominated the game, using smaller lengths of passing sequences resulted in more goals (Reep & Benjamin, 1968). Reep and Benjamin found that on average one goal was scored in every 10 shots (Reep & Benjamin, 1968). The results of this research are still influential in the modern game with teams trying to overcome the chance effect by using 'long-ball game' and 'direct-play' tactics to get the ball into shooting positions with as few passes as possible. The former tactic refers to sending long crosses from the defensive to offensive areas of the pitch with the aim of trying to catch the defence off guard. The latter refers to quick forward movements of play and the reduction of side and back passes. There were opposing views to this research, owing to evidence that possession football was in fact more successful, but the debate none the less had begun and analytics in football was born.

Opta provided data of the 2013-2014 season for this study. This was a season that saw Manchester City top the league with 27 wins, 5 draws and 6 losses and a final point tally of 86. There were 1052 goals, with Luis Suarez for Liverpool topping the goal-scoring table with 31 goals, 10 more than Liverpool's Daniel Sturridge in second place. It was an open season with the champion not decided until the final day. The first spot position changing hands 25 times (Wikipedia, n.d.), the most variety since the 2001-2002 season. Manchester City and Liverpool each scored more than 100 goals for the first time in league history (Wikipedia, n.d.).

2.1.4 Choosing a Player's Position

The competitive nature of Premiership football teams means that most of the reasoning behind choices and analytical information is undisclosed. This makes understanding the current state of the game at top levels difficult. Rare interviews with managers, coaches and players can provide some insight on the topic. An interview by the football magazine fourfourtwo quotes Arsene Wenger, arguably one of the most successful premiership managers, saying "*It's important you have the right players in the right positions*." (fourfourtwo, n.d.) . He suggests using personality as a method of choosing a player's optimum position "*I find that quiet, efficient guys make the best strikers*." (fourfourtwo, n.d.). However, with large amounts of

data becoming more readily available it seems only natural that this should be an aid in helping decipher where a player really performs best on the pitch. More recently Wenger was interviewed after moving Aaron Ramsey to a central midfield position. This is something the player himself had openly been pushing for. However, Wenger clearly stated that it was the data not Ramsey's persistence which got him the move: "*If you look at his Expected Goals when [Ramsey] is in a central position, it is among the best in the Premier League*" (Smith, 2015).

OptaPro (OptaPro, n.d.) host an annual event to try and bridge the gap between innovative analytical minds from outside clubs and professional minds inside clubs. One of the only events of its kind, OptaPro Analytics Forum produces interesting ideas and publicly available analysis. Club professionals choose the authors of the proposals they deem most relevant to speak at the forum. One of the successful speakers at the 2016 forum, Charles Neil, presented a tool that uses hierarchical clustering to group similar players (Charles, 2016) dependent on certain attributes e.g. position of the pass, passing accuracy, chances created. This tool was received well and is evidence that clubs do want further research into player characteristics (Charles, 2016).

2.2. Theory Surrounding Data Science Algorithms

2.2.1 Principal Component Analysis (PCA)

PCA is a method of dimensionality reduction that takes a number of related variables and orthogonally transforms them into a new smaller set of uncorrelated variables where the first few components try to retain as much of the information in the dataset as possible (Jolliffe, 2002). Principal Component 1 ($PC1$) ($\alpha_1^t x$) is a linear function which has maximum variance of the elements x (in this paper x refers to the log(average on-the-ball events per game) for each player), where $\propto_1$is a vector of p constants $a_{11}, a_{12}, \dots, a_{1p}$ and t denotes the transpose (Jolliffe, 2002). So that:

$$\alpha_1^t x = \propto_{11} x_1 + \propto_{12} x_2 + \cdots + \propto_{1P} x_P = \sum_{j=1}^{p} \alpha_{1j} x_j \qquad [1]$$

Then, PC2($\alpha_2^t x$) is found which has two key attributes; firstly it is uncorrelated with PC1 and second it also has maximum variance of the data. This continues until the z^{th} stage (PCZ) where $\alpha_z^t x$ has maximum variance and is uncorrelated with PC1, $PC2 \ldots . PC_{z-1}$. This can be visualised simply where $z = 2$.

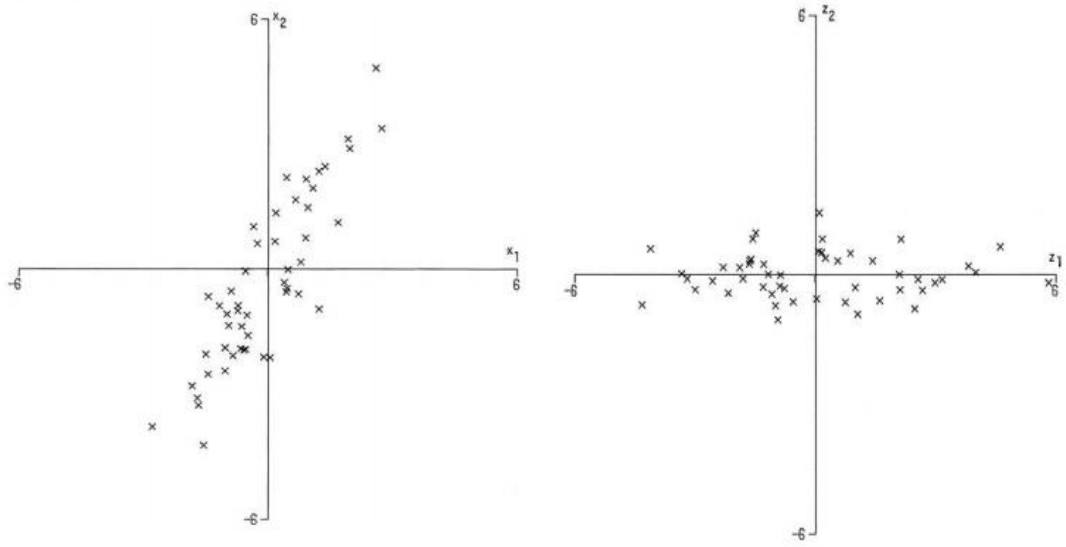


Figure 3 LHS shows variables x1, x2 RHS observations from LHS with respect to their PCs z1, z2

PCA aims to find the minimum sum of squared perpendicular distances from any input value to the principal components and therefore the inclusion of outliers in the dataset can have an effect on final result. It is usually advised to transform any highly skewed or long tailed data before performing the analysis so as not to distort results. This must be taken into account in this paper as the frequency of on-the-ball events may differ largely between different players.

2.2.2 Random Forests

Random Forests are a large ensemble of decision trees that can be used for classification or regression. They take the modal value of the output of the trees and remove the problem of overfitting (Breiman, 2001). Random Forests are a strong predictive tool, with their ability to find the importance of the input variables becoming a more recent popular asset of the model. (Louppe, et al., 2013).

The nature of this research aims to use the Importance Value outputs as Performance Indicators to describe the importance of variables. There are two main types of importance metrics

associated with Random Forests: the Mean Decrease Accuracy (MDA) and the Mean Decrease Impurity (MDI). The MDA uses out of bag (OOB) error calculation, with variables that reduce the error when included given higher importance (Louppe, et al., 2013). The MDI takes a weighted average of the impurity decreases at each split of a node on a variable. Equation [2] shows the importance of variable (X_m) for predicting Y. It is the average weighted impurity decrease for all the nodes t where X_m is used $(p(t)\Delta i(s_t, t))$ over all N_T trees in the forest (Louppe, et al., 2013). A widely used impurity index and the one used in this paper is the Gini index (denoted i in equation [2]).

$$Imp(X_m) = \frac{1}{N_T}\sum_T \sum_{\text{t}\in\text{T:v}(s_t)=\text{X}_m} p(t)\Delta i(s_t, t) \qquad [2]$$

A high decrease in Gini on a variable means that variable plays a large part in separating the data into distinct classes. In recent years there has been much debate on the bias of these importance measures. Strobl at al. stated that "*when a method is used for variable selection, rather than prediction only, it is particularly important that the value and interpretation of the variable importance measure actually depict the importance of the variable, and are not affected by any other characteristics*." (Strobl, et al., 2007). Their research uses simulated examples to show that variable importance is affected by an uneven number of categories for each predictor variable, as well as the scale of measurements of these values. A number of papers have been written in regards to solving this problem. Sandri et al. show that adding a set of artificial variables to the dataset can help estimate the bias (Sandri & Zuccolotto, 2008), while Clarke et al. address the bias due to grouping issues by developing a first-order correction term (Ourti & Clarke, 2009). To use this Gini method in this research these biases must be taken into account by ensuring the input categories of the Random Forest remain stable.

2.2.3 Visualisation of Multi-Dimensional Spatial Data

It was important to bear in mind the needs of the user of the research when interpreting results. This research is aimed at sporting professionals, as well as academics, and it was therefore important to ensure that the output reflected this requirement. This was made more difficult when high-dimensional datasets were used. For this reason heat maps were used extensively as a means of communicating with the intended audience. Heat maps are two-dimensional graphical representations of data and a popular way of representing large datasets. This is due

to their capacity to reduce large sets of data into a form that is more easily understood; the perceptive nature of their colour scale limits the work the user has to do and the spatial element relates to real life positioning which reduces the effort needed by the user to interpret areas of high and low activity (Bojko, 2009).

3 – Methods:

3.1. Data Gathering and Manipulation

The data for analysis has been provided by OptaPro (F24 package) (Opta, n.d.) in XML format. The dataset includes every on-the-ball event of the 2013-2014 English Premier League season, comprising 20 teams and 380 matches. The F24 feed listed all player action events within the game with player id, team id, event type, minute and second as well as a vast number of qualifiers describing each event. There were 65 events available for each timestamp and 229 qualifiers that could correspond to each event. The data was converted from XML into a SQL database using the R package *Mongodb* (Chheng, 2013). A SQL database was created for each match to increase the efficiency of importing data; one large database was created that merged all matches by *'*game_id*'*. Once converted every row of the dataset provided information on one 'event_id' and one 'qualifier_id', making the total size of the database for the full season 2,944,411 rows. The nature of the dataset meant it was important to create an index for easily searching *'*event_type_id' and *'*qualifier_id' to make sense of the results, Table 1 shows a sample of rows from each table[2]. The dataset includes **'**event_ids' and 'qualifier_ids'**.** 'Event_ids' explain on-the-ball events. 'Qualifier_ids' provide a more detailed understanding of the on-the-ball event; for example the 'event_id' may be a pass but within this there are 14 styles of passing such as a layoff or a chip.

On-the-ball event coordinates were provided and taken as 0-100 from left to right and 0-100 from top to bottom of the pitch. These coordinates were standardised so that attacking play is always from left to right, not dependent on team or period of the game.

A separate XML dataset provided information on the majority[3] of players involved in the season was also provided. This included information such as the player's name, positions, age, weight and height.

[2] Full tables available in F24: Feed appendices document.
[3] Not all players provided.

Event ID	Event Type	Event Description
1	Pass	Any pass attempted from one player to another free kicks, corners, throw ins, goal kicks and goal assists
2	Offside Pass	Attempted pass made to a player who is in an offside position
3	Take On	Attempted dribble past an opponent (excluding when qualifier 211 is present as this is overrun and is not always a duel
4	Foul	This event is shown when a foul is committed resulting in a free kick
5	Out	Shown each time the ball goes out of play for a throw-in or goal kick

Qualifier ID	Qualifier Type	Qualifier Value	Qualifier Description
1	Long ball		Long pass over 35 yards
2	Cross		A ball played in from wide areas into the box
3	Head pass		Pass made with a players head
4	Through ball		Ball played through for player making an attacking run to create a chance on goal
5	Free kick taken		Any free kick; direct or indirect
6	Corner taken		All corners. Look for qualifier 6 but excluding qualifier 2 for short corners
7	Players caught offside	Player ID	Player who was in an offside position when pass was made.

Table 1 Example of 'Event ID' (top) and 'Qualifier ID' (bottom) tables

3.2. Understanding the Characteristics of Different Footballer's Positions on the Pitch

Principal Component Analysis (PCA) was used to gain a broad understanding of whether players in different playing positions play different styles of football. The inputs were calculated as the average count over the season of each on-the-ball event for each player. For each player the average count would be as shown in Equation 3. Where, n = number of games and E_x = the count of each on-the-ball event type x. Therefore, for each player:

$$Average\ Count\ of\ Event\ x\ = \frac{1}{n}\sum_{i=1}^{n} E_{xi} \quad [3]$$

The events were filtered to include only those that related to on-the-ball events during the game. Metrics such as substitutions, weather events etc. were removed. The final 'event_id' table filtered to on-the-ball events can be seen in *Table 2*.

Event ID	Event Type	Event Description
1	Pass	Any pass attempted from one player to another; free kicks, corners, throw ins, goal kicks and goal assists
2	Offside Pass	Attempted pass made to a player who is in an offside position
3	Take On	Attempted dribble past an opponent
4	Foul	Shown when a foul is committed resulting in a free kick
5	Out	Shown each time the ball goes out of play for a throw-in or goal kick
6	Corner Awarded	Ball goes out of play for a corner kick
7	Tackle	Tackle = dispossesses an opponent of the ball - Outcome 1 = win & retain possession or out of play, 0 = win tackle but not possession
8	Interception	When a player intercepts any pass event between opposition players and prevents the ball reaching its target. Cannot be a clearance.
10	Save	Goalkeeper event: saving a shot on goal.
11	Claim	Goalkeeper event; catching a crossed ball
12	Clearance	Player under pressure hits the ball clear of the defensive zone
13	Miss	Any shot on goal which goes wide or over the goal
14	Post	Whenever the ball hits the frame of the goal
15	Attempt Saved	Shot saved - this event is for the player who made the shot.
16	Goal	All goals
17	Card	Bookings; will have red, yellow or 2nd yellow qualifier, plus a reason
42	Good Skill	A player shows a good piece of skill on the ball, such as a step over or turn on the ball
44	Aerial	Aerial duel 50/50 when the ball is in the air. Outcome will represent whether the duel was won or lost
45	Challenge	When a player fails to win the ball as an opponent successfully dribbles past them
49	Ball recovery	Team wins the possession of the ball and successfully keeps possession for at least two passes or an attacking play
50	Dispossessed	Player is successfully tackled and loses possession of the ball
55	Offside provoked	Awarded to last defender when an offside decision is given against an attacker
60	Chance missed	Used when a player does not actually make a shot on goal but was in a good position to score and only just missed receiving a pass
61	Ball touch	Used when a player makes a bad touch on the ball and loses possession. Outcome 1 Ð ball simply hit the player unintentionally. Outcome 0 Ð Player unsuccessfully controlled the ball.

Table 2 Table of events which relate to on-the-ball skill.

This data was transformed logarithmically to remove any skewed distributions before running the PCA. This analysis was run using the *prcomp* (3.4.0, 2016) package from the stats library in R alongside *ggbiplot* (Vu, 2011) for visualization of the output. The *prcomp* package provides output including: loadings of the variables, standard deviation of the components and value of the rotated data (centred data multiplied by the rotation matrix) (3.4.0, 2016). The value of rotated data for PC1 and PC2 (the variables with the highest variance) were then used in the *ggbiplot* function with each point representing a player and each colour representing their playing position. This use of dimensionality reduction alongside visualisations allowed

any patterns in the data to be seen by the appearance of clusters.

Pass _Long ball	Pass _Assist	Foul _Foul	Clearance _Blocked cross
Pass _Cross	Pass _2nd assist	Foul _Direct	Miss _Head
Pass _Head pass	Pass _In-swinger	Foul _GK Y Coordinate	Miss _Right footed
Pass _Through ball	Pass _Out-swinger	Foul _Defensive	Miss _Regular play
Pass _Free kick taken	Pass _Straight	Foul _Offensive	Miss _Fast break
Pass _Corner taken	Pass _GK Y Coordinate	Out _Player not visible	Miss _Set piece
Pass _Regular play	Pass _Offensive	Out _GK Y Coordinate	Miss _From corner
Pass _Fast break	Offside Pass _Long ball	Corner Awarded _GK Y Coordinate	Miss _Free kick
Pass _Set piece	Offside Pass _Cross	Tackle _Last line	Miss _Assisted
Pass _From corner	Offside Pass _Head pass	Tackle _Out of play	Miss _Strong
Pass _Throw-in	Offside Pass _Through ball	Tackle _GK Y Coordinate	Miss _Weak
Pass _Direct	Offside Pass _Players caught offside	Tackle _Defensive	Miss _Swerve Left
Pass _Intentional assist	Take On _Fast break	Tackle _Offensive	Miss _Swerve Right
Pass _Chipped	Take On _Overrun	Interception _Last line	Miss _Not past goal line
Pass _Lay-off	Take On _GK Y Coordinate	Interception _Head	Miss _Intentional assist
Pass _Launch	Take On _Defensive	Save _Last line	Miss _Big Chance
Pass _Flick-on	Take On _Offensive	Save _Def block	Miss _Individual Play
Pass _Player not visible	Foul _Penalty	Save _GK Y Coordinate	Post _Right footed
Pass _ Pull Back	Foul _Hand	Clearance _Head	Post _Regular play
Pass _Switch of play	Foul _Dangerous play	Clearance _Out of play	Post _Assisted
Post _Swerve Left	Attempt Saved _Big Chance	Goal _GK Y Coordinate	
Post _Swerve Right	Attempt Saved _Individual Play	Card _Foul	
Post _Intentional assist	Attempt Saved _GK Y Coordinate	Aerial _GK Y Coordinate	
Post _Individual Play	Goal _Penalty	Aerial _Defensive	
Post _GK X Coordinate	Goal _Head	Aerial _Offensive	
Attempt Saved _Head	Goal _Right footed	Challenge _GK Y Coordinate	
Attempt Saved _Right footed	Goal _Regular play	Challenge _Defensive	
Attempt Saved _Regular play	Goal _From corner	Challenge _Offensive	
Attempt Saved _Fast break	Goal _Free kick	Dispossessed _GK Y Coordinate	
Attempt Saved _Set piece	Goal _Assisted	Dispossessed _Defensive	
Attempt Saved _From corner	Goal _1 on 1	Dispossessed _Offensive	
Attempt Saved _Free kick	Goal _Strong	Chance missed _Assisted	
Attempt Saved _Assisted	Goal _Swerve Right	Chance missed _Intentional assist	
Attempt Saved _Strong	Goal _Deflection	Ball touch _Own shot blocked	
Attempt Saved _Weak	Goal _Keeper Touched		
Attempt Saved _Swerve Left	Goal _Intentional assist		
Attempt Saved _Swerve Right	Goal _Big Chance		
Attempt Saved _Deflection	Goal _Individual Play		
Attempt Saved _Hit Woodwork	Goal _ 2nd assisted		
Attempt Saved _Intentional assist	Goal _GK X Coordinate		

Table 3 All combinations of 'event_id' and 'qualifier_id' separated by '_'

The dataset was expanded to include 'qualifier_ids' to understand more fully the patterns in playing positions and styles of play. From here onward 'event_ids' that are combined with 'qualifier_ids' will be referred to as 'event_qualifier_id'. For example, if 'event_id' is a pass and 'qualifier_id' is a long ball the 'event_qualifier_id' will be a 'pass_longball'. The 'event_qualifier_ids' included can be seen in Table 3. There were 134 combinations of 'event_qualifier_id'.

Since a football match is a time series of spatial events it seemed important to include variables for the time the on-the-ball event took place during the game and where the event happened on the pitch. Since the introduction of player coordinate data, spatio-temporal analysis of the game of football has increased. However, since the flow of play means any player can go anywhere, for ease of analysis, it is necessary to transform these continuous coordinate values into discrete measures by gridding the pitch. Historically this has been done in many different shapes and sizes (Taki, et al., 1996). The dataset used in this paper is provided with a 'zone' metric which groups events into four main areas: back, centre, left and right. It was used for initial gridding and is shown in Figure 4.

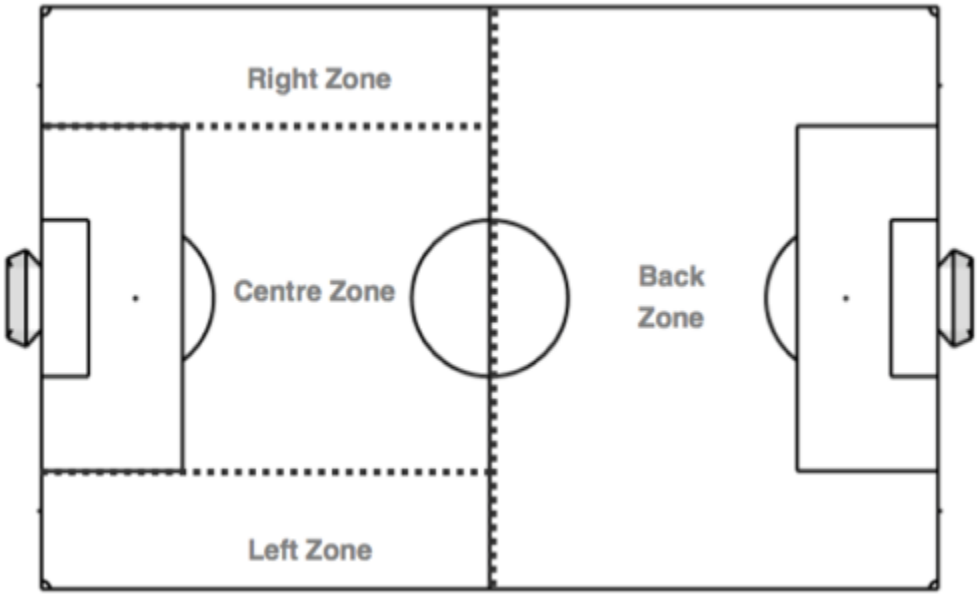


Figure 4 Zones of the pitch taken from Opta F24: Feed appendices document.

PCA was run separately for every 10 minutes of the match in these four main areas in order to provide a broad view of whether time and space were important in understanding a player's position.

3.3. Finding the On-The-Ball Events that are most Important in Defining a Player's Position.

With a general understanding of how key skills vary across playing positions it was then

important to understand which of these 'event_qualifier_ids' contained the most information relative to playing position. A Random Forest was used with the aim of filtering the 134 'event_qualifier_ids' into a smaller more manageable set. Due to findings explained in section 4.1. it was deemed adequate to run this model for the four different areas of the pitch but not for segmented times. This Random Forest model predicted the position of the player using the average count of completed 'event_qualifier_id' combinations. However, rather than use the algorithm for the more normal method of prediction it was instead used to find the importance of each of the 'event_qualifier_id' variables. This meant using a Mean Decreasing Impurity (MDI) calculation where the variables that contain the most information were given the highest weights. This importance variable was calculated by taking the average MDI across 'event_qualifier_id' for each area of the pitch. The 'event_qualifier_id' combinations with the highest importance were then used in future stages of analysis. This allowed events which had minimal importance when choosing players' positions to be discarded, reducing the dataset to a more manageable size.

3.4. Finding the On-The-Ball Events that are Most Important Relative to Success.

Up to this point all chosen 'event_qualifier_ids' had been decided dependent on a player's position. To try and assess a player's skillset and eventually find their optimal position it was important to understand what makes a player successful. Quantifying success in football is difficult, particularly when the decision making of managers is often informed by their own subjective criteria. This paper hopes to achieve a quantitative understanding of what makes one player perform better in a position than another. The difficulty with Performance Metrics in football is that the true decider of winning or losing is the number of goals scored and this value is very small relative to the number of games played[4] and the number of players on the pitch. It is also highly skewed towards strikers. To try and more evenly rank players this paper explores the idea of path analysis by converting the dataset into paths of 'event_qualifier_ids' in the hope of attributing value to the players that participated in the lead up to goals.

To understand what events lead to success the Opta 2013-2014 dataset was split into path level

[4] Average of 2.77 goals per match in the 2013-2014 season.

data using the filtered 'event_qualifier_ids' determined to be most important from the results in section 3.3. The path level data was initially created by taking every 30-second interval of each game and the previous 10 events that led to that point grouped by team. If the dataset was not split by teams it would be difficult to differentiate which team to award the value to. If there was a goal or a shot on target within this path, then it was labelled 1 and the path cut short at the time of this event. If there were no shots or goals, then the path was labelled 0. Due to the rarity of goals in football, the measure of success also included shots to try and increase the number of positive outcomes that went into the model. On average there were 1.5 events per second, which was the reason for choosing 10 events per 30-second interval as every event would be accounted for in at least one path.

An example of these paths can be seen in Table 4. This shows two timestamps (x and y) of a game for the two teams (1 and 2) competing. In the lead up to timestamp x, team 1 and team 2 both completed 10 events with neither producing a goal or shot in this time. In the lead up to timestamp y, team 2 again completed 10 events without a shot or goal whereas team 1 scored on their fourth event. This goal is labelled 1 in the last column and the path is cut short. This was to ensure that any events that happened after this '*success'* would not be attributed any value. This method was chosen to try and overcome the problem of goal scoring being the only quantifiable measure of success since the lead window allows not just shots/goals but all the events to have the potential to be attributed positive value.

Team	Timestamp	Event 1	Event 2	Event 3	Event 4	Event 5	Event 6	Event 7	Event 8	Event 9	Event 10	Goal =1, No Goal =0
1	x	Pass_Long ball	Pass_Throw-in	Take On	Pass_Long ball	Pass_Flick-on	Pass_In-swinger	Aerial	Dispossessed	Dispossessed	Ball Recovery	0
2	x	Pass_Throw-in	Challenge	Dispossessed	Ball Touch	Ball Touch	Pass_Straight	Take On	Take On	Pass_Cross	Pass_Cross	0
1	y	Pass_Throw-in	Take On	Take On	Goal							1
2	y	Pass_In-swinger	Aerial	Dispossessed	Dispossessed	Ball Recovery	Pass_Long ball	Pass_Throw-in	Take On	Pass_Long ball	Pass_Cross	0

Table 4 Example of constructing path-level data

To avoid any biases when calculating the importance of variables using the MDI method the paths were converted to binary with a 1 if the event happened in that path and a 0 otherwise,

and the target variable of 1 or 0 if there was a shot or goal or not. Although this removed the information surrounding the position of each 'event_qualifier_id' in the path this was considered to be acceptable considering that the duplicative nature of taking every 30 seconds produced an overlap of events which should account for this. For example, if an 'event_qualifier_id' appeared at position eight when a goal was scored in position ten for one 30 second interval and the goal was captured again in the next thirty second interval at position 9, the 'event_qualifier_id' would be recorded again at position seven increasing its value relative to shots and goals.

Once the importance of including spatial data was ascertained (as described in the results section 4.1.), the paths were broken down from the four zones provided in the dataset to 100 segments of 10x10 coordinates as shown in the bottom image in Figure 5. The (10,10) coordinate refers to the area of the pitch from 0-10 for the x-axis and 0-10 for the y-axis. Each event in the dataset has x and y coordinate values related to the positions of x and y described in the top image in Figure 5.

An example of the paths can be seen in Table 5. This looks at the (40,50) segment of the pitch with binary labelling of a subset of 'event_qualifier_ids' that happened in the lead up to timestamp x and y for teams 1 and 2.

Path level data was created for all teams for all matches of the season and used as the input for the Random Forest models. For this research Random Forests were not used for their predictive capabilities but instead to find the importance of the input variables relative to the target variables. For each 10x10 area of the pitch one Random Forest containing Importance Values for each 'event_qualifier_id' was created. The MDI method was chosen to calculate these importance variables for the computational speed that results when the Random Forest decides which node to split on.

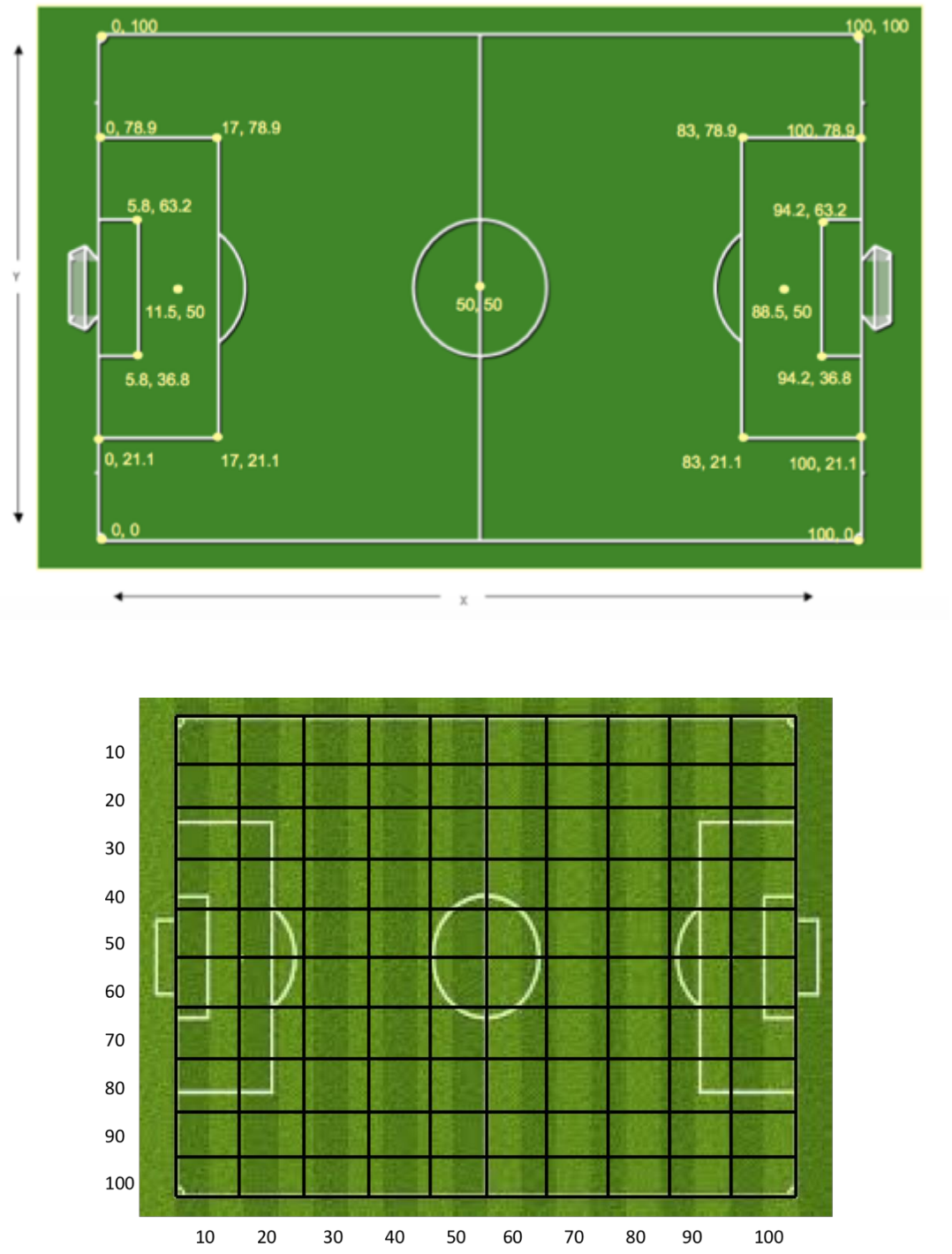


Figure 5 Diagrams showing the x and y coordinate values relative to areas on the pitch. Overall (top) and broken into 100 segments (bottom)

Due to the difference in magnitude of the Importance Values for each 'event_qualifier_id', and for each coordinate section of the grid, it was necessary to transform the values to the same scale so they could be compared in future stages. The values were standardised to between 0

and 1 for each 10x10 area of the pitch. So events with the most importance were closer to 1 and least importance 0. The importance of most events relative to scoring goals heavily increased when closer to the goal box, which produced a skew in the results. Removing this logarithmically was tested but it was decided that this skew was, in fact, important for understanding the importance of variables and therefore kept. The result of standardising these Importance Values was a Performance Indicator for each segment of the pitch and each 'event_qualifier_id'.

Team	Timestamp	X Coord	Y Coord	Pass_Long ball	Pass_Throw-in	Pass_Chipped	Pass_Lay-off	Pass_Launch	Pass_Flick-on	Pass_Pull Back	Pass_Switch of play	Pass_Cross	Pass_Assist	Goal =1, No Goal =0
1	x	40	50	1	0	0	0	1	0	0	0	0	0	0
2	x	40	50	0	0	1	0	0	0	0	0	0	0	1
1	y	40	50	1	1	0	0	1	0	0	0	0	0	0
2	y	40	50	1	0	0	1	1	0	0	0	1	0	0

Table 5 Example of binary data used as input to Random Forest.

3.5. Finding a Player's Optimum Position on the Pitch.

To relate the Performance Indicator to each player a multiplier of the player's performance was used; this methodology is discussed in detail in the next section. To visualise these results a heat map for each player showing each area of the pitch and their relative Performance Indicators was created to illustrate where they performed best. On first inspection it soon became clear that this initial model was too heavily weighted towards attack and gave much higher value to events in the (80,30) – (100,70) coordinate areas. To overcome these issues two solutions were tried.

1. **Include a defensive measure**. Running a new model with the same method as the first but add ball recovery to goals and shots as a target variable. Ball recovery was chosen because it was a common on-the-ball event and a good signal of a strong defensive player.

2. **Use longer path lengths in order to include more on-the-ball events in the lead up to goals**. Although the previous 10 on-the-ball events captured all the events in the game, it gave less emphasis to the beginning of the play that may have led to a goal. Increasing this to the previous 60 events insured that any event that may have caused the goal further down the pitch would also be accounted for.

3.5.1 The Frequency Problem

The creation of this new 3060 model vastly improved the Performance Indicator's versatility over the different areas of the pitch. However, the next step and a major challenge faced in this research, was trying to understand how best to relate the Performance Indicator back to the individual players. To do this it was first important to decide whether the aim of the research was to investigate the quantity or the quality of play. As the Performance Indicator's were comprised of standardised values between 0 and 2 (a summation of the normalised ball recovery model and the goals/shots model) a standard calculation of the number of times a player carried out an on-the-ball event in each segment of the pitch multiplied by the Performance Indicator would be heavily weighted to where the player had most possession removing any added value of the Performance Indicator. Similar to the tf-idf (term frequency – inverse document frequency) problem incurred in information retrieval it was necessary to find a way to scale performance and frequency. From a footballing perspective it was important not only to give value to the number of times a player completed an on-the-ball event, but also to what on-the-ball event they completed. With this view in mind, a player that completes 100 bad on-the-ball events should not be rated higher than a player that completed five good on-the-ball events. Similarly a player that only touched the ball twice should not have the same value as a player that made himself available for every ball. Calculating the average number of on-the-ball events for each position showed that although there was some discrepancy between the position with the most and least possession it was not a substantial difference. As would be expected if goalkeepers are excluded, it is the midfielders that have the most posession and the strikers the least.

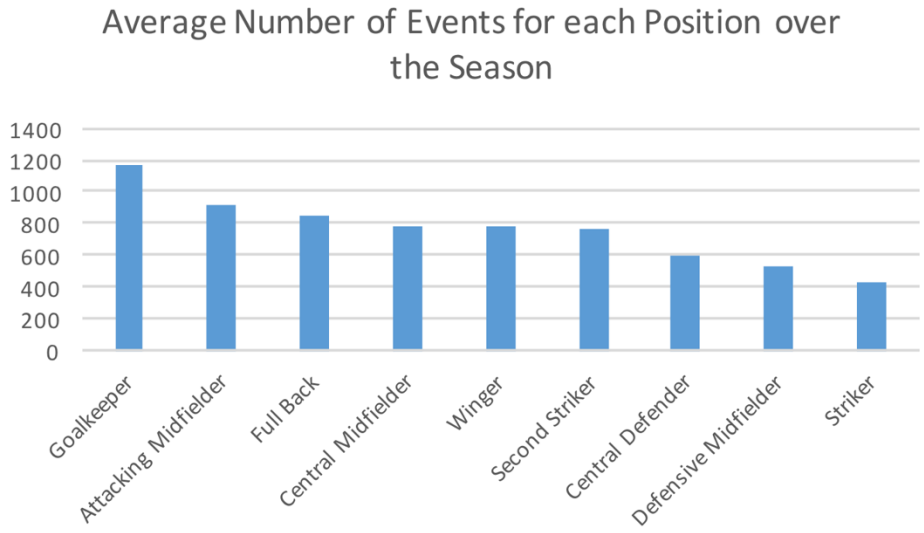


Figure 6 Average number of events completed in each position over the 2013-2014 season

Initially, to solve this problem the idea was to create the Player Metric as shown in Equation

$$Player\ Metric_{(x,y)} = \frac{E_{n(x,y)}}{\sum E_{(x,y)}} \quad [4]$$

4, where $E_{n(x,y)}$ is the number of times a player does Event n in the x and y coordinate area of the pitch. This showed a ratio of the number of times they carried out an on-the-ball event relative to the total number of on-the-ball events in each segment. Summing these Player Metrics for each on-the-ball event carried out in that area of the pitch and multiplying by the

$$Initial\ Performance\ Metric_{(x,y)} = \sum_{n=1}^{n=t} \left[\frac{E_{n(x,y)}}{\sum E_{(x,y)}} * (PI.E_{n(x,y)})\right] \quad [5]$$

Performance Indicator (PI) weighted the Performance Metric to the areas of the pitch the player showed the most ability (shown in Equation 5). However, after research into this function it appeared that areas of play that were less frequently reached were given too high an importance by the model, showing defenders as more capable in the offensive areas and attackers more capable defensively. To resolve this, it seemed important to add a scaling factor which included some value to the frequency of on-the-ball events without totally overriding

$$Final\ Performance\ Metric_{x,y} = \left[\sum_{n=1}^{n=t} \left(\frac{E_{n(x,y)}}{\sum E_{(x,y)}} * PI.E_{n_{x,y}}\right)\right] . \left[\frac{\log(E_{n_{x,y}})}{4}\right] \quad [6]$$

the Performance Indicator. Logarithmic and square root functions were experimented with and finally $f(\frac{\log(total\ events\ per\ coord)}{4})$ was chosen as a suitable scaling factor. To summarise where the player had been playing most effectively relative to ball recovery, shots and goals, all of the values for each area were summed to give one overall value per area, where (x, y) are coordinate values between 0 and 100 grouped by 10.

Figure 7 shows an example calculation with the blue squares representing the Player Metric values of three coordinate areas of the pitch (40,30) (40,40) and (50,30) and the green squares the same area of the pitch with the Performance Indicator values.

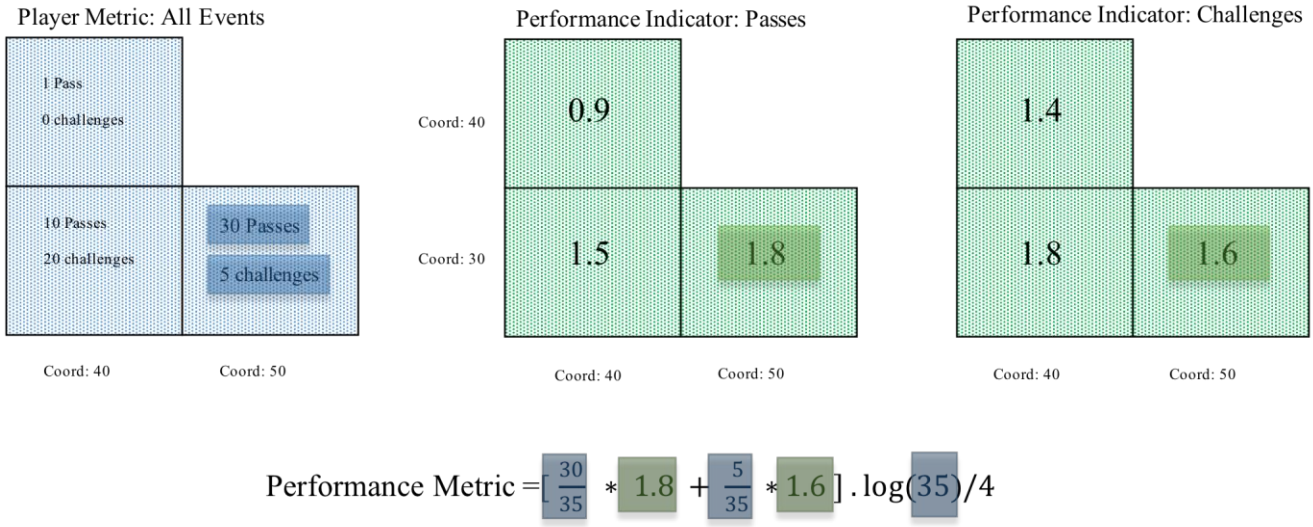


Figure 7 Example of Performance Metric calculation for (50,30) segment of the pitch.

3.6. Training/Validation/Testing and Evaluation

To ensure that the model was not overly influenced by the chosen time interval and previous on-the-ball event variables, a test for robustness of results was completed. This was done by running models for 30-second and 60-second intervals for the previous 10, 20, 30, 40, 50 and 60 on-the-ball events. Each model was run for each 10x10 segment of the pitch using 10-fold cross validation. Cross validation trains an algorithm on one part of the data and validates on the rest. Typically, 10-fold cross validation uses a 90/10 split so 90% of the data was used for training and 10% for validation where each fold uses a different 10%. This meant for every combination of seconds and previous events (12), each area of the pitch (100) and all cross

validations (10) there were in total 12,000 models run.

3.7. Model Implementation

Due to the size of the computations and number of models run to test the robustness two virtual machines of 4 CPUs and 15GB each were created to run the models using the Google Cloud Platform (Google, n.d.). R was installed on the virtual machines along with the appropriate packages and the SQL databases for each path level dataset were imported. This meant that four models could be run in parallel cutting down the computational time drastically. The time the models took to run ranged from 6-12 hours dependent on the path length of the data. For example, the 3060 model was far more computationally expensive than the 6030 model. The Random Forest outputs for each model were saved to the virtual machine and then transferred back to the local machine for analysis.

The first two objectives took approximately two months to complete from the beginning of July to the end of August 2016. This was in parallel to continuously exploring and understanding the dataset. The second two objectives took slightly longer as some methods had to be trialled and errored numerous times. They were completed between September and November 2016. The write up of this project was done throughout all stages with the main bulk completed towards the end of November and December 2016.

3.8. Dashboard Visualisation

Due to the number of football teams, players and the nature of the research it was deemed necessary to create an interactive dashboard to display the results clearly. The dashboard presented each player's individual on-the-ball event count on each area of the pitch compared to his calculated Performance Metric for the season. This allowed comparison of where the players actually played against where the model would have played them. There was also a filter to allow the break down of the Performance Metric on a match-by-match basis. A table showing the overall Performance Metric values for the team for the total season of matches chosen was also provided. These are summations of the Performance Metrics over all areas of the pitch. On a separate tab it was possible to view the Performance Indicator for each event to gain an overview of which events are more highly valued on which areas of the pitch.

Figure 8 Example of interactive dashboard used to view results. The top screen shot shows the Performance Metric for each player. The bottom screen shot shows the Performance Indicator for each on-the-ball event.

4 – Results:

4.1. Characteristics of Different Footballers' Positions on the Football Pitch

The objective of this section was to understand the type of on-the-ball events that define each position using PCA and clustering techniques on the average number of each on-the-ball event completed by a player over the season.

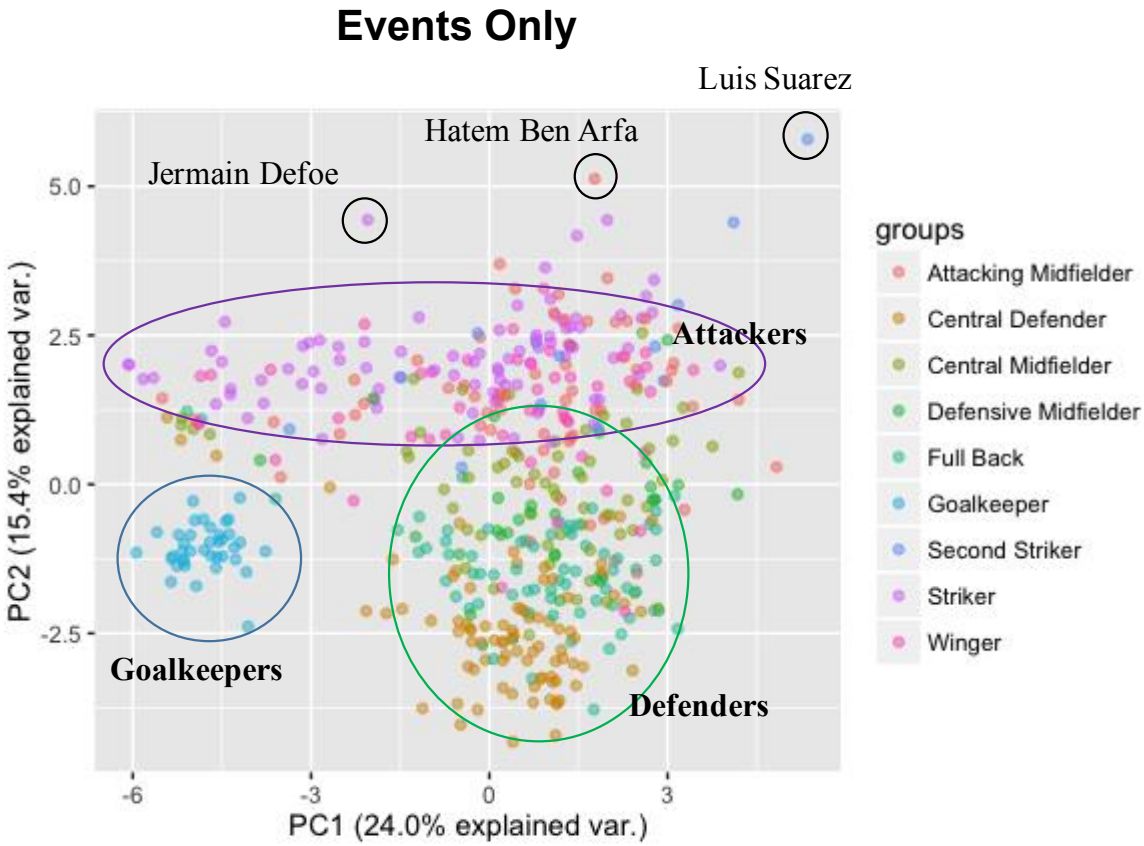


Figure 9 PCA output for all on-the-ball 'event_ids'. Each point represents a player and each colour represents a different playing position.

Unsurprisingly the least variance, and therefore the most obvious positional cluster, was seen for goalkeepers. This was shown by the tight clustering of blue points in the bottom left corner of Figure 9. This makes sense due to the distinctive qualities of goalkeeping activity during a

match and the inclusion of some on-the-ball events in *Table 2* that are goalkeeper specific. A pattern also emerged in the difference between offensive and defensive play. In general, PC1 was a more defensive component, hence the variance among the defensive players was less, shown in Figure 9 by the range of most defensive players lying between -2 and 2 on the PC1 axis. PC2 was a more attacking component and therefore higher variance was shown amongst the defensive players (green circle in Figure 9). Similarly, variance for offensive players was less for PC1 and higher for PC2 (purple circle in Figure 9). The other positions were grouped bearing some correlation to where the players were situated on the pitch, with central midfielders behaving most like defenders and least like strikers. However, attacking midfielders and wingers behaved more similarly to strikers. This suggests that these key clusters were dependent on the designated position of the player.

Zone	Events			
	PC1 - 24%		PC2 - 15%	
1	Ball recovery	7.3%	Clearance	9.6%
2	Foul	7.3%	Offside provoked	7.5%
3	Tackle	6.7%	Dispossessed	7.1%
4	Out	6.6%	Interception	6.8%
5	Take On	5.8%	Attempt Saved	6.8%
6	Ball touch	5.7%	Ball touch	6.6%
7	Challenge	5.7%	Take On	6.1%
8	Pass	5.6%	Pass	5.8%
9	Interception	5.4%	Save	5.6%
10	Dispossessed	5.2%	Miss	5.1%

Defensive Event	Offensive Event	Neutral

Table 6 Top 10 loadings shown as a percentage of total loadings for PC1 and PC2.

Table 6 breaks down the events by loading. This shows that ball recovery (team wins the possession of the ball and successfully keeps possession for at least two passes or an attacking play) has the highest loading in PC1 and clearances (player under pressure hits the ball clear of the defensive zone or/and out of play) the highest PC2 loading. This begins to show which type of event define which type of positions. From the clustering visualisation it is known that the events shown with the highest loadings in PC1 are the events which defenders behave most

similarly in completing. This corresponds with the red cells in Table 6 which represent defensive events. Although there are some defensive events in the highest PC2 loadings overall this again agrees with the clustering and is weighted towards offense.

As well as broader positional patterns the output of this analysis also shows some interesting insights at a player level. For example, Luis Suarez (the top goal scorer of the season) behaves differently, not only to the players who share his position but to all players during the season. These initial findings suggest there are definite patterns in the broad position of a player and their average number of 'event_ids' per game with further exploration needed.

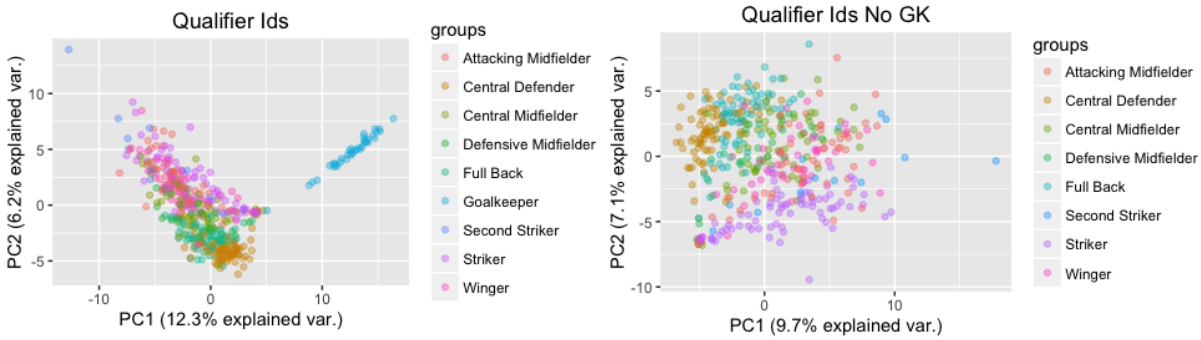


Figure 10 PCA output for all the on-the-ball 'event_qualifier_ids'. Each point represents a player and each colour represents a different playing position. LHS: includes goalkeeper, RHS: removes goalkeeper

Delving deeper into the data to include 'qualifier_ids' provided greater insight into the type of events that characterise a player's position. Due to the increase in variables the amount of variance covered by the top two principal components slightly decreased, but as expected the goalkeepers still created the most unique cluster.

Things became clearer and more interesting when goalkeepers were removed from the analysis. Table 7 shows a clear offensive defensive split between PC1 and PC2. The right-hand side chart in Figure 10 visualises this with attacking players (purple and red colour points) showing a a greater variance of skills relative to PC1 and defensive players (green and blue coloured points) showing a greater variance for PC2. Interestingly the model picks up the

'pass_long ball' (long pass over 35 yards) as behaving the most differently to the 'pass_lay off' (pass where player laid the ball into the path of a teammates' run) by grouping the 'pass_long ball' with the defensive 'event_qualifier_ids'. This could suggest the ineffectiveness of the long ball as an attacking move, instead indicating more behavioural likeness to a clearance or some type of defensive measure.

Zone	Events			
	PC1 - 9.7%		PC2 - 7.1%	
1	Pass _Lay-off	2.2%	Pass _Long ball	3.1%
2	Attempt Saved _Regular play	2.2%	ball_recovery	2.7%
3	Take On _Offensive	2.1%	Pass _Chipped	2.6%
4	Dispossessed _Offensive	2.1%	Tackle _Defensive	2.6%
5	Pass _Assist	2.1%	Pass _Launch	2.3%
6	Attempt Saved _Assisted	2.0%	Pass _Free kick taken	2.2%
7	Attempt Saved _Individual Play	2.0%	Pass _Direct	2.1%
8	Foul _Offensive	2.0%	Foul _Defensive	2.1%
9	Miss _Regular play	1.8%	Aerial _Defensive	2.0%
10	Attempt Saved _Right footed	1.8%	Clearance _Out of play	1.9%

Defensive Event	Offensive Event	Neutral

Table 7 Top 10 loadings shown as percentage of total loadings for PC1 and PC2 of 'event_qualifier_ids' without goalkeepers.

Since the game of football is a time series of spatial events it seemed necessary to break down the individual games into the time and area of the pitch where the events took place. The PCA was re-run for every 10-minute interval of the game using the pre-defined Opta segments of the pitch (left, right, back, centre) to understand if the addition of these variables affected the patterns in how player's in certain positions performed. Splitting the average on-the-ball events per player for every ten minutes of the game showed that in general there was minimal variation from one time segment to the next. The top loaded events were skewed towards offensive play as can be seen in Table 8. These were mainly shots, misses and goals with only a few defensive on-the-ball events (fouls and defensive tackles) appearing. Figure 11 verifies this with most of the information contained between -10 and 10 of both the x and y axis with approximately 6% of information contained in PC1 and 5% in PC2 for all time segments. There were some interesting outliers highlighted in Figure 11 of individual players who performed differently to the rest of their positional members. This was due to unique combinations of

events for those players at that time of the match that may indicate an optimal time to play them. Most of the outliers appeared for players who play striking positions. Further investigation into these results suggested that the number of goals scored was a large factor in producing these anomalies because they are such a rare event.

Removing the outliers as seen in Figure 12 reiterates that there were minimal changes to the clusters produced when incorporating a time variable and therefore it was not deemed an important property in understanding players and their positional attributes.

Time	0-10 Mins				10-20 Mins			
	PC1 - 6.3%		PC2 - 5.9%		PC1 - 6.9%		PC2 - 5.3%	
1	Foul _Direct	6.0%	Attempt Saved _Regular play	5.6%	Goal _Regular play	13.2%	Foul _Direct	10.3%
2	Foul _Foul	5.9%	Attempt Saved _Assisted	5.5%	Goal _Big Chance	12.9%	Foul _Foul	10.1%
3	Take On _Offensive	3.8%	Goal _Regular play	4.8%	Goal _Right footed	10.5%	Foul _Offensive	6.7%
4	Attempt Saved _Regular play	3.8%	Goal _Individual Play	4.8%	Goal _Assisted	8.0%	Foul _Defensive	4.6%
5	Attempt Saved _Assisted	3.7%	Foul _Foul	3.9%	Goal _Intentional assist	7.7%	Pass _Flick-on	3.5%
6	Foul _Offensive	3.6%	Attempt Saved _Individual Play	3.9%	Attempt Saved _Regular play	3.4%	Aerial _Offensive	3.5%
7	Foul _Defensive	3.3%	Foul _Direct	3.9%	Attempt Saved _Right footed	3.1%	Pass _Throw-in	2.8%
8	Attempt Saved _Right footed	3.2%	Attempt Saved _Right footed	3.6%	Take On _Offensive	2.7%	Attempt Saved _Regular play	2.6%
9	Miss _Intentional assist	2.9%	Attempt Saved _Intentional assist	3.5%	Attempt Saved _Assisted	2.5%	Dispossessed _Offensive	2.4%
10	Clearance _Head	2.8%	Pass _Long ball	3.4%	Pass _Long ball	2.2%	Pass _Lay-off	2.4%
Time	20-30 Mins				30-40 Mins			
	PC1 - 6.9%		PC2 - 6.5%		PC1 - 7.6%		PC2 - 6.5%	
1	Goal _Assisted	14.4%	Miss _Regular play	4.6%	Goal _Assisted	15.9%	Foul _Direct	6.1%
2	Goal _Intentional assist	14.4%	Miss _Assisted	4.5%	Goal _1 on 1	15.9%	Foul _Foul	5.9%
3	Pass _Flick-on	6.1%	Miss _Right footed	4.4%	Goal _Intentional assist	15.9%	Attempt Saved _Regular play	5.8%
4	Aerial _Offensive	5.8%	Attempt Saved _Assisted	4.0%	Goal _Big Chance	15.9%	Attempt Saved _Assisted	5.6%
5	Pass _Head pass	3.5%	Attempt Saved _Individual Play	4.0%	Tackle _Defensive	5.3%	Attempt Saved _Right footed	5.3%
6	Pass _Long ball	3.1%	Attempt Saved _Right footed	3.6%	Miss _Regular play	2.4%	Foul _Offensive	4.0%
7	Take On _Offensive	2.5%	Attempt Saved _Regular play	3.5%	Foul _Direct	1.7%	Dispossessed _Offensive	3.9%
8	Attempt Saved _Regular play	2.4%	Miss _Individual Play	3.4%	Foul _Foul	1.7%	Attempt Saved _Individual Play	3.9%
9	Tackle _Defensive	2.4%	Foul _Foul	3.3%	Attempt Saved _Regular play	1.7%	Foul _Defensive	3.8%
10	Attempt Saved _Assisted	2.3%	Attempt Saved _Intentional assist	3.3%	Attempt Saved _Assisted	1.6%	Miss _Assisted	3.1%
Time	40-50 Mins				50-60 Mins			
	PC1 - 7.0%		PC2 - 5.3%		PC1 - 5.8%		PC2 - 5.2%	
1	Attempt Saved _Assisted	4.4%	Foul _Direct	24.8%	Foul _Direct	5.8%	Post _Swerve Right	9.4%
2	Goal _Right footed	4.1%	Foul _Foul	24.8%	Foul _Foul	5.7%	Post _Individual Play	9.4%
3	Attempt Saved _Regular play	4.1%	Pass _Cross	20.7%	Attempt Saved _Zone	4.4%	Foul _Direct	5.5%
4	Goal _Big Chance	3.8%	Pass _Chipped	19.0%	Goal _Regular play	4.4%	Foul _Foul	5.4%
5	Goal _From corner	3.7%	Pass _Corner taken	17.2%	Foul _Offensive	4.3%	Pass _Corner taken	4.8%
6	Attempt Saved _Individual Play	3.6%	Foul _Offensive	15.5%	Attempt Saved _Regular play	4.1%	Attempt Saved _Zone	3.1%
7	Attempt Saved _Big Chance	3.6%	Take On _Offensive	15.4%	Attempt Saved _Right footed	3.9%	Attempt Saved _Regular play	3.0%
8	Attempt Saved _Right footed	3.5%	Pass _In-swinger	14.4%	Take On _Offensive	3.7%	Foul _Defensive	3.0%
9	Attempt Saved _Intentional assist	3.3%	Challenge _Defensive	12.6%	Attempt Saved _Assisted	3.7%	Aerial _Offensive	2.9%
10	Goal _Regular play	2.6%	Foul _Defensive	12.5%	Dispossessed _Offensive	3.4%	Pass _Cross	2.9%
Time	60-70 Mins				70-80 Mins			
	PC1 - 6.9%		PC2 - 5.1%		PC1 - 6.5		PC2 - 5.7%	
1	Goal _Regular play	9.6%	Foul _Direct	6.8%	Foul _Direct	26.3%	Foul _Direct	7.3%
2	Goal _Assisted	8.6%	Foul _Foul	6.6%	Foul _Foul	26.1%	Foul _Foul	6.8%
3	Goal _Intentional assist	8.6%	Miss _Assisted	4.8%	Attempt Saved _Regular play	20.7%	Foul _Offensive	5.6%
4	Goal _Big Chance	8.6%	Foul _Offensive	4.3%	Dispossessed _Offensive	20.5%	Attempt Saved _Assisted	5.5%
5	Attempt Saved _Regular play	5.5%	Miss _Regular play	4.1%	Attempt Saved _Assisted	20.3%	Attempt Saved _Regular play	5.5%
6	Attempt Saved _Assisted	4.4%	Foul _Defensive	4.0%	Foul _Offensive	18.3%	Attempt Saved _Individual Play	3.9%
7	Pass _Out-swinger	4.0%	Miss _Right footed	3.9%	Take On _Offensive	15.7%	Foul _Defensive	3.3%
8	Attempt Saved _Right footed	3.6%	Save _Def block	3.8%	Attempt Saved _Intentional assist	15.5%	Attempt Saved _Right footed	3.1%
9	Pass _Corner taken	3.6%	Miss _Intentional assist	2.6%	Attempt Saved _Individual Play	13.8%	Attempt Saved _Intentional assist	2.9%
10	Pass _Assist	3.0%	Pass _Head pass	2.5%	Foul _Defensive	13.8%	Pass _Assist	2.8%
Time	80-90 Mins				90-100 Mins			
	PC1 - 6.2%		PC2 - 6.1%		PC1 - 7.3%		PC2 - 6.2%	
1	Foul _Direct	5.7%	Foul _Foul	6.2%	Attempt Saved _Regular play	12.2%	Foul _Direct	14.7%
2	Foul _Foul	5.5%	Foul _Direct	6.1%	Attempt Saved _Assisted	12.2%	Foul _Foul	13.1%
3	Foul _Offensive	5.2%	Attempt Saved _Zone	5.3%	Attempt Saved _Intentional assist	11.4%	Foul _Offensive	10.4%
4	Goal _Regular play	4.6%	Attempt Saved _Regular play	4.4%	Attempt Saved _Right footed	9.0%	Foul _Defensive	8.4%
5	Attempt Saved _Zone	4.2%	Attempt Saved _Assisted	4.2%	Attempt Saved _Head	8.3%	Dispossessed _Offensive	3.6%
6	Goal _Big Chance	4.1%	Foul _Defensive	3.8%	Attempt Saved _Individual Play	4.4%	Pass _Switch of play	2.3%
7	Goal _Right footed	4.1%	Pass _Lay-off	3.5%	Pass _Lay-off	3.4%	Attempt Saved _Big Chance	1.7%
8	Goal _Assisted	3.9%	Dispossessed _Offensive	3.3%	Pass _Chipped	2.4%	Tackle _Defensive	1.6%
9	Attempt Saved _Regular play	3.4%	Foul _Offensive	3.3%	Pass _Cross	2.4%	Tackle _Zone	1.5%
10	Goal _Intentional assist	3.4%	Attempt Saved _Right footed	2.9%	Attempt Saved _Big Chance	2.3%	Tackle _Out of play	1.5%

Table 8 Top 10 loadings shown as a percentage of total loadings for PC1 and PC2 for each 10 minute interval of the game.

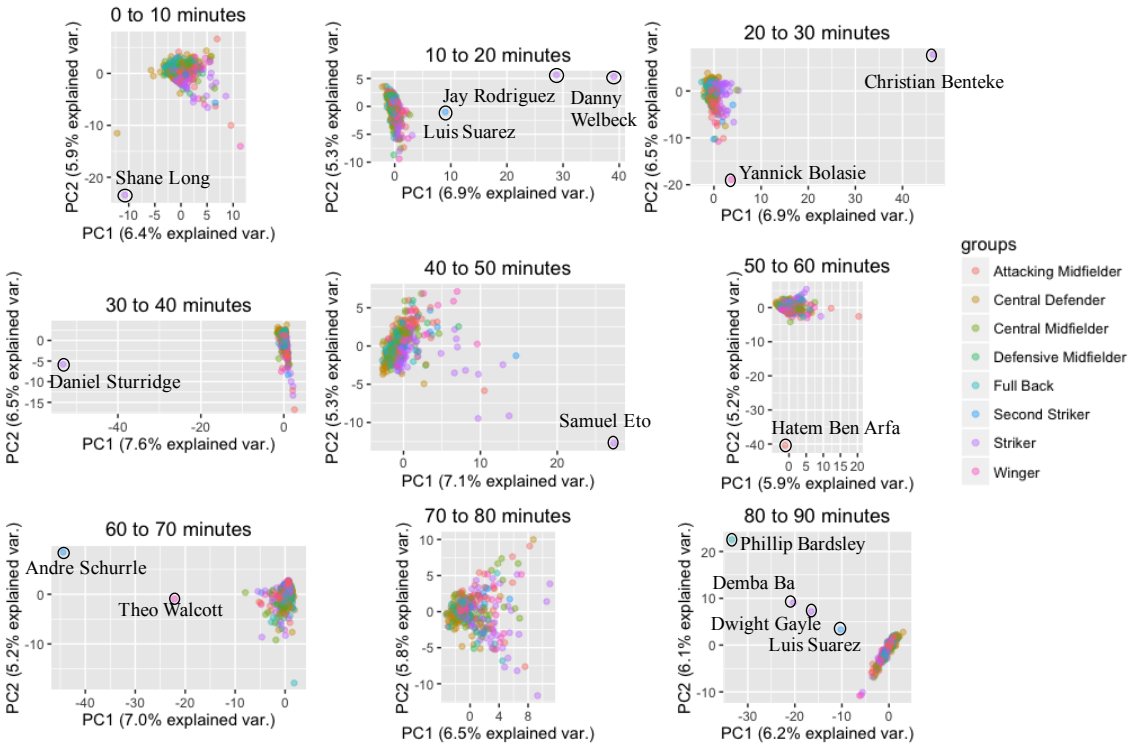


Figure 11 PCA output of each 10-minute interval of the game. Each point represents a player and each colour represents a different playing position. Outliers (unique players) labelled.

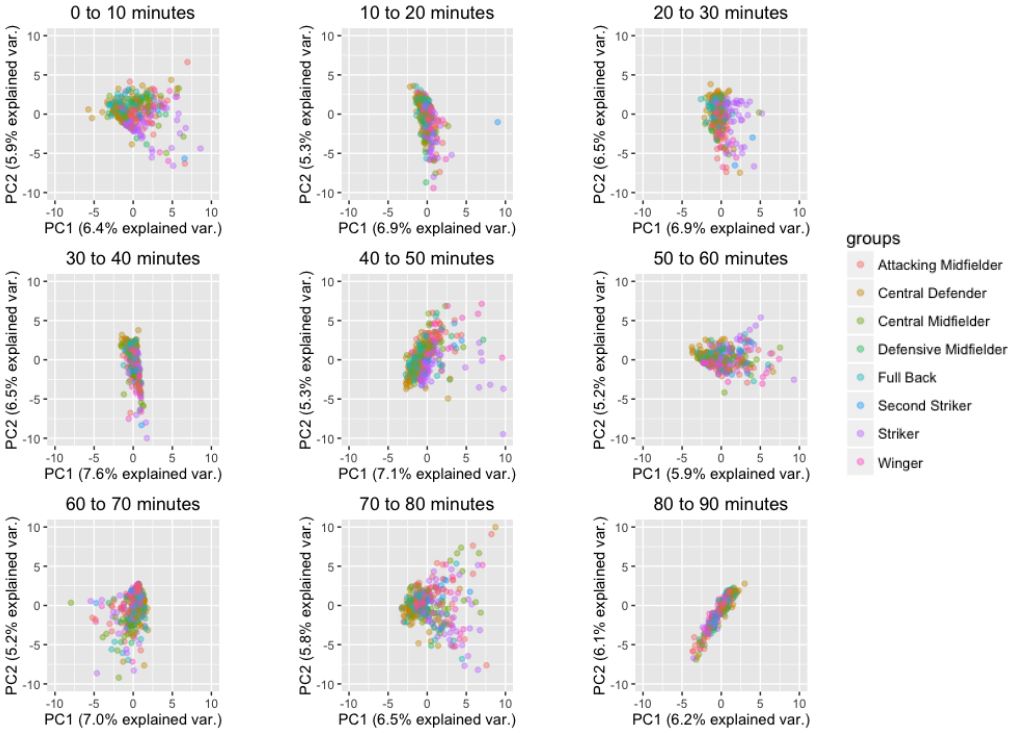


Figure 12 PCA output of each 10-minute interval of the game. Each point represents a player and each colour represents a different playing position. Outliers (unique players) not included.

To analyse the spatial element of the game relative to the players' positions the pitch was split into four segments, centre, back, left and right. This showed that the loadings were affected more by where they occurred than when they occurred. As expected more defensive on-the-ball events were dominant in both PC1 and PC2 for the back section and more attacking in the centre. The wings of the pitch seemed generally to veer towards attacking, with the left flank interestingly taking a slightly more defensive stance. The plots in

Figure 13 showed enough difference in clusters relative to the area of the pitch to conclude that it was an important factor that should be taken into account in future analysis.

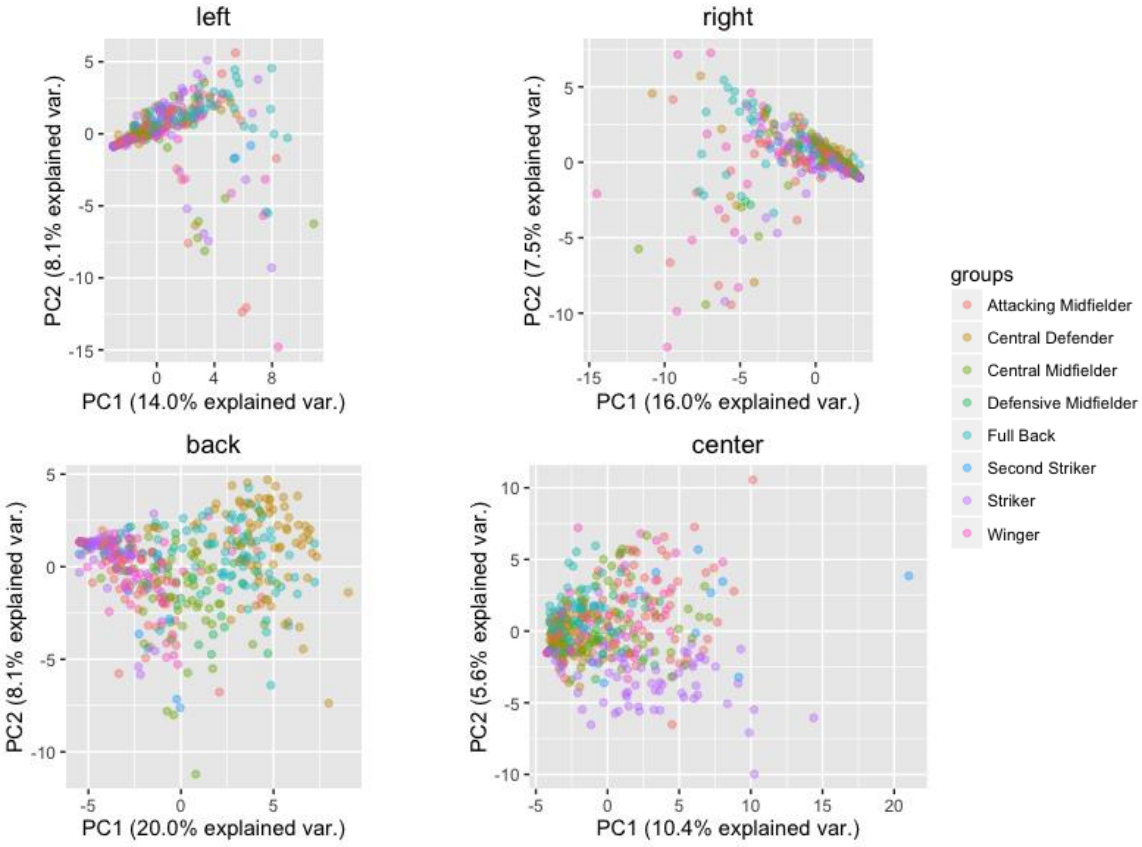


Figure 13 PCA output comparing four different areas of pitch. Each point represents a different player and each colour represents a different playing position.

Zone	Back				Centre			
PC	PC1 - 20%		PC2 - 8%		PC1 - 10%		PC2 - 5%	
1	Aerial _Defensive	7.2%	Foul _Foul	12.1%	Attempt Saved _Regular play	3.7%	Pass _Chipped	5.2%
2	Clearance _Head	6.8%	Foul _Direct	12.0%	Take On _Offensive	3.7%	Pass _Cross	5.0%
3	Pass _Long ball	6.6%	Foul _Defensive	8.9%	Pass _Lay-off	3.5%	Pass _Long ball	4.7%
4	Pass _Head pass	6.6%	Challenge _Defensive	6.4%	Attempt Saved _Assisted	3.4%	Pass _Corner taken	4.6%
5	Clearance _Out of play	5.9%	Pass _Lay-off	6.0%	Attempt Saved _Individual Play	3.1%	Aerial _Offensive	4.2%
6	Tackle _Defensive	5.5%	Foul _Offensive	5.9%	Attempt Saved _Right footed	3.1%	Pass _In-swinger	3.7%
7	Pass _Launch	4.8%	Clearance _Head	4.6%	Pass _Assist	3.0%	Pass _Direct	3.6%
8	Pass _Free kick taken	4.7%	Aerial _Defensive	3.9%	Foul _Foul	3.0%	Pass _Free kick taken	3.4%
9	Save _Def block	4.6%	Take On _Offensive	3.8%	Foul _Direct	3.0%	Pass _Out-swinger	3.4%
10	Pass _Throw-in	4.2%	Save _Def block	3.7%	Dispossessed _Offensive	2.9%	Pass _Flick-on	3.2%
Zone	Left				Right			
PC	PC1 - 14%		PC2 - 7%		PC1 - 16%		PC2 - 7%	
1	Take On _Offensive	8.7%	Foul _Foul	13.9%	Take On _Offensive	7.3%	Pass _Long ball	4.0%
2	Foul _Direct	8.0%	Foul _Direct	13.9%	Foul _Direct	6.7%	Pass _Cross	0.6%
3	Foul _Foul	7.5%	Foul _Offensive	12.0%	Foul _Foul	6.5%	Pass _Head pass	5.5%
4	Dispossessed _Offensive	7.2%	Aerial _Offensive	6.0%	Dispossessed _Offensive	6.0%	Pass _Through ball	0.7%
5	Pass _Throw-in	6.9%	Dispossessed _Offensi	5.3%	Pass _Throw-in	5.9%	Pass _Free kick taken	1.7%
6	Pass _Cross	6.1%	Pass _Chipped	4.4%	Foul _Offensive	5.7%	Pass _Corner taken	1.1%
7	Foul _Offensive	5.9%	Pass _Throw-in	3.8%	Pass _Head pass	5.3%	Pass _Zone	5.0%
8	Pass _Head pass	5.1%	Pass _Flick-on	3.7%	Aerial _Offensive	5.1%	Pass _Throw-in	5.1%
9	Pass _Chipped	5.0%	Pass _Launch	3.7%	Pass _Offensive	4.9%	Pass _Direct	1.8%
10	Aerial _Offensive	4.5%	Pass _Long ball	3.1%	Pass _GK Y Coordinate	4.8%	Pass _Chipped	5.0%

Table 9 Top 10 loadings shown as a percentage of total loadings for PC1 and PC2 of four different areas of the pitch

To confirm these results, further analysis was carried out to explore each time segment with each area of the pitch individually, i.e. back with 0-10, 10-20, 20-30; left with 0-10, 10-20, 20-30 etc. The results showed stability over time once again. The mean and variances of each 'event_qualifier_id' loading for each combination of time and area of the pitch is shown in Table 10. The low variance supports the argument that inclusion of a time component does not provide any substantial additional information.

	PC1	PC2
Average Variance	0.005	0.006
Average Mean	-0.015	-0.004

Table 10 Mean and Variance of each 'event_qualifier_id' for each combination of time (every 10 minutes) and area (back, centre, right, left)

In conclusion, the average number of times a player committed an on-the-ball event was a good metric for predicting their position. Segmenting this by the time the event was completed did not appear to add any further information, whereas adjusting for a spatial component did. PCA is interesting for exploratory analysis and grouping of elements but to really understand which on-the-ball events were most important in defining a player's position on the pitch further analysis was required.

4.2. The Most Important Events for Defining a Player's Position

Running a Random Forest of 500 trees with average count of 'event_qualifier_ids' per game as the inputs and the position of the player as the target variable was used to determine the variables deemed most important when understanding a player's position. Understanding the predictive power of the model gave some further insight into the properties of playing positions. Table 11 shows the error rate of the training set with central defenders and strikers having less than 0.2 errors on nearly all areas of the pitch. The precision and recall[5] values shown in Table 12 support this with approximately 0.8 precision for both positions (striker slightly highly) and approximately 0.6 recall. This shows that these two positions are the most distinct in their playing styles. Some of the positions have very high errors dues to their similarities to each other and therefore the model finds difficulty in predicting the position of the player. Combining certain positions such as striker and second striker would vastly improve its predictive capabilities.

[5] Precision: number of retrieved instances that are relevant. Recall: Number of relevant instances that are retrieved (formulas found in glossary).

Left	Attacking Midfielder	Central Defender	Central Midfielder	Defensive Midfielder	Full Back	Second Striker	Striker	Winger	Classfication Error
Attacking Midfielder	22	4	9	0	2	0	12	7	0.61
Central Defender	0	67	2	0	11	0	2	0	0.18
Central Midfielder	13	11	25	0	3	0	8	0	0.58
Defensive Midfielder	2	11	10	0	0	0	2	1	1.00
Full Back	0	15	2	0	49	0	1	1	0.28
Second Striker	6	1	0	0	0	0	6	0	1.00
Striker	3	2	4	0	1	0	63	5	0.19
Winger	7	3	8	0	5	0	12	9	0.80

Centre	Attacking Midfielder	Central Defender	Central Midfielder	Defensive Midfielder	Full Back	Second Striker	Striker	Winger	Classfication Error
Attacking Midfielder	26	1	8	1	3	0	12	6	0.54
Central Defender	0	67	1	0	14	0	1	1	0.20
Central Midfielder	13	3	31	1	2	0	9	2	0.49
Defensive Midfielder	0	6	14	2	1	0	4	0	0.93
Full Back	1	14	0	0	53	0	1	0	0.23
Second Striker	7	0	0	0	0	0	5	1	1.00
Striker	1	1	2	0	1	0	75	0	0.06
Winger	14	0	5	0	9	0	6	12	0.74

Right	Attacking Midfielder	Central Defender	Central Midfielder	Defensive Midfielder	Full Back	Second Striker	Striker	Winger	Classfication Error
Attacking Midfielder	8	4	15	0	2	0	16	12	0.86
Central Defender	1	56	4	0	17	0	6	1	0.34
Central Midfielder	13	10	25	3	3	0	5	2	0.59
Defensive Midfielder	2	8	11	0	1	0	5	0	1.00
Full Back	0	15	3	0	45	0	5	1	0.35
Second Striker	6	0	0	0	0	0	6	1	1.00
Striker	2	2	2	0	2	0	68	1	0.12
Winger	14	3	5	0	6	0	11	6	0.87

Back	Attacking Midfielder	Central Defender	Central Midfielder	Defensive Midfielder	Full Back	Second Striker	Striker	Winger	Classfication Error
Attacking Midfielder	18	2	10	0	1	0	14	12	0.68
Central Defender	0	72	1	0	10	0	2	0	0.15
Central Midfielder	13	2	31	5	2	0	7	1	0.49
Defensive Midfielder	5	4	15	1	0	0	1	1	0.96
Full Back	1	10	0	0	57	0	2	0	0.19
Second Striker	6	0	2	0	0	0	5	0	1.00
Striker	3	0	3	0	0	0	71	3	0.11
Winger	13	0	6	0	3	0	18	6	0.87

Table 11 Confusion Matrix and classification errors of the Random Forest training set used to predict player positions.

	Attacking Midfielder		Central Defender		Central Midfielder		Defensive Midfielder	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Left	0.39	0.42	0.82	0.59	0.42	0.42	0.00	NA
Right	0.14	0.17	0.65	0.57	0.41	0.38	0.00	0.00
Center	0.46	0.42	0.80	0.73	0.51	0.51	0.07	0.19
Back	0.32	0.31	0.85	0.80	0.51	0.46	0.04	0.24

	Full Back		Second Striker		Striker		Winger	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Left	0.72	0.69	0.00	NA	0.81	0.59	0.20	0.39
Right	0.65	0.59	0.00	NA	0.88	0.56	0.13	0.25
Center	0.77	0.64	0.00	NA	0.94	0.66	0.26	0.55
Back	0.81	0.78	NA	NA	0.89	0.59	0.13	0.26

Table 12 Precision and Recall values of the Random Forest training set used to predict player positions.

To find the most important variables relative to players' positions the Importance Values of variables were explored. Four separate models were created for the different areas of the pitch for the entirety of the games (0-100 minutes to include any injury time). The way in which the dataset was constructed meant that only certain on-the-ball events contained a 'zone' variable. Therefore this information could not give an overview of all 'event_qualifier_ids' but would give a good understanding of the majority. Overall, there were 134 'event_qualifier_ids' included in the model. To aggregate the Importance Values an average value for each 'event_qualifier_id' was calculated across the different areas of the pitch with the 'event_qualifier_ids' with the highest average chosen as most relevant. The highest 10 on-the-ball events all included passing, clearly showing this event to be the most important relative to defining player positions. Table 13 shows the top 20 'event_qualifier_id' and their average Importance Values across the four areas of the pitch.

Average Importance Value	Event and Qualifier ID
14.19	Pass_Long ball
14.02	Pass_Throw-in
12.21	Pass_Lay-off
11.73	Pass_Launch
10.31	Pass_Chipped
10.15	Pass_Cross
10.00	Pass_Head pass
6.47	Pass_Flick-on
6.03	Pass_Corner taken
5.36	Take On_Offensive
5.34	Pass_Free kick taken
5.09	Aerial_Offensive
4.94	Pass_Direct
4.88	Dispossessed_Offensive
4.81	Clearance_Blocked cross
4.13	Pass_Switch of play
4.06	Tackle_Defensive
4.01	Clearance_Head
3.88	Take On_Defensive

Table 13 Top 20 'event_qualifier_ids' and their Importance Values averaged across the four areas of the pitch.

Up until this stage the areas of the pitch had been segmented into four sections. However, after concluding that the area of the pitch that a player completed an on-the-ball event was a key component in defining what type and how many 'event_qualifier_ids' were completed, it seemed suitable to break these areas down further into 10x10 segments. To get an overview of where most on-the-ball events were carried out on the pitch visualisations were created which aggregated the frequency of on-the-ball events for all players over the whole season for each 'event_id' of interest. These heat maps represent different segments on the pitch with the defensive goal situated on the left of the map and the offensive goal situated on the right (as shown in Figure 18).

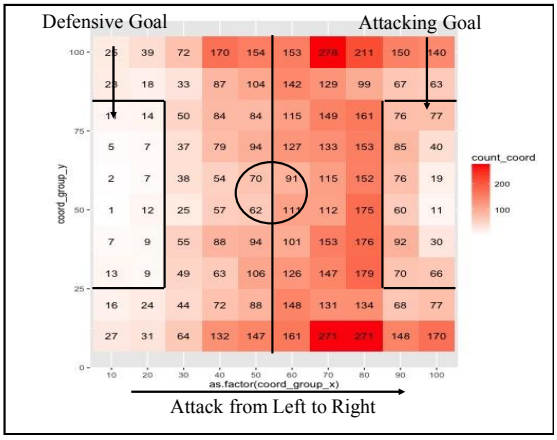


Figure 14 Example of how the heat maps relate to the different areas of the pitch. Attacking always from left to right with no regard to the team or period of the game.

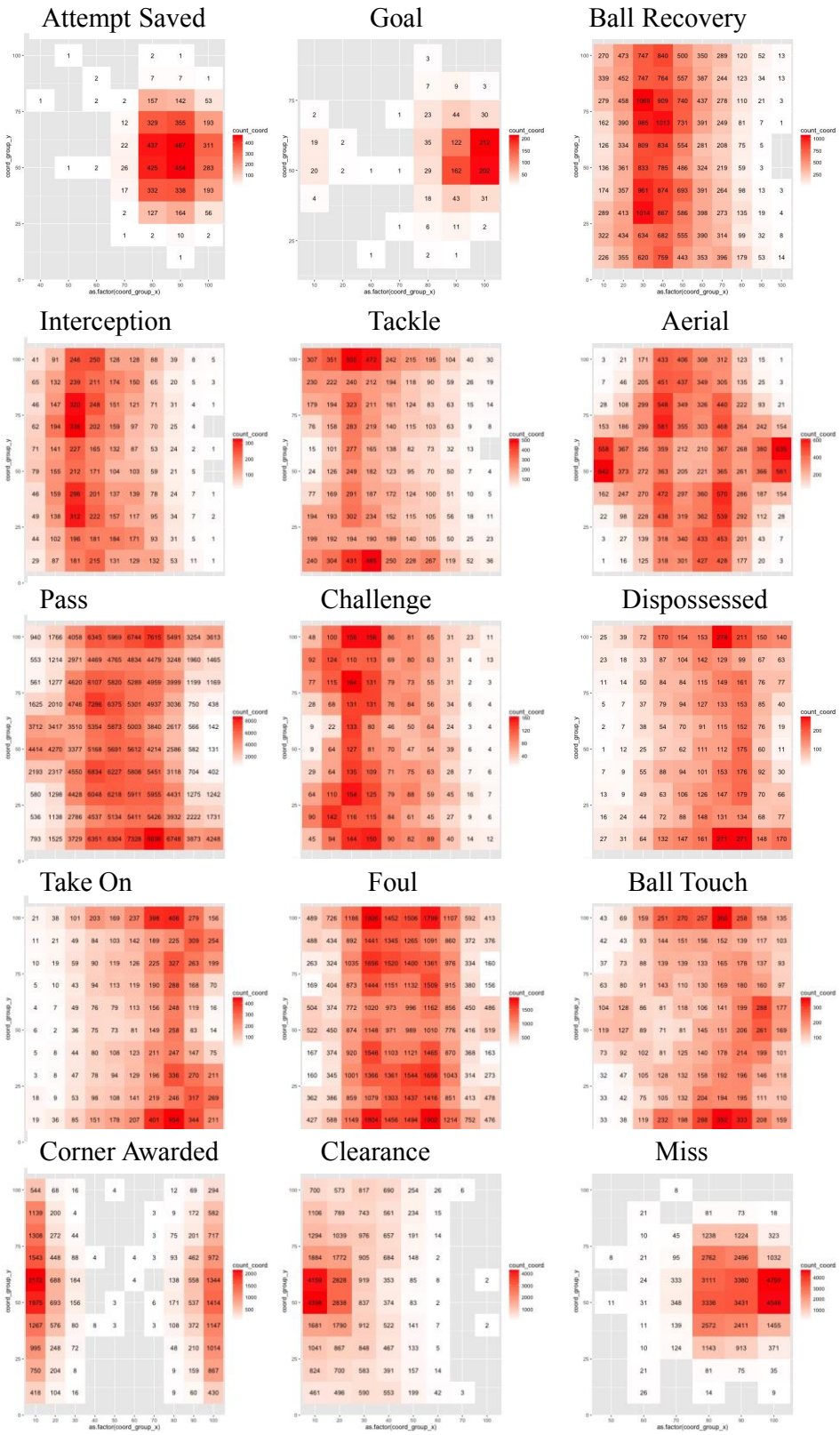


Figure 15 Heat maps of total frequency of 'event_ids' for all players over the season. Defensive goal situated on the left and attacking goal on the right.

The gradient effect from red to white of the heat maps showed that the 10x10 partitions distributed the frequency of events well. An interesting result of viewing the heat maps in Figure 15 was the symmetrical properties shown between the left and right wings for all the 'event_ids'. These heat maps also showed a much higher frequency (average of approximately 5,000) of the number of 'passes' compared to any other 'event_id'.

To explore this further Figure 16 shows the 'event_qualifier_id' for passes with a variety of different 'qualifier_ids'. Breaking passes down in this way transformed the frequencies to a more normally distributed level compared to the other 'event_ids'. The results of the variable selection method from the Random Forest heavily favoured the different types of passing and the results of the frequency heat maps suggested the passing event should be broken down to 'qualifier_id' detail to ensure there were no skewed frequencies. Therefore, it was decided the variables chosen to be used in the models would be mainly 'event_ids' with only 'qualifier_ids' relating to passing included. The 'event_qualifier_ids' shown in Table 14 were the final list chosen.

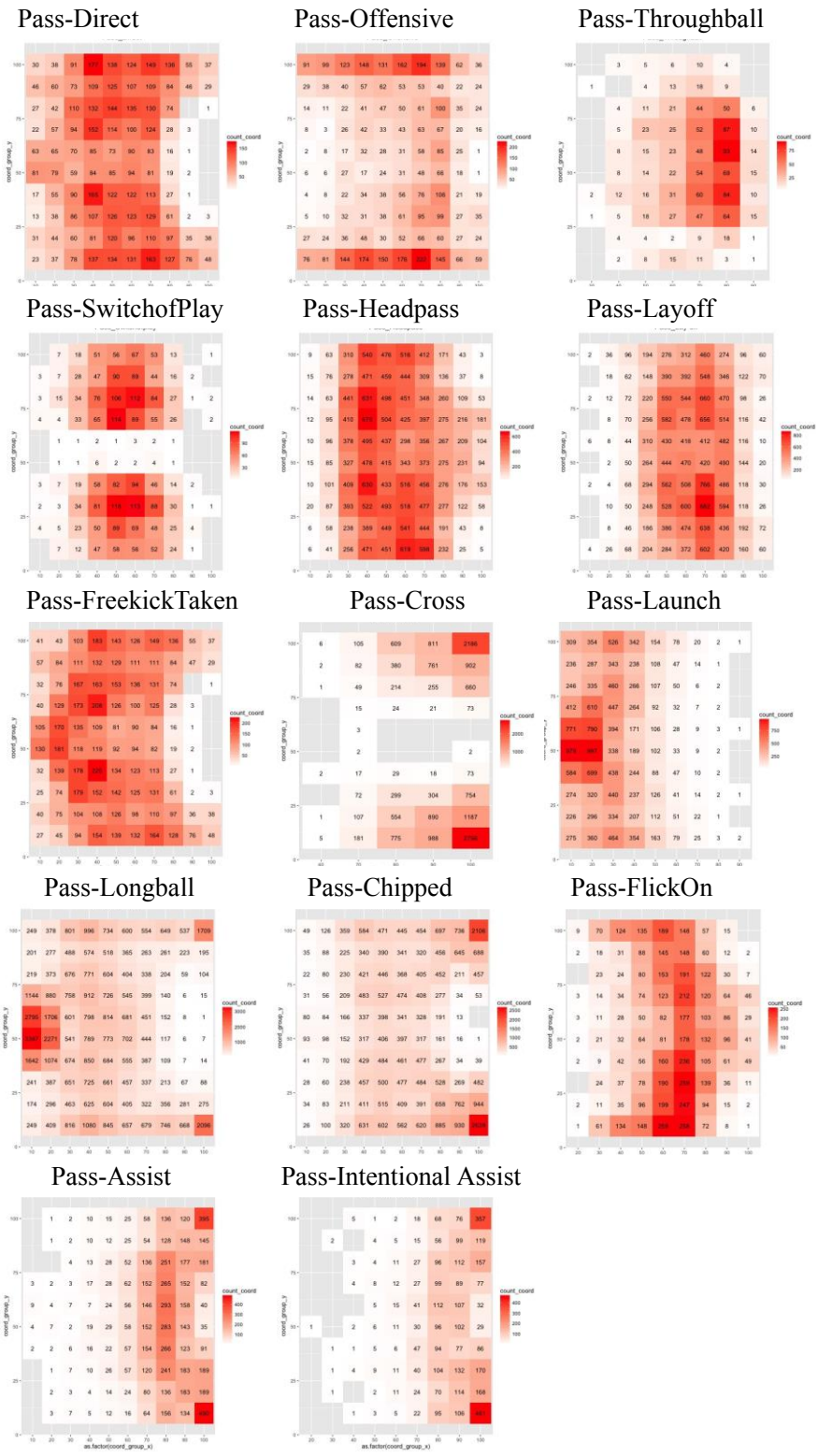


Figure 16 Heat maps of total frequency of 'event_qualifier_ids' related to passing for all players over the season. Defensive goal situated on the left and attacking goal on the right.

Event_Qualifier Id	Description	Average Importance
49	Ball Recovery	NO ZONE PROVIDED
1_1	Pass_Long ball	14.19
1_107	Pass_Throw-in	14.02
1_156	Pass_Lay-off	12.21
1_157	Pass_Launch	11.73
1_155	Pass_Chipped	10.31
1_2	Pass_Cross	10.15
1_3	Pass_Head pass	10.00
1_168	Pass_Flick-on	6.47
1_6	Pass_Corner taken	6.03
3	Take On	5.36
1_5	Pass_Free kick taken	5.34
44	Aerial	5.09
50	Dispossessed	4.88
1_196	Pass_Switch of play	4.13
7	Tackle	4.06
1_210	Pass_Assist	2.94
45	Challenge	2.89
4	Foul	2.88
1_223	Pass_In-swinger	2.02
15	Attempt_saved	1.95
2	Offside	1.80
1_224	Pass_Out-swinger	1.79
1_4	Pass_Through ball	1.09
16	Goal	0.99
1_195	Pass_Pull Back	0.61
61	Ball Touch	0.47
1_225	Pass_Straight	0.42
6	Corner Awarded	0.01

Table 14 Average Importance Values of final chosen variables. 'event_id' 49 had no zone provided in the dataset but was still chosen as a variable for future stages.

4.3. The On-The-Ball Events that are Most Important Relative to Success

Using the 'event_qualifier_ids' found from Section 4.2. and the method of manipulating the dataset into path-level data described in Section 3.4. a Performance Indicator was created for each 10x10 area of the pitch and each 'event_qualifier_id'.

4.3.1 Robustness Test

To test whether the chosen timestamp intervals or the number of previous events had any substantial impact on the Importance Values the model was run for 12 combinations of time/previous events. The average Importance Value for each 'event_qualifier_id', segment of the pitch and cross validation (10-fold) was taken for each XY model. The Importance Values were then ranked for each model with the smallest ranking (1) being the highest value. The aim was to test whether there was stability across the models for each of the

'event_qualifier_ids'. The complexity of analysing the output of combining 39 'event_qualifier_ids', 12 models and 100 coordinates was considerable. However, it was decided the best way to do this was to plot the rankings for each 'event_qualifier_id' against each other for every model, in the hope that similar patterns would emerge if the models were similar and therefore robust. Figure 17 showed positive results, with similar patterns observed in all the models. Where there was variance this was largely explained by the difference in time intervals, with 30-second models behaving slightly differently to the 60-second models, but the increase of previous on-the-ball events having minimal effect.

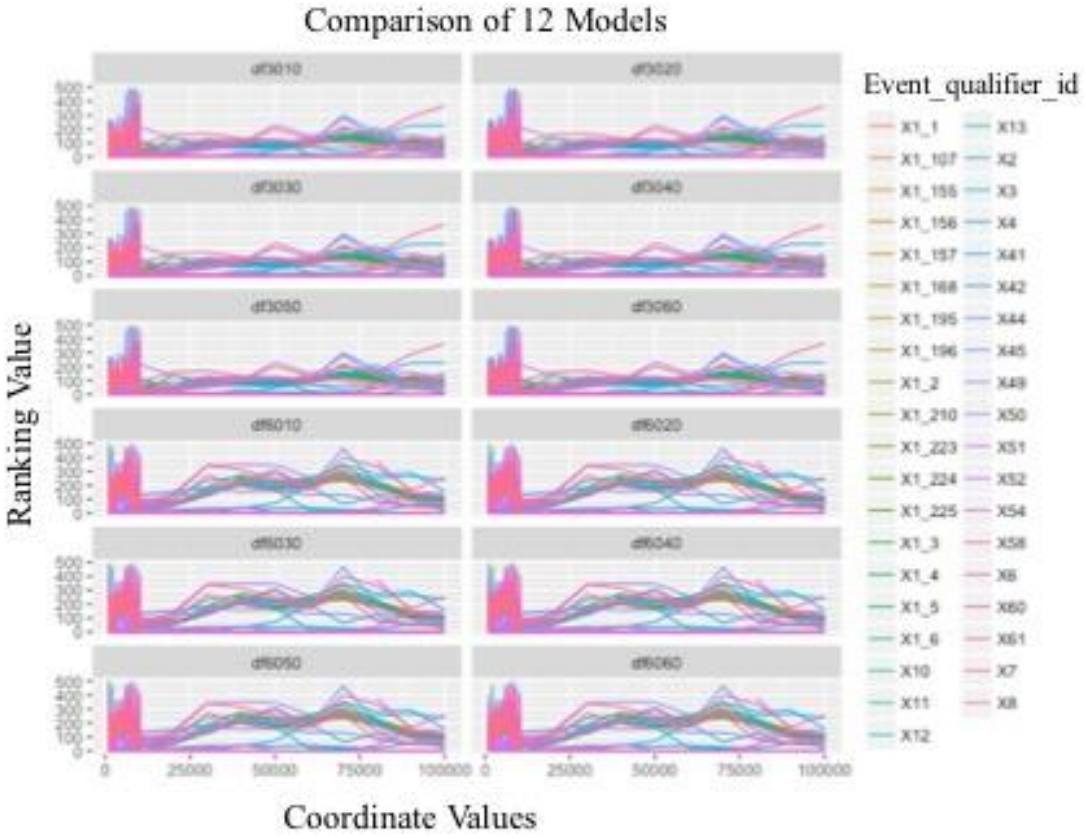


Figure 17 Output of test for robustness; y-axis shows the ranking value, x-axis shows the different coordinate areas of the pitch. 30-second models labelled df30XX and 60-second models labelled df60XX, where XX ranges from 10 to 60. The different coloured lines represent the different 'event_qualifier_ids'.

4.3.2 3010 Model

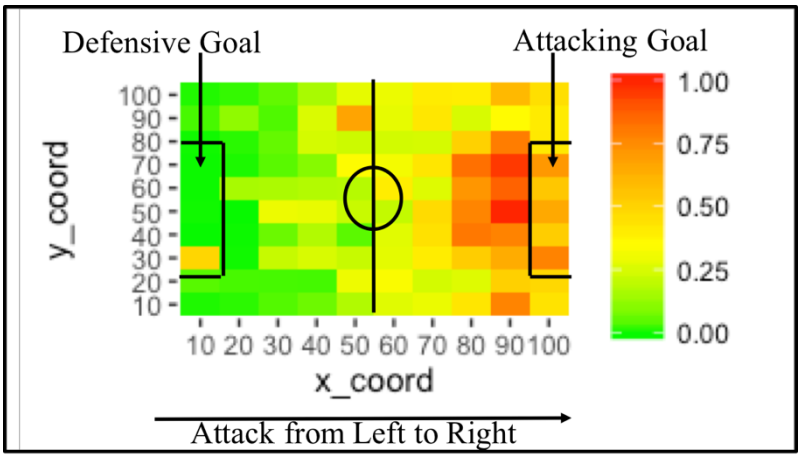


Figure 18 Example of how the Performance Indicator and Performance Metric heat maps relate to the different areas of the pitch. Attacking always from left to right with no regard to the team or period of the game. .

The initial model (3010 Model) included every 30-second interval and the previous 10 on-the-ball events that occurred with the target variable as shots and goals. It was soon clear that the Performance Indicator was heavily weighted towards attacking players with all areas of the pitch near the goal having the highest values. Analysing the Performance Indicator heat maps for each of the events (example shown in Figure 19) clearly showed the model had over valued the areas nearest the goal. There were two key reasons for this. First, although using every 30-seconds of each game and the last ten on-the-ball events captured all of the events at least once in the match, it did not fully account for the lead up to the play. If a shot or goal was recorded it was inevitable that it would have been in the areas of yellow and red shown on the heat maps in Figure 19 and none of these 'successes' would be attributed to other areas of the pitch. This was again a consequence of the low frequency of positive target variables available in football analytics. The second reason was that the model was strongly weighted towards attacking players, resulting in strikers and attacking players being given the highest value. To account for these issues, the number of previous events was increased to try and model a longer lead up of play and a second model was created that included a defensive measure.

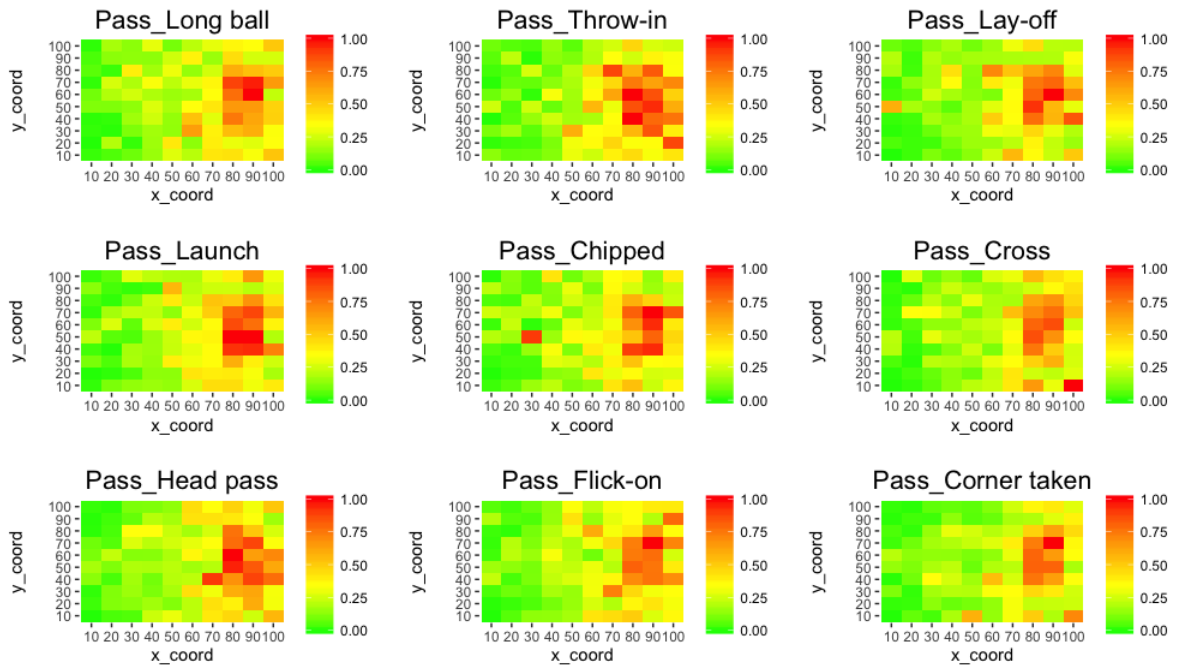


Figure 19 Example of nine Performance Indicator heat maps of 3010 model; showing a bias towards the offensive areas of the pitch.

4.3.3 3060 Model

Defence in football is a much harder metric to measure and define success. There are clear events that highlight success, such as a clean sheet[6] during a game or a count of successful dispossessions[7]. Visualising the frequency of certain 'event_qualifier_ids' as shown in Figure 15 provided insight into some variables that could be used as a defensive measure.

Event	Description
Ball Recovery	Team wins the possession of the ball and successfully keeps possession for at least two passes or an attacking play
Shot Saved	This event is for the player who made the shot.
Goal	All Goals

[6] Not conceding any goals during a game
[7] Taking possession from the other team

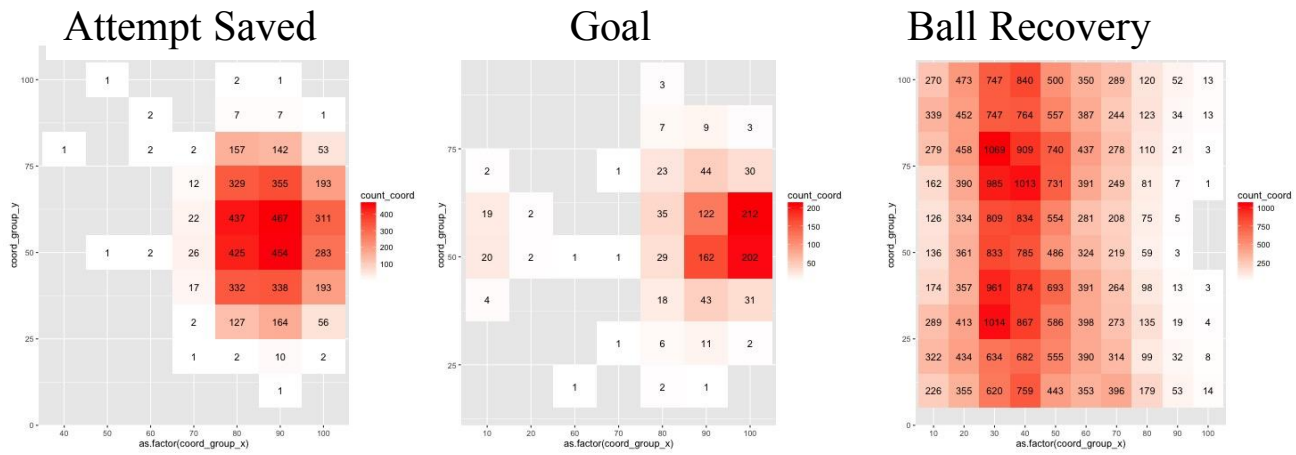
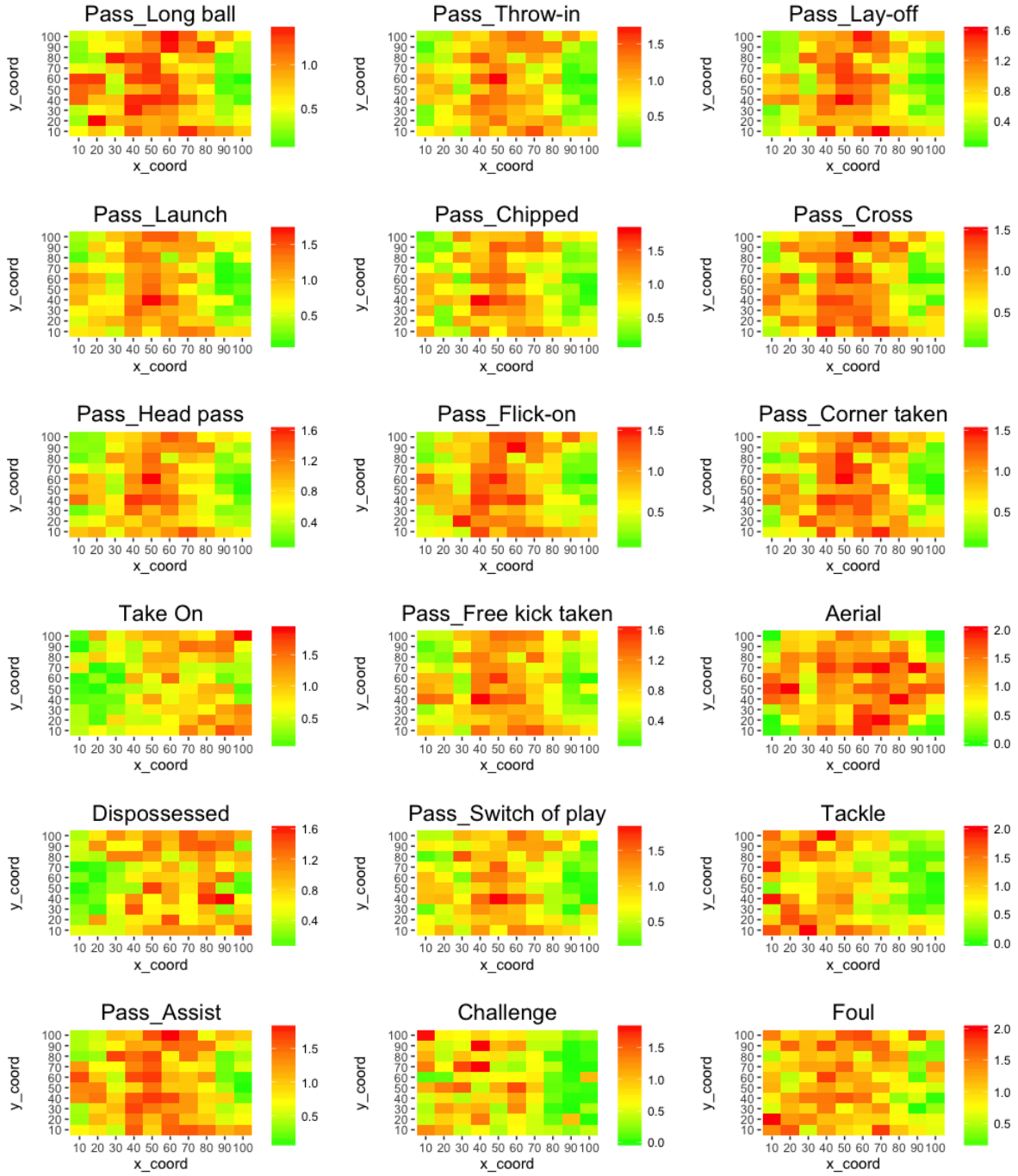


Figure 20 Heat maps of the three 'event_ids' used as target variables in the final model

Figure 20 shows the frequencies of the three main on-the-ball events used as the final target variables. The aim was to find on-the-ball events that when combined had similar frequencies on all areas of the pitch to remove any biases this may incur on the model. Ball recovery seemed a suitable choice for the defensive measure as it occurred frequently and not only captured a change in possession of the ball but also the ability to retain this possession for a minimum of two passes. Rather than create a single model including paths labelled with either a goal, shot or ball recovery it was decided to create separate models; one measuring offense and one measuring defence. This was done to keep interpretation of results as simple as possible.

Analysing the Performance Indicator values (calculated by standardising the Importance Values attained from the Random Forest models) gave an insight into where on the pitch the model attributed value to each event. Initially it was clear that most of the passing events behaved similarly with highest values shown in the midfield area. This suggested that creating a pass in midfield was more important when producing shots/goals or recovering the ball than in other areas of the pitch. However, within this midfield area there were some 'event_qualifier_ids' which favoured the flanks of the pitch. This was most noticeable in a 'pass_long ball', 'pass_lay off' and 'pass_cross' For the non-pass related on-the-ball events there were some interesting findings such as dispossession (player is successfully tackled and loses possession of the ball) having higher importance at the attacking end of the pitch suggesting the value of a striker dispossessing the ball is higher than a defender. Conversely,

a challenge (when a player fails to win the ball as an opponent successfully dribbles past them) is given higher importance at the defending area of the pitch.



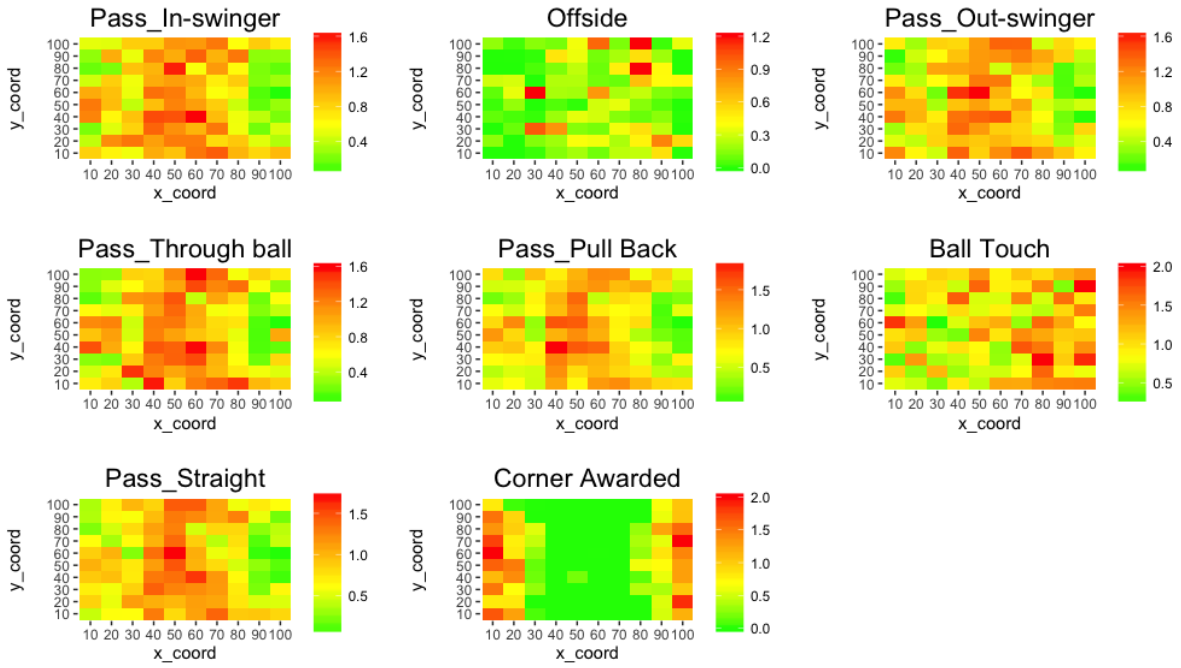


Figure 21 Heat maps of Performance Indicator with red areas displaying the areas of the pitch with the highest importance for different events.

4.4. A Player's Optimum Position on the Pitch.

4.4.1 Top/Bottom League Comparison

To test the results of the Performance Indicator and Performance Metric it was decided to compare the team that finished top and the team that finished bottom of the Premier League in the 2013-2014 season. To get a spatial understanding of the players' positioning the results were visualised and compared on heat maps. The top heat map shows the frequency of total player 'event_qualifier_ids' for each pitch area and the bottom the relative Performance Metric as shown in Figure 22.

To understand quantitatively whether the model believed a player played well an aggregate of the Performance Metric for the whole pitch was created. Table 15 shows the total Performance Metrics for each team in the league. Those highlighted green finished first and second and those red came last and second last in the 2013-2104 Premier League. These are positive results because the model's Performance Metric confirmed these top and bottom placing teams in the top and bottom four ranked positions. However promising these initial findings were, in order

to answer the research question it was necessary to break these results down further to understand those players who performed the best or worst and then relate this back to their position on the pitch.

Team ID	Team	Average Performance Metric per Team
14	Liverpool	30.39
80	Swansea City	30.14
6	Tottenham Hotspur	28.96
43	Manchester City	28.71
20	Southampton	27.77
4	Newcastle United	27.45
1	Manchester United	26.99
21	West Ham United	26.87
45	Norwich City	26.78
3	Arsenal	25.65
11	Everton	25.07
7	Aston Villa	24.95
88	Hull City	24.75
8	Chelsea	24.54
110	Stoke City	24.5
35	West Bromwich Albion	24.39
56	Sunderland	24.31
97	Cardiff City	23.78
31	Crystal Palace	22.71
54	Fulham	22.54

Table 15 Average performance of each team over the entire 2013-2014 season. Manchester City and Liverpool (highlighted green) finished in first and second place. Cardiff City and Fulham (highlighted red) finished in the last two positions.

The top and the bottom placing clubs in the 2013-2014 season were Manchester City and Cardiff City, respectively. An average Performance Metric across all the players for the entire season gave Manchester City 28.71 and Cardiff City 23.78. Table 16 shows the aggregate results for each player on the pitch for the whole season. Again positive results can be seen for Manchester City, with the top six players from the model appearing as 6 of the 8 rewarded 'Etihad's player of the month' award[8] throughout the season. Not so surprisingly, considering their season, most of Cardiff's strikers appear in the bottom half of the table.

[8] Álvaro Negredo, Yaya Touré x2, Sergio Agüero, Samir Nasri, Fernandinho, Edin Džeko, David Silva, Martín Demichelis

Player ID	Player Name	Player Position	43:Manchester City
28554	Samir Nasri	Attacking Midfielder	50.60
14664	Yaya Toure	Defensive Midfielder	50.35
20664	David Silva	Attacking Midfielder	48.84
27789	Fernandinho	Defensive Midfielder	47.06
42544	Edin Dzeko	Striker	42.00
15312	Martin Demichelis	Central Defender	39.17
15157	James Milner	Attacking Midfielder	37.19
19534	Javi Garcia	Defensive Midfielder	36.26
42892	Ilvaro Negredo	Striker	35.88
20658	Pablo Zabaleta	Full Back	34.44
17740	Jesls Navas	Winger	32.33
17476	Vincent Kompany	Central Defender	30.20
37572	Sergio Aguero	Striker	29.66
17336	Gail Clichy	Full Back	29.12
42593	Aleksandar Kolarov	Full Back	28.90
7551	Joleon Lescott	Central Defender	22.21
84702	Matija Nastasic	Central Defender	20.61
50442	Stevan Jovetic	Second Striker	15.60
15749	Joe Hart	Goalkeeper	9.95
20492	Micah Richards	Full Back	6.76
56827	Costel Pantilimon	Goalkeeper	5.30
49384	Jack Rodwell	Defensive Midfielder	4.58
80235	Dedryck Boyata	Central Defender	3.30

Player ID	Player Name	Player Position	97:Cardiff City
49438	Jordon Mutch	Central Midfielder	50.03
15282	Peter Whittingham	Attacking Midfielder	47.25
48853	Gary Medel	Defensive Midfielder	45.63
28541	Fraizer Campbell	Striker	43.79
77877	Kim Bo-Kyung	Attacking Midfielder	41.48
68815	Steven Caulker	Central Defender	38.44
49845	Aron Gunnarsson	Defensive Midfielder	35.48
1231	Craig Bellamy	Striker	34.29
36956	Ben Turner	Central Defender	33.12
58636	Klevin Thleophile-Catherine	Central Defender	29.59
26900	Peter Odemwingie	Striker	29.15
55913	Craig Noone	Winger	28.10
44925	Don Cowie	Central Midfielder	25.76
54771	Fabio	Full Back	25.48
82403	Wilfried Zaha	Attacking Midfielder	22.03
18145	Andrew Taylor	Full Back	21.43
65972	Declan John	Full Back	21.14
18215	Kenwyne Jones	Striker	19.72
15144	David Marshall	Goalkeeper	12.58
27698	Matthew Connolly	Central Defender	8.91
114517	Andreas Cornelius	Striker	6.80
7638	Mark Hudson	Central Defender	6.71
49207	Rudy Gestede	Striker	2.80
20441	Nicky Maynard	Striker	2.47
19686	Joe Lewis	Goalkeeper	2.40
131403	Rhys Healey	Striker	0.19

Table 16 LHS: Manchester City who finished top of the 2013-2014 Premiership and RHS: Cardiff City who finished bottom. The values in column 4 show the total Performance Metric for the individual players in each team.

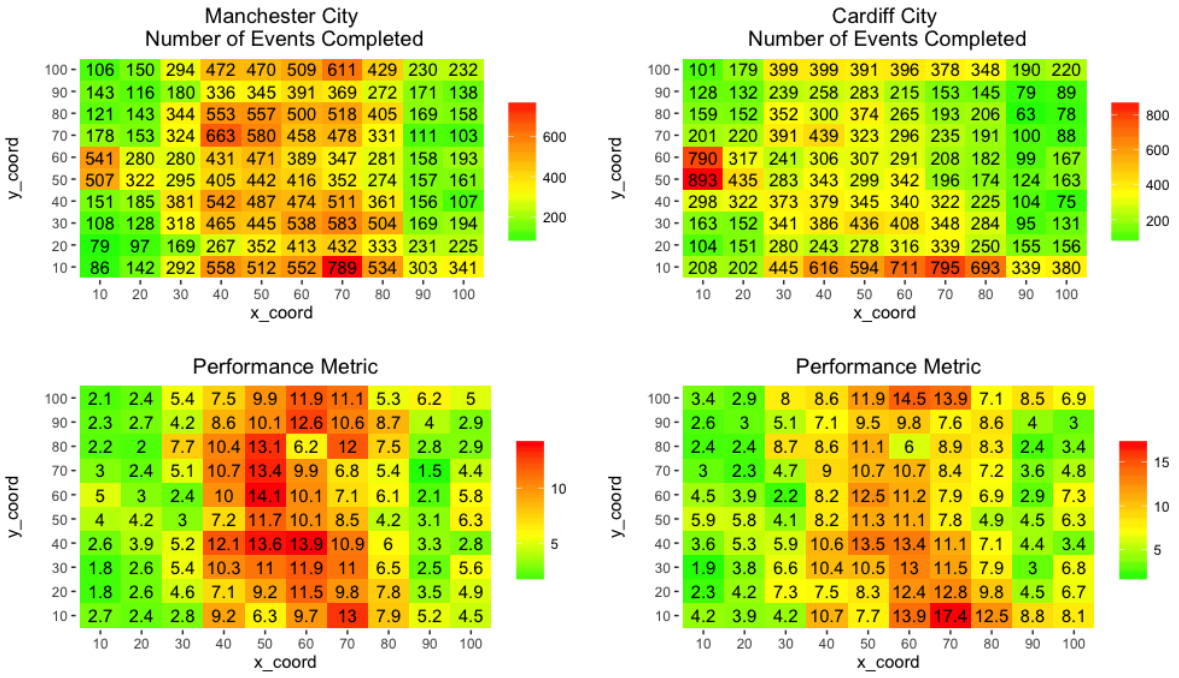


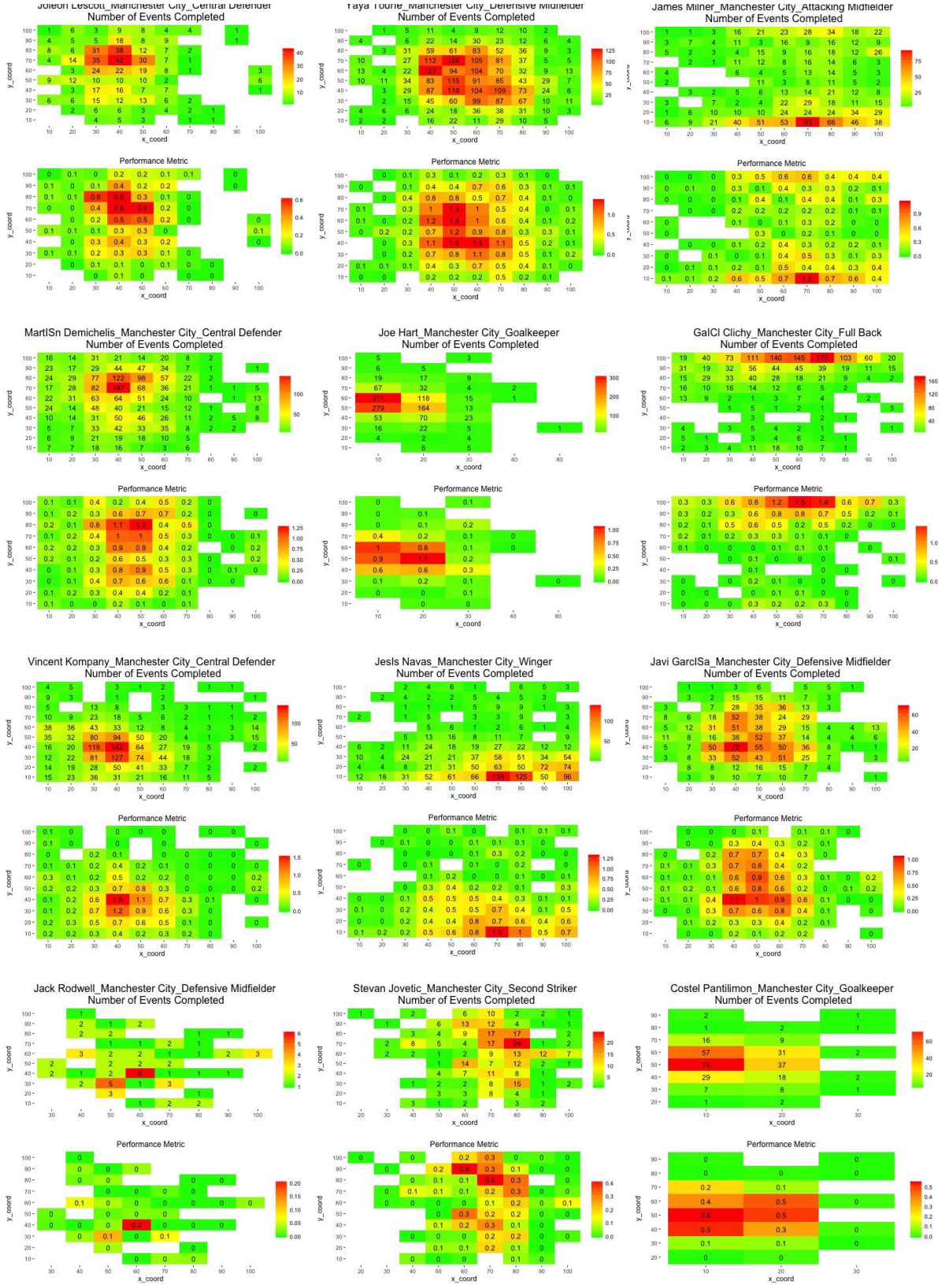
Figure 22 Frequency heat map (top) and Performance Metric values (bottom) for all players over the entire season. LHS: Manchester City RHS: Cardiff City

Figure 22 shows the differences over the whole season for all the players for Manchester City and Cardiff City. The top charts for both indicate the frequency of on-the-ball events and show that Cardiff's goalkeeper was tested substantially more than Manchester City's. The much

darker red areas of both the model's Performance Metric and frequency heat maps shows that Manchester City had more possession and used this well. This suggests that possession is an important factor for success. The Performance Metric heat map also shows that areas where there were similar amounts of possession between the two teams, such as on the right wing, there is evidence to suggest that Cardiff actually performed better than Manchester City. This is indicated by the difference in Performance Metric in the (70,10) coordinate giving Manchester City 13 and Cardiff City 17.4. This could suggest that gaining and keeping possession was Cardiff's biggest problem, rather than on-the-ball skill. Although Cardiff may have shown greater value on the wing, overall Manchester City performed better in most other areas, agreeing with the result of the season[9].

It was clear that over the entirety of the season Cardiff City favoured possession on the wings with nearly double the frequency of all other events on the pitch being performed on the right wing (attacking from left to right). Although the model picks this up as being favourable, there is not a large difference to the Performance Metric values on the left wing (coordinates (70,10) and (70,100) showing 17.4 and 13.9 respectively). This suggests that although they had more plays of the ball on the right they may actually have created more opportunity on the left. To try and figure out if it was the position of Cardiff's players which let them down it was necessary to view each individual player's Performance Metric for the whole season as seen in Figure 23 and Figure 24. Manchester City clearly showed that where players played they performed well, with minimal variation shown in the heat maps between their frequency of on-the-ball events and their Performance Metric.

[9] Appendix B shows Performance Metric values for the rest of the teams for the 2013-2014 season.


Joleon Lescott_Manchester City_Central Defender
Number of Events Completed
Performance Metric
Yaya Toure_Manchester City_Defensive Midfielder
Number of Events Completed
Performance Metric
James Milner_Manchester City_Attacking Midfielder
Number of Events Completed
Performance Metric
MartISn Demichelis_Manchester City_Central Defender
Number of Events Completed
Performance Metric
Joe Hart_Manchester City_Goalkeeper
Number of Events Completed
Performance Metric
GaICl Clichy_Manchester City_Full Back
Number of Events Completed
Performance Metric
Vincent Kompany_Manchester City_Central Defender
Number of Events Completed
Performance Metric
JesIs Navas_Manchester City_Winger
Number of Events Completed
Performance Metric
Javi GarcISa_Manchester City_Defensive Midfielder
Number of Events Completed
Performance Metric
Jack Rodwell_Manchester City_Defensive Midfielder
Number of Events Completed
Performance Metric
Stevan Jovetic_Manchester City_Second Striker
Number of Events Completed
Performance Metric
Costel Pantilimon_Manchester City_Goalkeeper
Number of Events Completed
Performance Metric
x_coord
y_coord

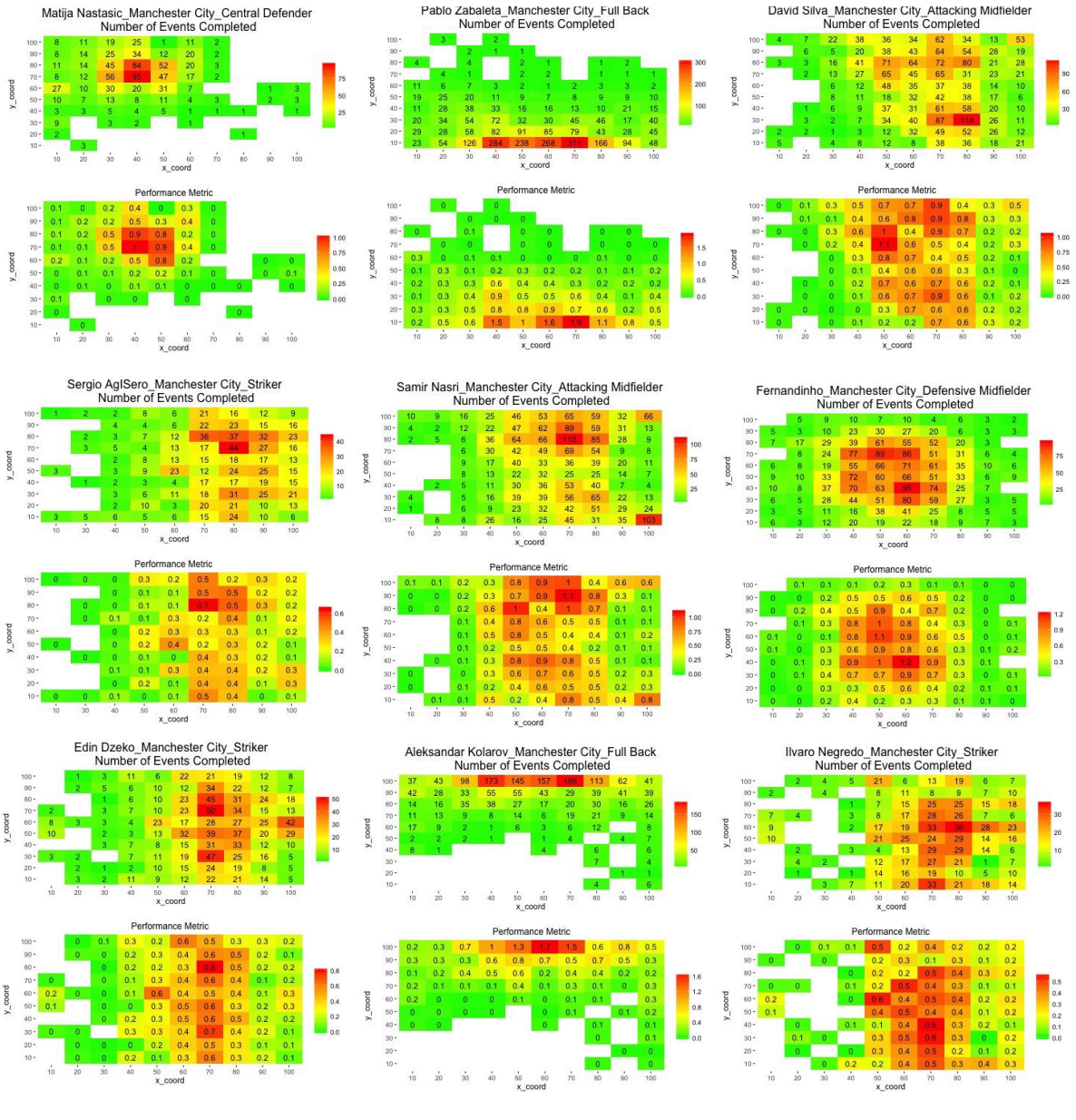
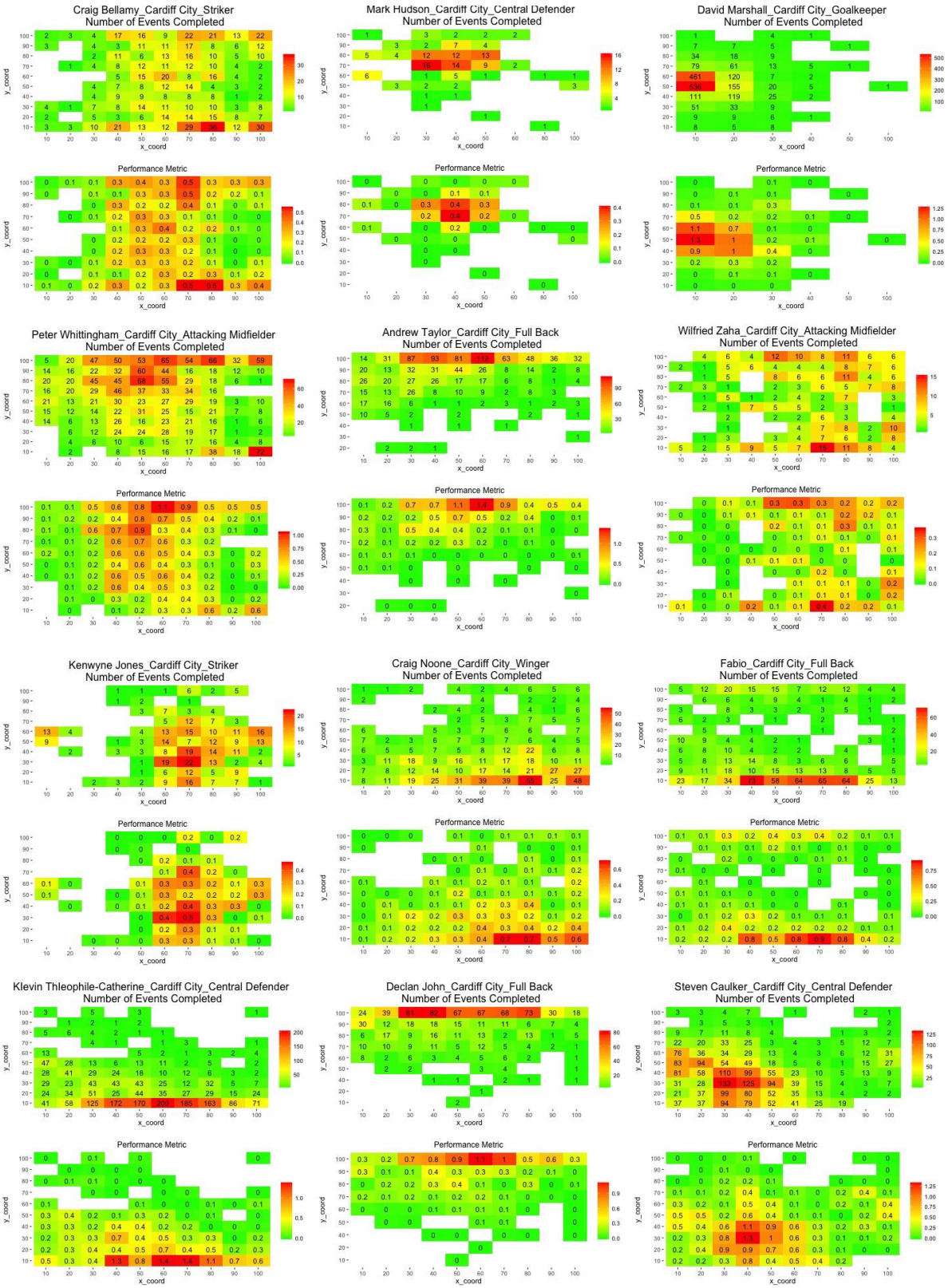


Figure 23 Heat maps of frequency (top heat map) and Performance Metric (bottom heat map) of all Manchester City's players.


Craig Bellamy_Cardiff City_Striker
Number of Events Completed
Performance Metric
Mark Hudson_Cardiff City_Central Defender
Number of Events Completed
Performance Metric
David Marshall_Cardiff City_Goalkeeper
Number of Events Completed
Performance Metric
Peter Whittingham_Cardiff City_Attacking Midfielder
Number of Events Completed
Performance Metric
Andrew Taylor_Cardiff City_Full Back
Number of Events Completed
Performance Metric
Wilfried Zaha_Cardiff City_Attacking Midfielder
Number of Events Completed
Performance Metric
Kenwyne Jones_Cardiff City_Striker
Number of Events Completed
Performance Metric
Craig Noone_Cardiff City_Winger
Number of Events Completed
Performance Metric
Fabio_Cardiff City_Full Back
Number of Events Completed
Performance Metric
Klevin Thleophile-Catherine_Cardiff City_Central Defender
Number of Events Completed
Performance Metric
Declan John_Cardiff City_Full Back
Number of Events Completed
Performance Metric
Steven Caulker_Cardiff City_Central Defender
Number of Events Completed
Performance Metric
x_coord
y_coord

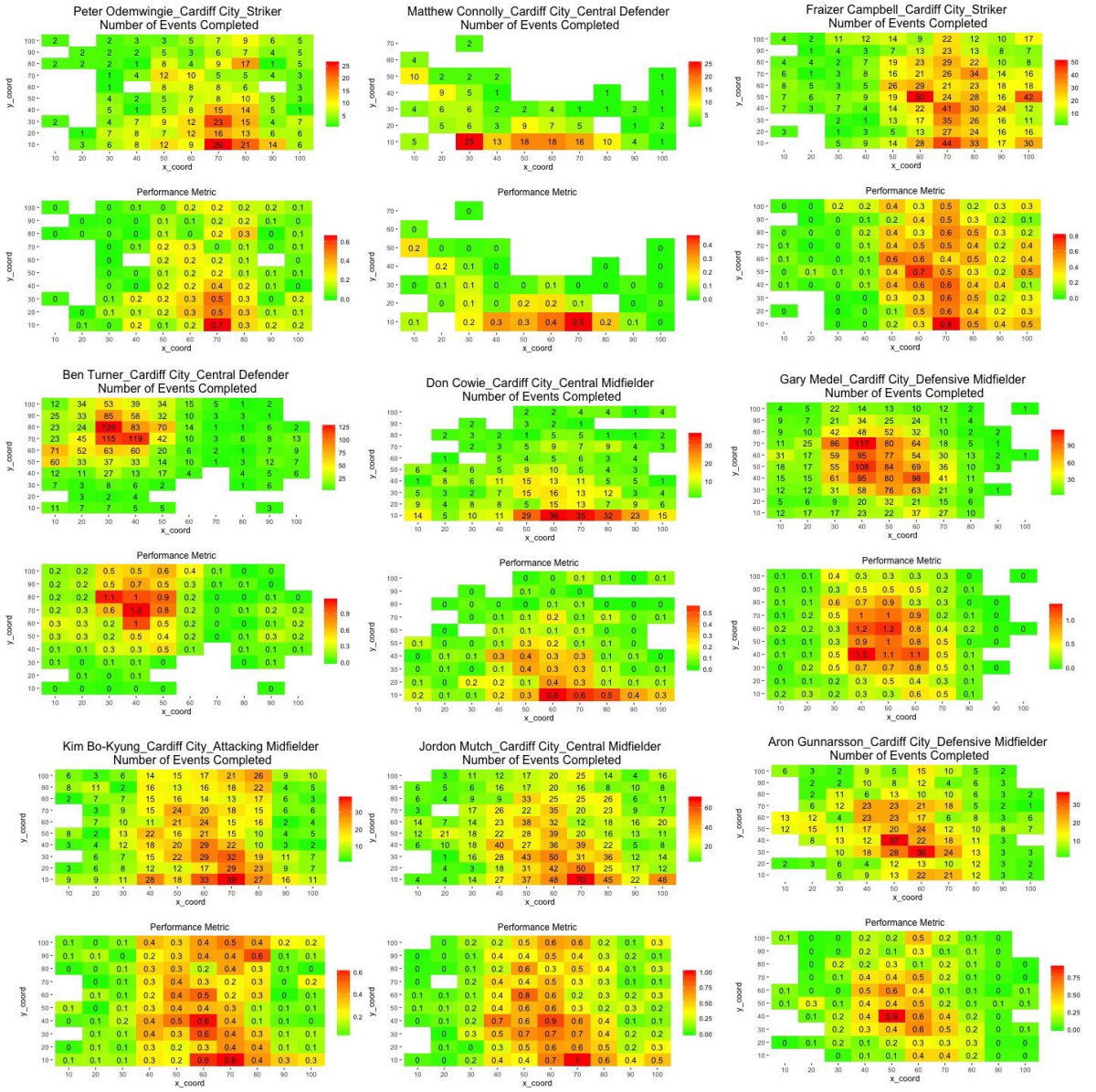


Figure 24 Heat maps of frequency (top heat map) and Performance Metric (bottom heat map) of all Cardiff City's players.

Similar results can be seen for some Cardiff City players but for others the model reveals that a different strategy may have been more effective. For example, the attacking midfielder Kim Bo-Kyung plays more on the right wing yet his Performance Metric gives him equal value on the left. Similarly, the striker Craig Bellamy is prone to a right wing attack but the model suggests varying this to include the left flank more may actually have provided more value. The dashboard outputs of these two players can be seen in Figure 25 and Figure 26.

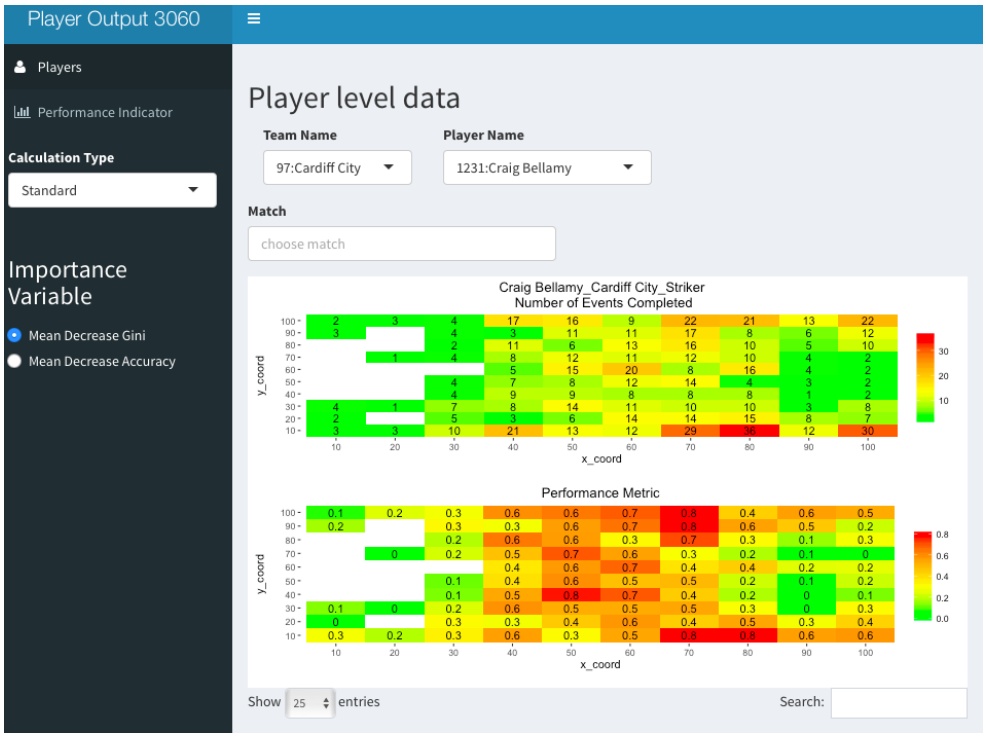


Figure 25 Heat map and Performance Metric for Cardiff City's Craig Bellamy taken from dashboard.

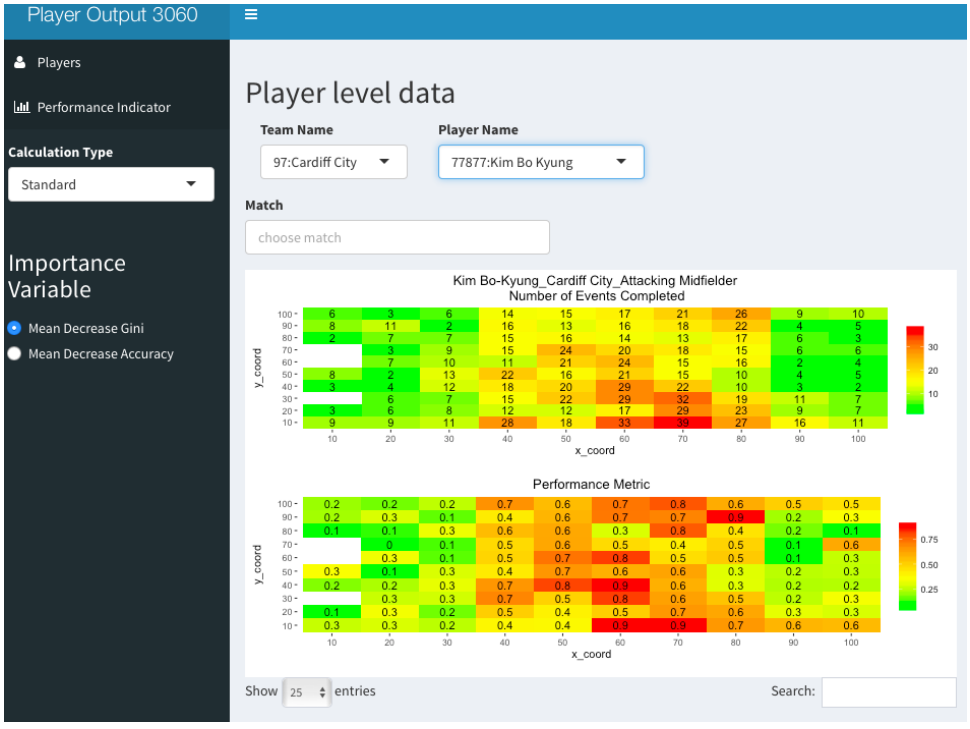


Figure 26 Heat map and Performance Metric for Cardiff City's Kim Bo Kyung taken from dashboard

4.4.2 Player Analysis

There were two key components to look for when comparing results: 1) overall high values of the Performance Metric and 2) areas of the pitch where the player performed the best. In an attempt to determine a player's optimum position points 1 and 2 must be taken into account. Interesting properties of the model were highlighted when looking at Liverpool F.C., the team that contained top goal scorer and player with the most assists over the season. Even though Luis Suarez was the top goal scorer and PFA Player's Player of the Year, he was only placed fourth on the aggregated Performance Metric, suggesting a possible negative bias in the model against strikers. This can be explained by the nature of the Performance Metric. Since the target variables are shots/goals and ball recoveries they were not given Importance Values and not used in the Performance Indicator. This means that the frequency of these on-the-ball events for each player was not taken into account in the final Performance Metric. Evidence that the model attributes greater value to the lead up to goals can be seen with Steven Gerrard, he was given second place, having created the most assists over the season.

Player ID	Player Name	Player Position	14:Liverpool
56979	Jordan Henderson	Central Midfielder	54.06
1814	Steven Gerrard	Attacking Midfielder	53.20
84583	Philippe Coutinho	Attacking Midfielder	47.69
39336	Luis Suarez	Second Striker	46.54
103955	Raheem Sterling	Winger	44.36
43191	Lucas Leiva	Central Midfielder	43.97
9047	Glen Johnson	Full Back	41.98
40555	Joe Allen	Central Midfielder	38.86
26793	Martin Skrtel	Central Defender	37.52
40755	Daniel Sturridge	Striker	37.41
91979	Jon Flanagan	Full Back	37.09
12450	Kolo Toure	Central Defender	32.52
40784	Mamadou Sakho	Central Defender	26.81
49013	Victor Moses	Attacking Midfielder	25.78
21094	Daniel Agger	Central Defender	25.01
44354	Aly Cissokho	Full Back	19.16
26725	Josle Enrique	Full Back	15.18
40270	Iago Aspas	Striker	15.12
82451	Luis Alberto	Attacking Midfielder	9.94
66797	Simon Mignolet	Goalkeeper	9.83
58786	Martin Kelly	Full Back	5.58
103912	Jordon Ibe	Winger	0.30

Table 17 Liverpool's overall Performance Metric values for the each player over the entire season.

Analysing Suarez's possession on the different areas of the pitch over the season showed that he was not only valuable in striking positions at the top of the pitch but performed equally well

in the midfield areas. It showed that his strengths near the goal were greatest when the ball was coming from the right, as shown by the higher Performance Metric value in coordinate (100,30). These are interesting results that could be used by Liverpool F.C. to ensure balls are played to that right area of the pitch more frequently as this is clearly his optimum position for scoring goals.

Comparing this to Wayne Rooney, the top goal scorer for Manchester United for the season, also provided some interesting findings. As mentioned in Section 2, Rooney has recently[10] been moved from a striker to midfield position. Interestingly the findings from this model would support the argument for such a move, with his Performance Metric showing higher value in the midfield of the pitch rather than at the top. This is evident from the darkest red values between 50 and 70 on the x-axis in Figure 27.

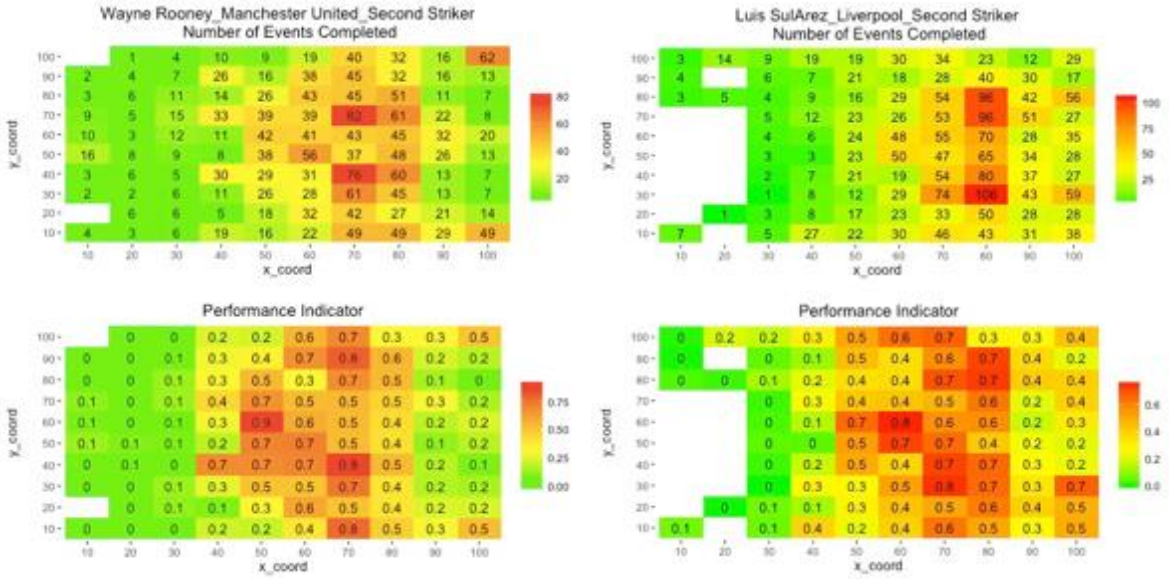


Figure 27 Frequency heat map and Performance Metric values for Wayne Rooney (LHS) and Luis Suarez (RHS) for the entire season

With the longest losing run of seven games, Crystal Palace was also an interesting team for analysis. Yannick Bolasie played predominantly on the left wing, as shown by the higher frequency values in the top heat map of Figure 28. However, the higher Performance Metric values shown on the right wing of the bottom heat map suggest he may have provided more

[10] 2016-2017 season

value if played on the right. Similarly, the full back Joel Ward spent most of his time as a right back but the model would suggest he showed more ability on the left or even midfield areas.

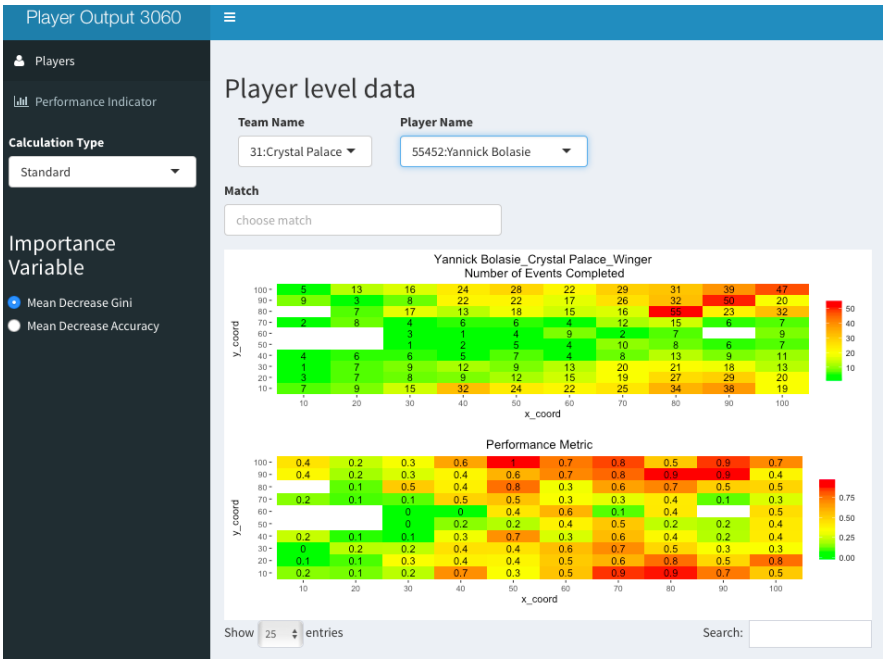


Figure 28 Yannick Bolasie's Performance Metric and frequency heat map taken from the dashboard

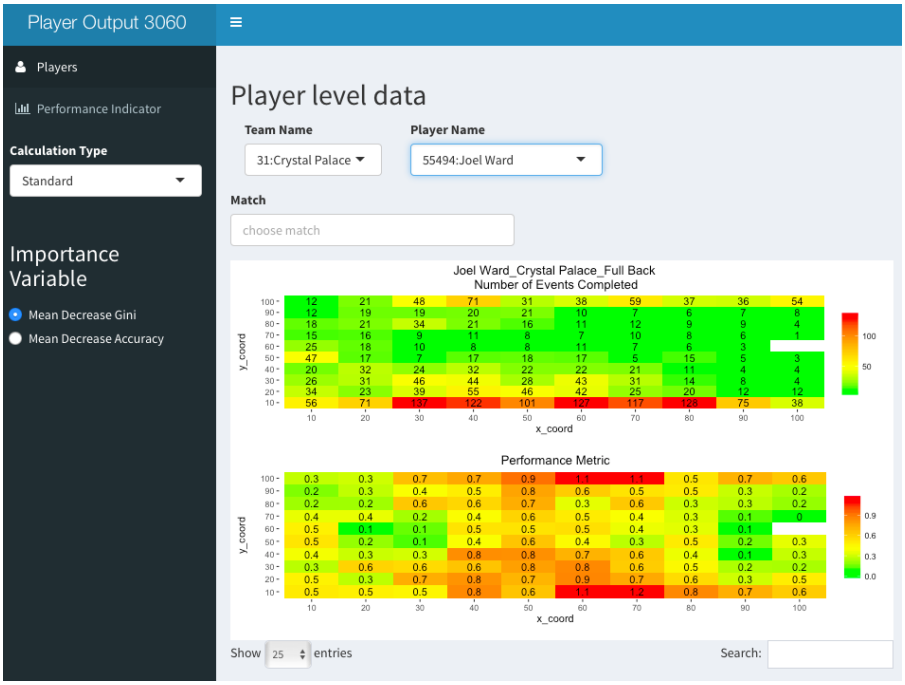


Figure 29 Joel Ward's Performance Metric and frequency heat map taken from the dashboard.

4.4.3 The Matches with Biggest Win-Loss Discrepancies

To understand if player positioning is a key factor in winning and losing games two games that saw the widest range in goals for and against each team were explored.

The aim of these heat maps was to compare where on the pitch a team had most possession relative to where the model would have played them. For the two winning teams shown below and on the following page, Manchester City and Chelsea, it was clear that the areas of highest possession were also the areas where the Performance Indicator placed them. By contrast, the losing teams, Norwich City and Arsenal, generally had sporadic areas where they performed well which did not have any correlation to the frequency of 'event_qualifier_ids' they played.

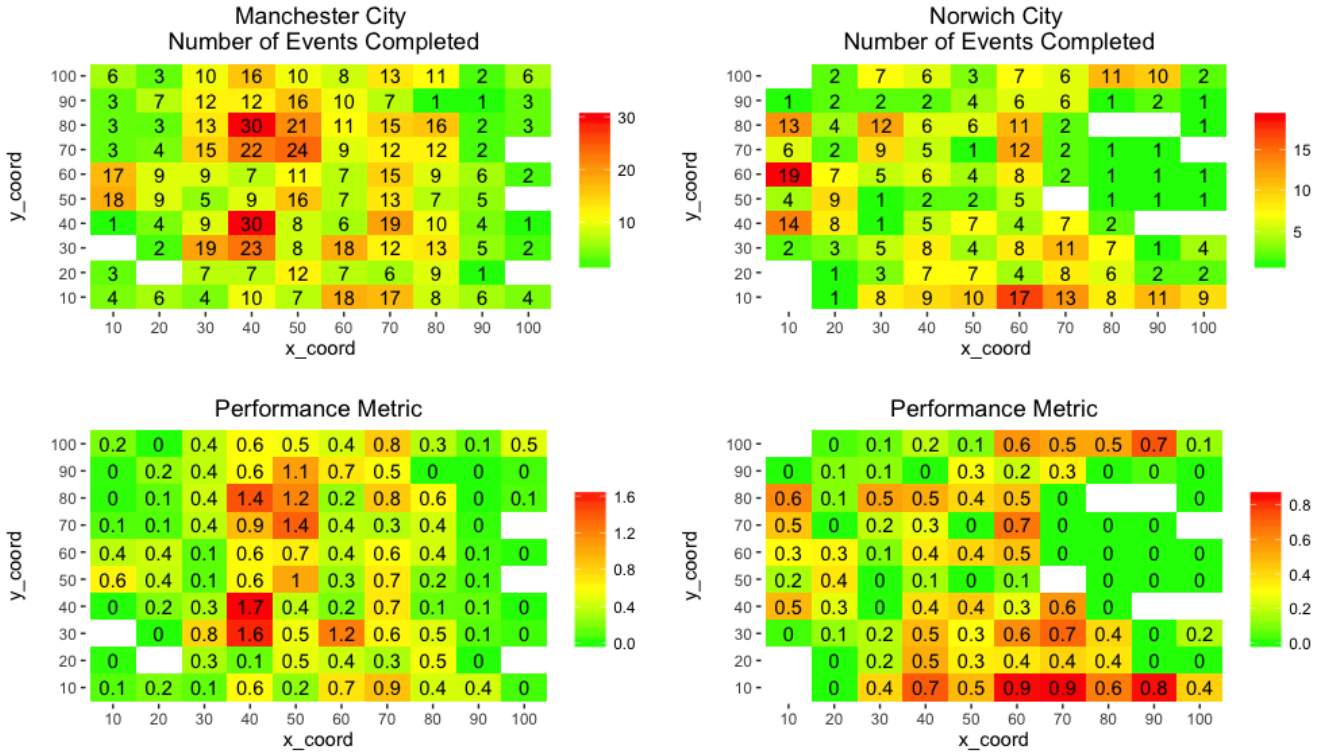


Figure 30 Frequency and Performance Metric heat maps of Manchester City Vs. Norwich City (7-0)

Manchester City in general show much higher Performance Metric values than Norwich City, especially in the middle of pitch (40,30 – 60,80) coordinate areas. Part of this will be due to the higher frequency of possession throughout the match. However, the areas where Manchester City had most possession were where they achieved the highest Performance Metric values. In contrast, relative to their possession on-the-ball, Norwich City could only get

a maximum Performance Metric of 0.9 shown for the wing which left the midfield area weak with values only reaching approximately 0.5. If presenting these findings to a football club professional, this model would suggest that since Norwich were unable to keep possession in the middle of the pitch they should utilise the wings where they were gaining the most results, possibly ensuring their strongest players were available here.

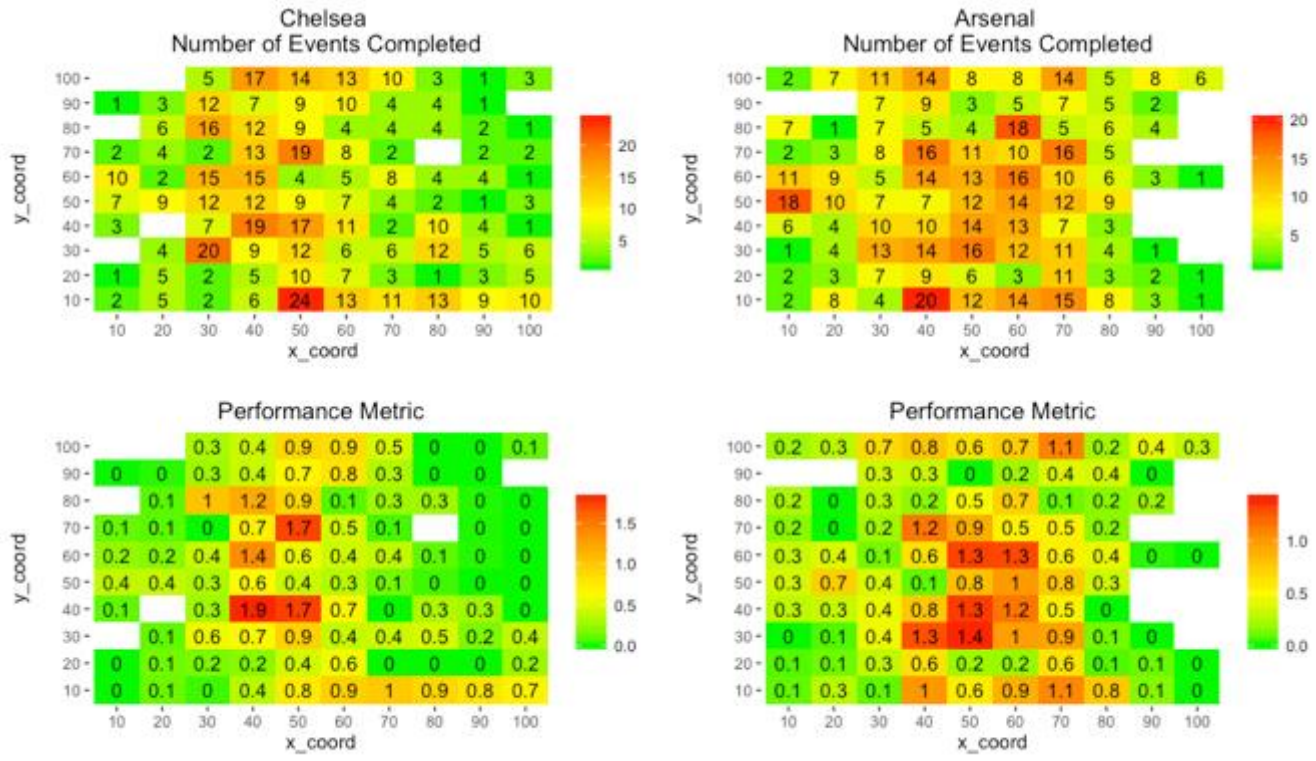


Figure 31 Frequency and Performance Metric heat maps of Chelsea Vs. Arsenal (6-0)

Six goals, one red card and a penalty would have made for exciting viewing but where did Arsenal go wrong in this thrashing by Chelsea? This was an interesting game for analysis as, unlike the other studied matches where possession was clearly one sided, in this game it seemed evenly spread. At first inspection it appeared that Arsenal had more value in the middle areas of the pitch with a larger group of dark red cells. However, their actual Performance Metric values were lower than Chelsea's in these areas. The heat maps show that Arsenal had most of their possession on the right wing but that this was clearly ineffective in producing value. The results of these analyses suggest that many of the Chelsea goals may have come from counter-attacking play rather than from passing and build up play leading to goals. From this Arsenal

should be advised to work on their defence in a bid to stop such counter attacks.

Breaking down both of these matches by player would provide further insight into who were the strong and weak links for each team as shown in Figure 32. However, in doing this for Arsenal it became clear how rarely most players actually touch the ball the during a game. The Numbers Game (Anderson & Sally, 2013) discussed Chris Carling's research showing that on average, players only have the ball for 53.4 seconds and only cover approximately 191 metres of the pitch during any game. This means that a player has the ball for less than 1% of the total time they are on the pitch with 98.5% of their movement done without the ball. This heavily suggests that the game of football is not just about what happens when a player has the ball but also what they are doing when they do not. To fully model the game effectively, off-the-ball event tracking should be considered. This is outside the scope of this research.

Figure 32 Frequency and Performance Metric heat maps of individual Arsenal players from Chelsea Vs. Arsenal (6-0). Evidence for the need of off-the-ball tracking as too few on-the-ball events available for analysis for one game.

5 – Discussion:

5.1. Objectives

5.1.1 Understanding the Characteristics of Different Footballers' Positions on the Football Pitch

The first objective of understanding the characteristics of football players' positions on the pitch was completed with the use of PCA and clustering analysis on the average count per game of completed on-the-ball events of each player. A test of the suitability of results was carried out by the inclusion of goalkeepers in the first stage of analysis. The obvious clusters of this position provided a suitable benchmark and confirmed that the average count of player on-the-ball events was an appropriate metric to use. However, experimenting with different metrics such as the ratio of on-the-ball events completed relative to areas on the pitch would be interesting to see if different results were obtained.

The overall objective results produced top level findings showing offensive and defensive players to have differing skillsets made evident by the difference in clusters produced from the output of PCA loadings. As more granular data was provided (i.e. The inclusion of 'qualifier_ids') the difference between the two groups of positions increased further.

In some cases, the 'event_qualifier_ids' were labelled in the dataset provided by Opta as offensive or defensive moves. However, in most cases this was decided subjectively by the author of this paper through knowledge of the game. For example, a 'pass_long ball' was assumed to be an attacking skill, but the results grouped it with the defensive events. This could suggest that plays such as these were not as effective in attacking as previous research would argue (Reep & Benjamin, 1968); implying most attempts at a 'pass_long ball' are used to remove the threat of the other team scoring rather than impose the threat of your own team scoring goals. It would be interesting to expand the dataset further than one season of football to understand if the style of the game has changed over the years.

Another key characteristic concluded from this section was that in general the time during the game that a player completed an on-the-ball event had no major effect on the difference in their skillsets, whereas the spatial coordinates of where they completed the event did. This was

shown by evaluating both the loadings and resulting clusters of differing space and time components. The lack of variation between 'event_qualifier_ids' for each time segment of the game may be in direct correlation to the high levels of fitness of top-flight players since their average on-the-ball event count does not appear to change over time. Repeating the analysis on Sunday league or semi-professionals whose fitness may not be quite as high would be a good method for testing this.

Completing this objective did provide interesting findings at an individual player level highlighting players who could be considered to behave differently to other players who play the same position as outliers. These findings could help clubs when valuing players based on their unique capabilities on the pitch. Although the time variable was dismissed as not being a useful positional attribute there were some interesting insights into striker performance such as Daniel Sturridge's distinctive qualities between 30 and 40 minutes or Andre Schurrle's between 60 and 70 minutes. This is information that could be directly used by clubs when deciding on game changing substitutions.

5.1.2 Finding the On-The-Ball Events that are Most Important in Defining a Player's Position

The second objective of finding the on-the-ball events that were most important in defining a player's position was completed with the use of Random Forests. This objective contained two parts. The first was to understand the skills that defined different positions and the second was to use these findings to reduce the size of the dataset. Using the output of the Random Forest in a traditional manner showed that central defenders and strikers had the most unique positions on the pitch. This was shown by the stronger predictive capabilities of the model for these positions. Initially, it was believed the reason for this was because a striker's skillset is different to other positions because they score more goals and defenders similarly perform more dispossessions and tackles. However, the use of the Random Forest to select the most important variables using the MDI gave passing and all variations of passing the highest importance when predicting players' positions on the pitch. To help strengthen these arguments these results were joined with heat map visualisations of the frequency of on-the-ball events in each of the 100 segments to choose the final variables that were believed most important when defining a player's position.

An interesting finding off the back of these visualisations was the clear symmetrical nature of the frequency of on-the-ball events. It would be interesting to explore this further to find out if there was a relationship between the symmetry of play and winning games or whether it is merely a side effect of chosen formations.

Cutting down the dataset from 134 'event_qualifier_ids' to 29 was important in maintaining a manageable set of variables for further stages of this research. For footballing coaches and managers the 'event_qualifier_ids' shown in Table 14 that were finally chosen are a good indication of what skills are important when choosing a player's position.

5.1.3 Finding the On-The-Ball Events that are Most Important Relative to Success

The third objective of finding on-the-ball events that were most important relative to success was completed using path analysis and Random Forests. Tackling the problem of the small number of relative successful events (such as goals) in football was overcome by converting the dataset into path-level events data. In this case shots/goals and ball recoveries were the skills chosen as successful events. Modelling every 30 seconds of the game as a unique row of data produced different paths which included overlapping on-the-ball events and increased the number of successful on-the-ball events per game. Taking the previous 60 on-the-ball events of each 30-second interval also assigned value to the lead up to 'success' with the hope of effectively modelling the continuous nature of the game. For example, if a 'pass_chip' event happens two moves prior to a goal in 50% of matches, a high proportion of the value assigned to the successful event (a goal in this case) would be given to the 'pass_chip'. This was not an approach that had been seen before with most similar work favouring the use of networks where players are the nodes and event types the edges (Duch, et al., 2010). Due to this unique methodology, testing its success in accurately modelling the game of football was difficult. However, the test for robustness was used to remove doubt that varying the main inputting factors, such as time interval and previous on-the-ball events, would have any large effect on the overall results. This is positive as the computational power needed to run these models was large. If implemented at a club where up-to-date findings were needed more frequently, it may be worthwhile looking at longer time intervals to reduce this complexity.

The choice of splitting the pitch into 10x10 segments was chosen in this instance as a solution

that would incorporate enough frequency of on-the-ball events in each segment without biasing towards any areas and was tested by visualising the frequency of on-the-ball events on a heat map. However, a number of papers have been written providing different views on this methodology. Using some of the results in this model would be an interesting area for further investigation (Gudmundsson & Horton, n.d.) (Lucey, et al., 2013) (Taki, et al., 1996).

The MDI outputs of the Random Forest models were used to create the final Performance Indicators. This was not a common technique and therefore needed considerable thought to understand the best way to use the results effectively. Since each of the 100 Random Forest models were calculated separately the range of results of the Importance Values varied. Therefore, one main aim was to ensure that all the values for each 'event_qualifier_id' were comparable across the 100 models. For example, the importance value for the 'pass_chip' in segment (10,10) of the pitch may have been 100 whereas the importance value for the 'pass_chip' in segment (90,10) may have been 0.9. Without knowing the values of the other 'event_qualifier_ids' in each segment we were unable to draw any conclusions from these values alone. This is because although a higher absolute value the 'pass_chip' may have had the lowest importance in the (10,10) segment relative to the other 'event_qualifier_ids' and the highest importance in the (90,10) segment relative to the other 'event_qualifier_ids'. Many different attempts were made to transform these MDI values to understand their actual importance relative to each other so that all 100 models would be comparable. In the end, standardising all values to between 0 and 1 was chosen with the aim of maintaining maximum interoperability.

Research was also done into the best way of aggregating all the 'event_qualifier_id' Importance Values when trying to find an overall performance metric for each player. In doing this calculation there was evidence to suggest that because the MDI method relies on splits at each node for its calculation, it was important when summing these values to include a weighting of the ratio of these splits $\frac{\sum nodes\ involving\ a\ split\ on\ \ variable\ X}{\sum nodes\ \ involving\ a\ split\ on\ all\ variables}$. However, by the time the summation across events was completed the initial Importance Values had already been manipulated to such a degree that there did not appear to be any value in doing so.

In the creation of the path-level data as the input for the Random Forest models, shots/goals

and ball recoveries were used as target variables. This meant they were not included as input variables therefore not producing an Importance Value for the Performance Indicator. The effects of this may have been unfair for strikers since their defining skills are shots and goals. One idea of accounting for this was to add a measure of number of goals scored. However, this may have had the opposite effect, resulting in an unfair advantage in favour of strikers. So in trying to keep the model as equal for all positions this was decided against.

5.1.4 Finding a Player's Optimum Position on the Pitch

The aim of this paper was to delve deeper into conventional positions (defender, midfielder, striker) adding a quantifiable measure to how players have been labelled relative to their position on the pitch in the hope of finding where *within* these positions players perform their best. For example, do some strikers perform better outside the box or do some central midfielders perform better on one side of the central areas of the pitch? These questions were answered with the help of the final objective; finding the player's optimum position on the pitch.

As a starting point there was a clear clear indication that top-flight football players have a position and stick to it. This was most evident for full back or wing players, with their frequency heat maps showing near to no possession in other areas of the pitch, as shown in Figure 33.

Figure 33 An example of the possession of a full back heavily weighted towards the left wing

Because of this it was important to understand that the results of this study were not to show for example, that a striker should be a defender, but instead that the broad label of a striker can be broken down to understand where within the striking position that player shows the most ability. Saying this if a player has played in multiple positions the model has the ability to pick up where it believes they played the best. This was evident when looking at wingers such as Yannick Bolasie who had clearly been used on both the left and right wing. The model favouring playing him on the right.

Selecting a couple of individual players over the season provided some key recommendations like this for footballing professionals. It was found that Suarez performed better as an attacking striker on the right and Rooney was stronger in the midfield areas of the pitch. These were just two recommendations exemplifying of the power of the model. The high dimensionality of the results required the creation of an interactive dashboard which would enable footballing professionals to filter through the Performance Metrics and find similar conclusions about their own players. When using the heat maps created to compare multiple players/teams it was important to take into account that different segments of the pitch were given different colours dependent on the highest (red) and lowest (green) values for that player/team only. They were not relative to the highest and lowest of all players/teams combined. For example, a player with a generally low Performance Metric may see more dark red areas than a player with on

average a higher Performance Metric. It is important therefore to make sure the values as well as the colours are taken into account. If this study were to be done again this may have been something that would have been changed to allow for easier interpretation of the results.

A main challenge in this paper was the method used to relate the Performance Indicator back to each player, the issue being that the frequency of 'event_qualifier_ids' in some cases overrode the indicator of performance. Considerable experimentation was done to find a metric that would account for this but since it is a unique problem there is always room for further research and optimisation. The results show enough variation between the frequency heat maps and Performance Metrics to make reasonable conclusions but there is still evidence of a slight bias towards the frequency of on-the-ball events.

5.2. Answering The Research Question

Can Data Science techniques be used to help understand a player's skillset and find their optimum playing position?

The aim of this paper was to help understand whether data science techniques could be used as an explorative tool to help understand a football player's skillset and aid football professionals in finding a player's optimum position on the pitch. Although there will always be subjectivity in sport, this work aims to remove the reliance and pressure on a small group of individuals in the hope of producing better results for the team. There were definite outputs that could be taken from this paper and, with more data incorporated into all stages of this research the results, and therefore the outcomes will grow. It is clear from the outcomes discussed that the use of data science techniques can definitely help explorative research in the area of football specifically relating to the positioning of players.

6 – Evaluation, Reflections and Conclusions:

6.1. Overall

This study indicates that players are given certain positions to maintain throughout the game as described by Wayne Rooney (Taylor, 2016) and evidenced by the high-frequency on-the-ball events of players in certain positions, e.g. heavily weighted towards the left wing. However, using the framework that this research paper has established a Performance Metric that is not reliant only on the frequency of on-the-ball events has been created. It shows when high values are revealed in low frequency areas that the player actually performed better in that segment of the pitch, suggesting they may be better suited to playing in that area. This study not only uses the large complex dataset to present cogent arguments to academics, but the creation of the dashboard allows those arguments to be presented very clearly to non-academics and footballing professionals.

Overall the results of this work have been positive in producing a metric that allows footballing professionals to add a quantifiable nature to their positional choices. The key objectives were outlined in the proposal and maintained throughout the paper. The use of this structure was invaluable in ensuring the research question was answered.

The choice of dataset was detailed and took considerable time to understand and manipulate into a suitable structure for use. This meant that the initial stages of the project that imported the XML files and created SQL databases of the match-by-match events were critical. Having seen the dataset in the lead up to the project it was known that this would be the case and therefore time was suitably adjusted for in the project plan.

Creating path-level databases to model the game created difficulties in relation to the length of time taken to run the models. Larger datasets than first anticipated called for the use of virtual machines so multiple models could be run in parallel. Creating dynamic code for this environment that could run any model with any time interval or previous event combination was extremely useful in later stages when certain problems became apparent that demanded the model be re-run, such as realising that the 3010 model did not suffice and 3060 model would have to be calculated.

The main methods used in this paper were PCA, clustering and Random Forests. These differed slightly from those discussed in the project proposal which also included unsupervised classification and self-organising maps (SOM). SOM's were experimented with when trying to understand the characteristics of football players on the pitch. However, due to the nature of the metric used as input (average count of completed on-the-ball events per player) they did not appear to add any value and so were excluded from the paper. Using the Random Forest Importance Values to create the Performance Metric was a novel use of these output values which meant there was no similar work to compare to. Much experimenting with how best to transform the outputs into a usable form was therefore required. Importance Values are more generally used for variable reduction so it was really exciting to experience the results of this unique approach.

This paper is very much an ongoing research problem and labelling its findings as successful or not was a difficult thing to do. It was decided to compare the team that finished top to the team that finished bottom of the 2013-2014 Premier League to try and understand whether the results showed more players out of position in the worst performing team. This provided a couple of recommendations, such as Cardiff City understanding the value in their winged play as well as considering both Craig Bellamy and Kim Bo Kyung play more predominantly on the left flanks. With approximately 20 players in each squad for the season there were over 400 players available for analysis. This meant that the creation of the dashboard was crucial in providing a simplified method for viewing them all.

The large size of the dataset meant that relating index values such as 'game_id' to the actual names of the teams that were involved in the match was time consuming. Although an index of all 'event_qualifier_ids' was created which described what each number meant it would have been useful to expand this idea to include index tables for all the key indices, such as 'team_id'. This would have saved significant time in the latter stages of interpreting results. There were many times that similar code was used for calculations, such as creating the average event count per player metric. If I were to start again I would have created functions that could be called, rather than having to search through each script or in some cases re-write the code. A key taking from completing this work is how important it is to keep all of the data, code and outputs in a tidy and easily accessible format.

Modelling football with data is a complex problem with an infinite number of methods and decisions needed. However, the findings in this paper show that the use of data science techniques can help to provide value that may not be available through other means. The ability to now model the continuous and complex nature of the game paired with the results of this research suggest that the standard labelling of players' positions may no longer suffice. Instead there is a need for a more granular understanding of the exact areas of the pitch the players perform the best.

6.2. Proposals for Further Work

6.2.1 Data Expansions

This paper looks at one season worth of Premiership data and a key proposal for further work would be to expand this to include other tournaments as well as other seasons. It would be interesting to understand whether the key on-the-ball events relative to choosing a player's position has changed over time. There were also a number of results which may be unique to professional footballers, such as the time of the game not being an important factor in defining player's positions. It was concluded this may be due to the high levels of fitness of professional footballers and would therefore be interesting to extend this study to include semi-professionals or Sunday League players.

With women's football becoming more popular it would also be interesting to compare the results to see if there were any key similarities or differences between the two sexes.

One issue with the model was that most players at top level only play in one position, meaning they carry out most of their on-the-ball events in the same areas of the field. This creates a slight problem as it is then difficult to understand if they might perform better in other positions. This was evidenced by the model performing best on midfielders as they cover the most ground. An interesting extension to this research would be to run the model on younger players that are still unsure of their position.

This paper only looks at on-the-ball events with the results in the section 4.4.3 clearly showing that compared to the length of the game the amount of time any player is actually on-the-ball is minuscule. Exploring the use of tracking devices to try and incorporate an off-the-ball

element would be a really interesting and no doubt extremely useful development to this study.

Since the nature of this research was to understand players' skillset, the dataset of 'event_qualifier_ids' were reduced to not include off-the-ball events such as weather etc. It may be interesting to run another model that would incorporate these events to see if they have any direct effect on the game.

6.2.2. Performance Metric Extensions

There were three key variables which largely dictated the output of the Performance Metric. It would be interesting to compare and contrast the results of changing these variables, as well as help test the robustness of the model.

The choice of 10x10 segments – This was chosen to try and model as much of the continuous nature of the game as possible. Decreasing the size of the squares would create an even closer model to the continuous elements of the game. However, the computational complexity would increase substantially as the size of the segments decreased. Breaking down the areas by other shapes may also be an answer, as has been done previously (Gudmundsson & Horton, n.d.).

The model used to create the Performance Indicator – This paper only looks at the MDI outputs from the Random Forest model. It would be interesting to look for different methods other than Random Forests that could provide values for the importance of variables and compare the results to try and understand the generalisability and dependency of the model.

The method used to create the Performance Metric which relates the player back to the Performance Indicator – as discussed this was a difficult problem which took some experimentation before finally choosing the end result. Experimenting with different methods may provide further insight into other solutions.

7 – Glossary:

7.1. Technical Terms

Word/Phrase	Definition
Performance Indicator	Bespoke measurable value that shows how important an on-the-ball event is in one area of the pitch.
Performance Metric	Bespoke measure that shows where on the pitch a player's optimal position is.
Clustering	Grouping data points with similar properties
Variable Importance Values	Value showing the importance of variables from a Random Forest model
Precision	$\frac{True\ Positives}{True\ Positives\ +\ False\ Positives}$
Recall	$\frac{True\ Positives}{True\ Positives\ +\ False\ Negatives}$
PCA Loadings	Weight by which individual scores should be multiplied to get component value.

7.2. Football Terms

Word/Phrase	Definition
Top-Flight Team	Playing at the highest level possible
Break	To counter attack
Clean-sheet	Not concede any goals

8 – References:

Anderson, C. and Sally, D., 2013. *The Numbers Game: Why Everything You Know About Football is Wrong*. Penguin Books Ltd.

Bojko, A., 2009. Informative or Misleading? Heatmaps Deconstructed. *Human-Computer Interaction*. *New Trends,* 5610(1), pp. 30-39.

Breiman, L. 2001. Random Forests. Machine Learning, [Online] Volume 45(5), p.5-32. Available at: http://link.springer.com/article/10.1023/A:1010933404324 [Accessed October 2016]

Bull, J., 2016. *How Claudio Ranieri's tactics put his rivals to shame at Leicester City*. [Online] Available at: http://www.telegraph.co.uk/football/2016/04/30/how-claudio-ranieris-tactics-put-his-rivals-to-shame-at-leiceste/ [Accessed November 2016].

Charles, N., 2016. *Find me a player like Andrés Iniesta: Part 2*. [Online] Available at: http://www.hilltop-analytics.com/football/opta-conference-2016-find-me-a-player-like-andres-iniesta-part-2/ [Accessed October 2016].

Chheng, T. 2013. RMongo: MongoDB Client for R. R package version 0.0.25. Available at: https://CRAN.R-project.org/package=RMongo

Duch, J., Waitzman, J. S. and Amaral, L. A. N., 2010. Quantifying the Performance of Individual Players in a Team Activity. [Online] Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0010937 [Accessed November 2016]

Ducker, J., 2016. *Jose Mourinho insists Wayne Rooney won't play in midfield for Manchester United next season: 'He will never be a No 8*'. [Online] Available at: http://www.telegraph.co.uk/football/2016/07/05/jose-mourinho-insists-wayne-rooney-wont-play-in-midfield-for-man/ [Accessed 12 November 2016].

fourfourtwo, n.d. *wenger-how-improve-your-5-side*. [Online] Available at: http://www.fourfourtwo.com/performance/skills/wenger-how-improve-your-5-side [Accessed October 2016].

Google, n.d. *Google Cloud Platform*. [Online] Available at: https://console.cloud.google.com/

Gudmundsson, J. & Horton, M., n.d. *Spatio-Temporal Analysis of Team Sports – A Survey*. [Online] Available at: http://arxiv.org/pdf/1602.06994v1.pdf

Jefferson, L., 2016. Paul Pogba Joins Manchester United For World Record 89m. *Sky Sports News*. [Online] Available at: http://www.skysports.com/football/news/11667/10528626/paul-pogba-joins-

manchester-united-for-world-record-89m
[Accessed October 2016].

Jolliffe, I., 2002. *Principal Component Analysis, Second Edition*. Second ed. Aberdeen: Springer.

Lewis, M., 2003. *Moneyball: The Art of Winning an Unfair Game*. s.l.:Norton.

Louppe, G., Wehenkely, L., Sutera, A. and Geurts, P, 2013. Understanding variable importances in forests of randomized trees. Belgium: s.n.

Lucey, P. et al., 2013. Assessing Team Strategy using Spatiotemporal Data.

Mike Hughes, I. F., 2005. Analysis of passing sequences, shots and goals in soccer. *Journal of Sports Sciences,* 23(5), pp. 509-514.

Oliver, D., 2005. *Basketball on Paper: Rules and Tools for Performance Analysis*. Washington: Potomac Books, Inc.

OptaPro, n.d. *Opta Pro*. [Online] Available at: http://www.optasportspro.com/
[Accessed 30 October 2016].

Ourti, T. and Clarke, P. 2011. A Simple Correction to Remove the Bias of the Gini Coefficient due to Grouping.*The Review of Economics and Statistics* [Online] Volume 93(3), p.982-994. Available at: http://www.mitpressjournals.org/doi/abs/10.1162/REST_a_00103?journalCode=rest#.WD0gB6KLSu4 [Accessed November 2016]

Pollard, R. and Reep, C., 1997. Measuring the Effectiveness of Playing Strategies at Soccer. *Journal of the Royal Statistical Society*. *Series D (The Statistician),* 46(4), pp. 541-550.

R Core Team 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Available at: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/prcomp.html Vienna, Austria. URL https://www.R-project.org/.

Reep, C. and Benjamin, B., 1968. Skill and Chance in Association Football. *Journal of the Royal Statistical Society*. *Series A (General),* 131(4), pp. 581-585.

Sandri, M. and Zuccolotto, P., 2008. *A bias correction algorithm for the Gini variable importance measure in classification trees,* Brescia: s.n.

Smith, R., 2015. *Bespoke data gives Arsene Wenger the tactical edge at Arsenal*. [Online] Available at: http://www.thetimes.co.uk/tto/sport/football/article4638099.ece
[Accessed November 2016].

Sports, O., n.d. [Online]
Available at: http://www.optasports.com/services/media/data-feeds/opta-data-feeds-overview.aspx

STATS, n.d. *Sports Data Company*. [Online]
Available at: www.stats.com [Accessed October 2016].

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 25 January.8(25).

Taki, T., Hasegawa, J. and Fukumura, T., 1996. Development of motion analysis system for quantitative evaluation of teamwork in soccer games. *In Proceedings of 3rd IEEE International Conference on Image Processing*, September, Volume 3, pp. 815-818.

Taylor, D., 2016. *Wayne Rooney says Allardyce left him 'battered' with carte blanche remark*. [Online]
Available at: https://www.theguardian.com/football/2016/oct/04/wayne-rooney-sam-allardyce-england-left-me-battered [Accessed November 2016].

Vincent Q. Vu (2011). ggbiplot: A ggplot2 based biplot. R package version 0.55. [Online]
Available at: http://github.com/vqv/ggbiplot

Wikipedia, n.d. *2013-2014 Premier League*. [Online]
Available at: https://en.wikipedia.org/wiki/2013%E2%80%9314_Premier_League
[Accessed October 2016].

Appendix A – Project Proposal

Project Proposal: Data Driven Techniques to Support Football Coaches in the Assessment of Players' Skill Sets

INTRODUCTION

The aim of this paper is to understand whether data science techniques can be useful in providing football coaches and managers with additional information to help assess their player's skill sets. The data provided by Opta sports (Opta, n.d.) includes the full database of premiership data for the 2013/2014 season. This will be analysed using both supervised and unsupervised learning methods and a variety of feature selection and pattern detection techniques. The techniques employed in this project are chosen for their suitability and ability to gain as much useful information from the dataset as possible.

Quantitatively analysing football matches has always been a more challenging task than in other sports such as baseball and American football. One reason behind this was lack of data and an inability to measure success on anything apart from goals. Now top sport statistics companies such as Opta (Opta, n.d.) are able to record and store high level time series data down to x,y coordinate level (Opta, n.d.) and with the growing success of data science techniques such as artificial intelligence, machine learning, dimensionality reduction and feature selection there is now, more than ever, scope for merging these techniques to enable football analytics to grow.

Research question

Football at top levels is a sport with many vested interests in both a monetary and, as one of the most popular sports in the world, an emotional sense. "*In the top leagues, it is all about the last five percent. Technical skills and athletic performance have probably reached their peak…. But analytics help stretch beyond that limit. We can tap into new potentials*." (Website, n.d.). Any advantage that can be gained is critical in winning the top flight leagues. This leads to the question:

How can Data Driven Techniques Support Football Coaches in the Assessment of Player's Skill Sets?

Objectives

- Understand the characteristics of different footballer's positions on the football pitch.
- Group players into positions dependent on their playing style.

- Research suitable performance metrics, either *externally* using known metrics or *internally* creating new metrics to allow ranking of players.
- Produce a classification system that categorises a player into their optimum playing position using their football characteristics.

Outcomes

- An understanding of the most crucial features that make a successful player.
- An unsupervised method of grouping players by their characteristics on the pitch into playing positions.
- Performance metric to rank players.
- A supervised classification method to categorise players into their optimum positions.

Beneficiaries

Player coaches – The Numbers Game (Anderson & Sally, n.d.) describes how approximately 50% of match results come down to chance. The aim here is to make sure that the other 50% of results can be optimised with the help of efficiently using large spatio-temporal data sets. If coaches can understand the features of a player that are most important when playing certain positions, they can also use this as a training tool and focus on improving these skills.

Team managers – With performance analytics becoming more important, coaches and managers are analysing these statistics much more frequently. Arsene Wenger was recently interviewed after moving Aaron Ramsey to a central midfield position. This is something the player himself had openly been pushing for. However, Wenger clearly stated that it was the data not Ramsey's persistence which got him the move: "*If you look at his Expected Goals when [Ramsey] is in a central position, it is among the best in the Premier League*" (Wenger, n.d.).

Scouts – A player's strengths could more easily be recognised when trying to purchase a player for a certain position. On the flip side it would also be an advantageous tool for scouting the opponent's players to understand the strengths and weaknesses of the opposition.

Gambling companies - Gambling and betting companies would be interested in this research as the performance metric would allow them to amend odds dependent on the predicted results.

CRITICAL CONTEXT

The Need for Data Driven Approaches in Football

Football Analytics

Numerically analysing football is not a new concept. Half-time and post match analysis are now full of statistics of individual as well as team performance. Shots on target, possession and corners are all examples of fixed events commonly used in football analysis. These basic statistics, although useful in some senses, rarely provide enough information to accurately predict the correct result for the team or overall performance of the player. In terms of data analytics football is miles behind other sports such as baseball, ice hockey and American

football. The book *Moneyball* (Lewis, 2003) started a new craze of sport analytics, showing how successfully value can be added to baseball by data analytics. Danny Willett's golf Masters victory is another example of how data can be used to help sporting achievements. During his public thank you tweet he gave a shout out to 15th Club for their involvement in preparing for the course strategy. 15th Club, a British analytics company, state they *"help professional golfers win by applying intelligence and context to performance data*" (Ingle, 2016). In *The Numbers Game* (Anderson & Sally, n.d.) Mark Brunkhart, founder of Match Analysis, a football data collection company (Analysis, n.d.), is quoted as saying "*In comparison to historical medicine, football analytics is currently in the time of leeches and blood lettings*" (Anderson & Sally, n.d.). There are three main reasons for this: the first is the complexity of the play and the continuous movement of the ball (Duch, et al., 2010); the second is the inability to quantitatively measure success of individual players and the third is the lack of performance labels available to describe the captured events (Lucey, et al., n.d.).

Performance Metrics

One of the key problems when analysing football is the lack of frequent performance metrics. Duch et al. (Duch, et al., 2010) discuss how the play of the game, not just the 'simple statistics', provides the true measure of performance. There has been previous research into methods of doing this. Duch et al. (Duch, et al., 2010) created a performance metric that uses networks with players as the nodes and number of passes between each player as the arc weights to look at the relationship between players. Lucey et al. (Lucey, et al., n.d.) chose to create labelled events using trained analysts to reflect levels of success (Lucey, et al., n.d.).

Data

Understanding Spatio-Temporal Data

Now that computer vision and tracking methods are available and companies such as Opta (Opta, n.d.) and Prozone (Prozone, n.d.) record detailed event data the problem is no longer that there is not enough data but instead how to correctly use it. Lucey et al. (Lucey, et al., n.d.) state that at the time of writing the article (2013) there were no successful methods of "*utilising spatio-temporal data in continuous sports*". They agree that being able to deal with noisy data, draw information via unsupervised techniques and then predict future performances would be an invaluable asset to coaches. The real advantage to spatio-temporal data is that it not only allows us to understand that an event happens but also where and why (Lucey, et al., n.d.). Although very useful, the data mining of spatio-temporal data is non-trivial. The Andrienkos (Andrienko & Andrienko, n.d.) propose a method of dealing with spatio-temporal data which they call Exploratory Data Analysis (EDA). This method has three main stages: viewing the data set as a whole to gain an overview, dividing into sub categories to understand individual elements and finally looking more deeply at interesting properties such as outliers. The tools they describe as best for doing this are visualisations, aggregation, filtering, marking and data transformations. They argue that a combination of all of these gets the best results (Andrienko & Andrienko, n.d.). These initial analyses allow for preliminary insights into the dataset to make sure it is suitable for the more complex computations to be carried out (Andrienko & Andrienko, n.d.). Three methods argued to be the most effective for gaining knowledge from spatio-temporal data are pattern recognition, clustering and predictive modelling (Andrienko, et al., n.d.). Pattern recognition is used to understand common events occurring within the data, clustering is used to highlights key groups and their properties; and predictive modelling, also

known as classification, is used to discover certain behaviours with the aim of predicting future events (Andrienko, et al., n.d.).

Machine Learning

Principal Component Analysis

Principal Component Analysis is a popular statistical method for transforming a large set of variables into a significantly smaller set of uncorrelated variables, whilst still keeping as much information as possible. It works by choosing the principal components which account for the highest variance in the dataset (Dunteman, 1989). This is a useful tool since working with a smaller set of variables is much more manageable for more complex analyses (Dunteman, 1989).

Clustering

Clustering aims to create groups of similar points from a given dataset (Likasa, et al., 2003). The most popular method uses the k-means algorithm which assigns each point in the dataset to a cluster k with the nearest mean. K-means is an efficient clustering technique. However because of its local search method the end result is highly dependent on the initial starting conditions (Likasa, et al., 2003). Different techniques can be adopted for the clustering of time series data, with the chosen method dependent on whether they work directly on the raw data, or indirectly on derived features or models created from the raw data (T. Warren Liao, 2005).

Self-Organising Maps

Self-organising maps are unsupervised neural networks which map multi-dimensional datasets to two-dimensional form in order to allow the complex data to be easily visualised. They are useful for both classification and clustering and are popular because of their ability to highlight unique variables in the dataset (Malone, et al., 2005).

Unsupervised Classification

Restricted Boltzmann Machines (RBM) are two layer undirected neural networks that can be used for both supervised and unsupervised classification (Salakhutdinov, et al., n.d.). Unlike Boltzmann Machines, RBM's have no connections between the hidden layers, this allows for speedier and more efficient algorithms to be used. They are trained using a gradient based algorithm named "*contrastive divergence*" (Salakhutdinov, et al., n.d.).

Supervised Classification

Decision Trees (classification trees) are a method of predictive modelling. The leaves represent target variables and the branches are conditions that need satisfying to reach these variables. In general Decision Trees are an efficient method of classification. However they can sometimes grow to be so complex it is hard to interpret their meaning.

Decision Making

Visual Analytics

Due to the complexity of the dataset, visualisations will be important not only to display results but also during analysis to gain a deeper understanding of its features. Visual analytics researches the benefits of merging "*cognitive reasoning and domain knowledge*" (Sacha, et al.,

n.d.) from the human, with the ability to store and compute mass data from the machine in the hope of producing effective visualisations that reveal information that may have not been found by other methods (Sacha, et al., n.d.). Sacha et al. (Sacha, et al., n.d.) highlight the importance of human and computer interaction in their knowledge generation model which shows that employing a visualisation system helps the user to work iteratively, exploring and verifying their findings in the hope of improving the results on each iteration. An example of a visualisation can be see in figure 1. This heat map shows the on the ball events of the Arsenal team normalised from left to right in a game in the 2013/2014 season.

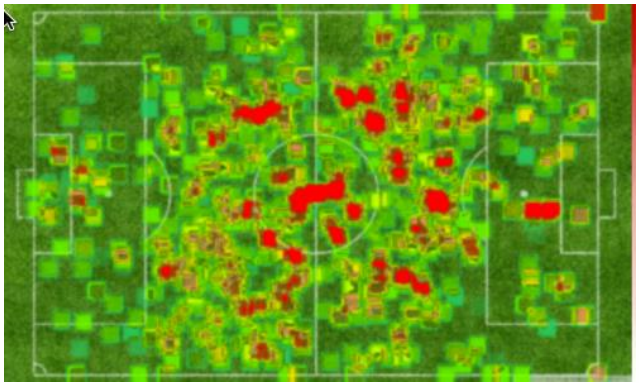


Figure 34: Visualisation created in processing of on-the-ball events of the Arsenal team normalised from left to right of a game in 2013/2014 season.

APPROACHES: METHODS & TOOLS FOR ANALYSIS & EVALUATION

Gathering Data

The data for analysis has been provided by Opta Pro (F24 package) (Opta, n.d.) in XML format. The dataset includes 20 teams and 380 matches. The F24 feed lists all player action events within the game with player id, team id, event type, minute and second as well as a vast number of qualifiers describing each event. There are 65 events available for each timestamp and 229 qualifiers that could correspond to each event.

Parsing Data

The first phase of the project will involve parsing this data from XML format into csv using R. From here the individual match data can be imported into Sequel Pro so that it can be easily queried. This is important as queries will be done to process the data into a form that can be used in future calculations. This will be necessary to answer the project objectives; for example when analysing the data on a player by player level rather than by match. The dataset is vast, including x,y positional information and time-stamps on every on-ball event (Opta, n.d.). To make sure that the dataset is fully understood an index of event types and qualifiers will be made that can be used as a lookup table when necessary.

Obtaining Features

The vast nature of the raw data means the next stage will be spent deriving features from the dataset. To find the important characteristics of players relative to their positions the data will be manipulated into variables on a player-by-player basis. The key variables of interest (averaged by game) will be completed passes; assists; successful tackles; area of the pitch

covered; amount of on the ball possession; crosses and goals. As well as these a level of more specific features will be included. These will include offensive/defensive/sideways passes; aerial/ground duels won and lost; headers; clearances and shots on/off target. This will create as many features as possible from the dataset so that the algorithms used in later methods can work on as much data as possible. To account for the spatio-temporal nature of the data these features will also be accounted for dependent on when and where they happened. To do this the pitch will be split into regions and the match time into groups (eg. 0-15 mins, 15-30 mins and 30-45 mins). There has been some research done into the best way of splitting a pitch for analysis (Gudmundsson & Horton, n.d.). Some methods involve an equal splitting (Lucey, et al., n.d.) whereas another common approach is to look at the player's dominant regions. "*The dominant region of a player p is the region of the playing area where p can arrive before any other player*." (Taki, et al., 1996). Researching and experimenting with both will provide evidence of the most effective way to use the spatio-temporal data.

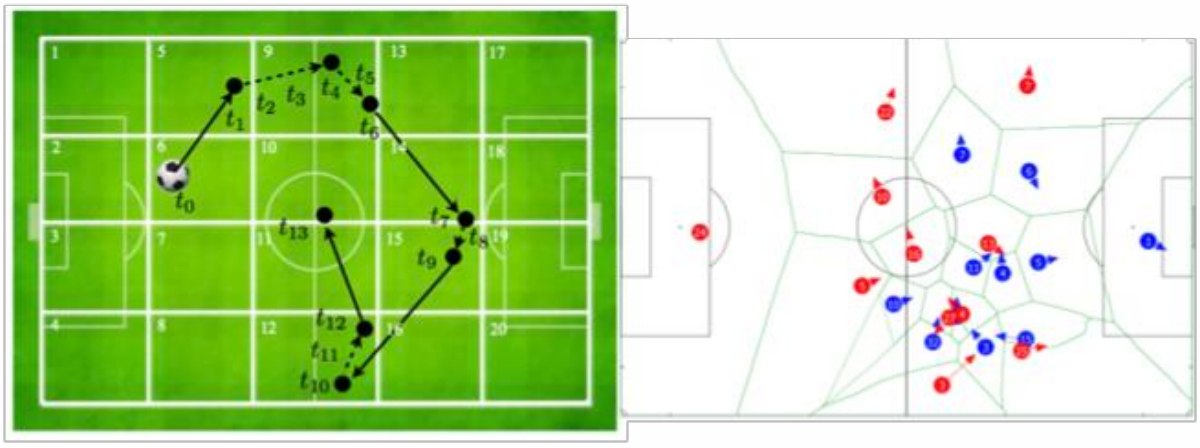


Figure 35 LHS (Lucey, et al., n.d.)equal splitting of pitch, RHS dominant regions (Gudmundsson & Horton, n.d.)

Goal keeper features such as number of goals saved will be excluded from the feature table; however they will be left in as players in the overall analysis to test the robustness of the results in the hope they are clearly grouped as their own positional category.

Feature Selection and Dimensionality Reduction

The third stage will include researching and experimenting with feature selection techniques to try and find the most important features and reduce the dimensions of the dataset to a more controlled size. The aforementioned method of gridding the pitch and turning the spatio-temporal data into discrete features will provide a feature table that can be used with dimension reduction techniques. The aim is to find the relationship of the features and remove any that are highly correlated in a bid to maintain the most amount of information in a more manageable set. To do this principal component analysis in R will be used, clustering the results using k-means algorithm (experimenting to find a suitable value for k). Self organising maps (SOM) will also be used as another method of visualising the high dimensional feature set. The results of the k-means and SOM will be compared in a bid to attain the most information relative to categorising football players into their positions. From here the aim is to delve deeper into these results to understand whether there are any obvious groupings of types of players within their already known positions. For example, within a group of defenders are there any obvious

differences between attackingly minded defenders and defensively minded defenders? Similar methods to the previous stage will be used here but with sub groups of players determined by the stage three results.

Unsupervised Classification
Restricted Boltzmann machines will be used to create a model that allows the input of a player's skill set delivering an output of their optimum playing position.

Create Performance Metric for Supervised Classification
The final exploratory phase of the project will be creating a performance metric to rank players. From here we can understand the most important features on certain positions as well as use classification techniques to predict how successful a player may be in the future. Researching around this topic will provide ideas of how others have approached this problem and will allow an informed decision of the best performance metric to use. Once a performance metric has been chosen and players have been labelled a supervised learning method can be created. Decision Trees and Naïve Bayes will be the main methods compared.

Evaluation
Evaluation will be carried out throughout the project in the form of statistical outputs and visualisations. This will allow iteration between successful and unsuccessful methods. Research at different stages will be discussed with an Opta pro contact, receiving feedback where necessary. This will be done at the end of every key section described in the work plan by emailing a small table of results with a small summary of the findings. Alongside this some questions such as: are these results novel? Are these results useful? The answers to these questions will help to guide the research and answer the research question most efficiently.
Processing visualisation as shown in figure 1 will be used to help ensure the numerical results make sense. For the feature selection and derivation section results will be graphed to more easily understand any groupings of players that are found. For the classification section cross validation will be used, with results provided in a table with percentage of error in the training, validation and test sets compared to find the best method.

Reporting
With so much trial, error and experimentation involved it is crucial to record each stage as it develops. The report will be started when the project commences. However since not all can be written until final results are achieved there will be a six-week window for the final write up.

Ethical, Legal and Professional issues
It must be ensured that all research is carried out without any chance of ethical, legal or professional issues occurring. When reviewing the City University Ethics Review Questionnaire there were no questions which required further action. Due to the nature of the data provided from Opta a Non-Disclosure Agreement has been signed and must be upheld to avoid any legal issues occurring.

WORK PLAN

The work plan is shown in figure 1. The main tasks have been identified and split into sub

tasks. The project will start in April 2016 and finish in January 2017. If there is any time deviation from the current plan this will be adapted to show this. If there is deviation above reasonable bounds a new plan will be produced. Six weeks has been left for the write up incorporating extra time incase any of the risks discussed occur.

Work Breakdown Structure

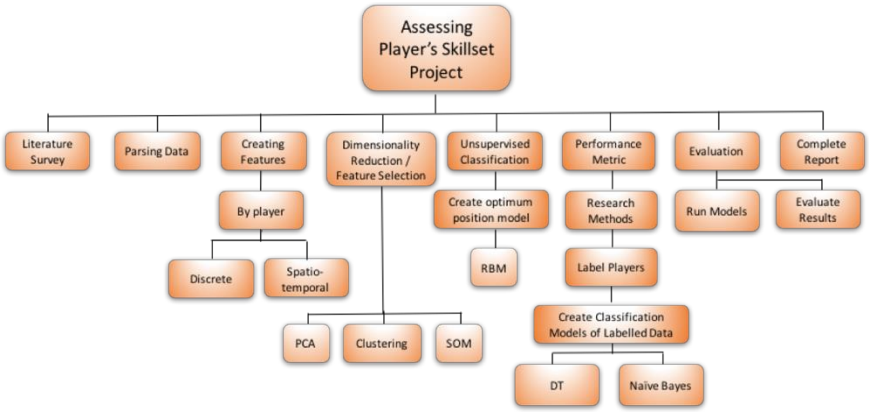
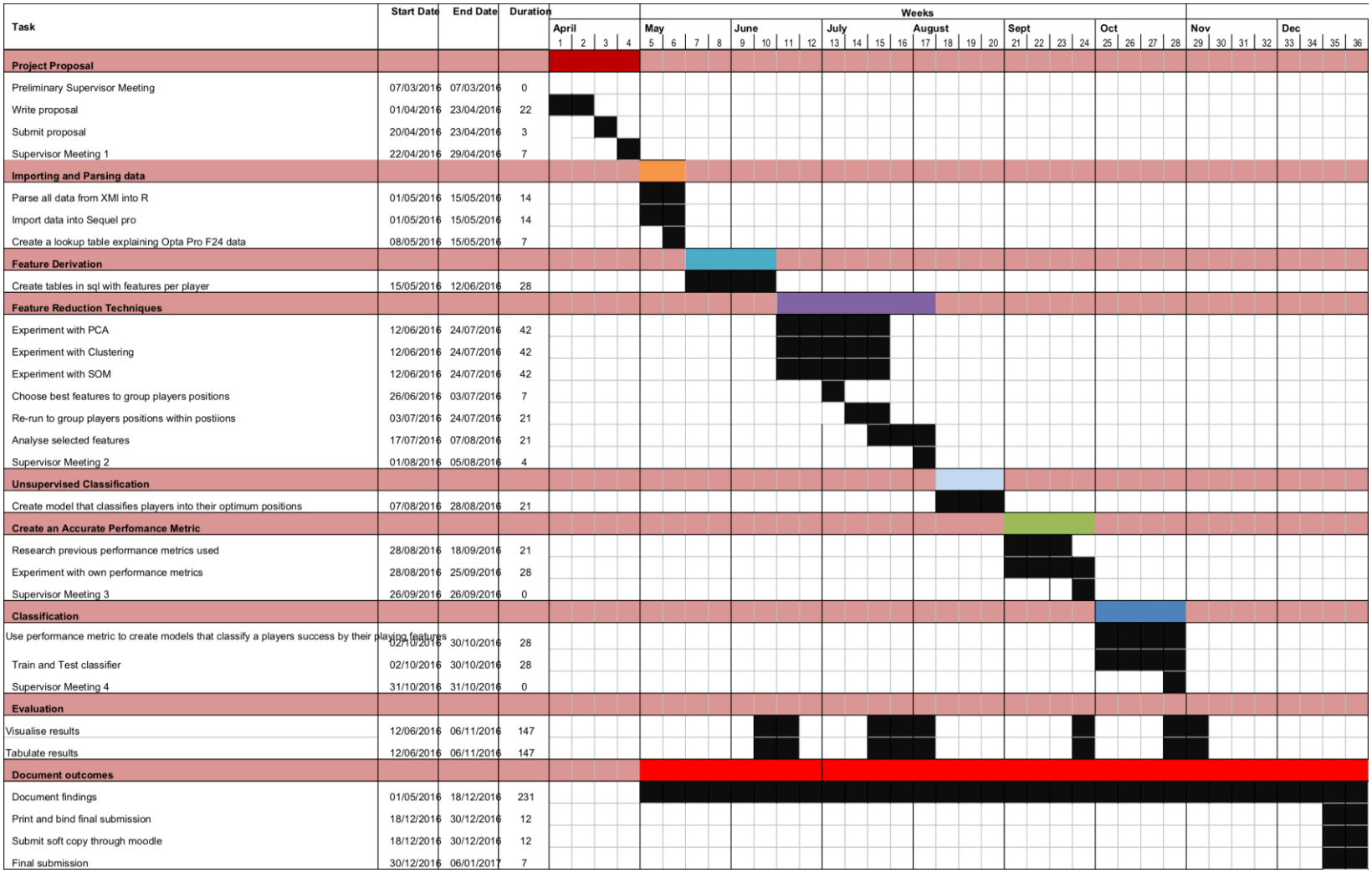


Figure 36 Work Breakdown Structure



Task	Start Date	End Date	Duration
Project Proposal			
Preliminary Supervisor Meeting	07/03/2016	07/03/2016	0
Write proposal	01/04/2016	23/04/2016	22
Submit proposal	20/04/2016	23/04/2016	3
Supervisor Meeting 1	22/04/2016	29/04/2016	7
Importing and Parsing data			
Parse all data from XMl into R	01/05/2016	15/05/2016	14
Import data into Sequel pro	01/05/2016	15/05/2016	14
Create a lookup table explaining Opta Pro F24 data	08/05/2016	15/05/2016	7
Feature Derivation			
Create tables in sql with features per player	15/05/2016	12/06/2016	28
Feature Reduction Techniques			
Experiment with PCA	12/06/2016	24/07/2016	42
Experiment with Clustering	12/06/2016	24/07/2016	42
Experiment with SOM	12/06/2016	24/07/2016	42
Choose best features to group players positions	26/06/2016	03/07/2016	7
Re-run to group players positions within postiions	03/07/2016	24/07/2016	21
Analyse selected features	17/07/2016	07/08/2016	21
Supervisor Meeting 2	01/08/2016	05/08/2016	4
Unsupervised Classification			
Create model that classifies players into their optimum positions	07/08/2016	28/08/2016	21
Create an Accurate Perfomance Metric			
Research previous performance metrics used	28/08/2016	18/09/2016	21
Experiment with own performance metrics	28/08/2016	25/09/2016	28
Supervisor Meeting 3	26/09/2016	26/09/2016	0
Classification			
Use performance metric to create models that classify a players success by their playing features	02/10/2016	30/10/2016	28
Train and Test classifier	02/10/2016	30/10/2016	28
Supervisor Meeting 4	31/10/2016	31/10/2016	0
Evaluation			
Visualise results	12/06/2016	06/11/2016	147
Tabulate results	12/06/2016	06/11/2016	147
Document outcomes			
Document findings	01/05/2016	18/12/2016	231
Print and bind final submission	18/12/2016	30/12/2016	12
Submit soft copy through moodle	18/12/2016	30/12/2016	12
Final submission	30/12/2016	06/01/2017	7

Figure 37 Project Timeline

RISK

Risk management includes a quantitative assessment of any risks that may occur during the project. The risk table will be updated throughout the project, adding any new risks that may develop as well as removing risks that are no longer a threat. The likelihood of the risk will be estimated between 1 and 3 and the consequence between 1 and 5, with the impact being the multiple of these two (Dawson, n.d.). The response provides a solution if these risks were to happen.

Risk	Likelihood (1-3)	Consequence (1-5)	Impact (LxC)	Response
Withdrawal of data from Opta	1	5	5	In constructing research make note of open sources of data available
Terms of usage changed by Opta	1	3	3	In constructing research make note of open sources of data available
Tasks taking longer to complete than initially thought	3	3	9	Allow enough time to account for this
Work pressures leave less time for project than initially anticipated	2	3	6	Add 10% to time estimates
Interesting topics arise and take extra time to research	2	3	6	Add 10% to time estimates, stick to plan
Data too complicated to fully understand	2	4	8	Keep up relationship with Opta to ask questions and get help with data if needed.
Data not sufficient to answer research question	1	3	3	In constructing research make note of open sources of data available
Laptop lost/stolen or crashes	2	5	10	Back up once a week
Loss of Motivation	1	3	3	Chosen a topic that is of the upmost interest so this is extremely unlikely
Classification models take too long to run	2	3	6	Reduce number of models and folds

Figure 38 Risk Breakdown

REFERENCES

3.4.0, P. s. v., n.d. *Principal Components Analysis*. [Online]
Available at: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/prcomp.html
[Accessed 12 October 2016].
Analysis, M., n.d. [Online]
Available at: http://matchanalysis.com
Anderson, C. & Sally, D., n.d. *The Numbers Game: Why Everything You Know About Football is Wrong*. s.l.:Penguin.
Andrienko, N. & Andrienko, G., n.d. *Exploratory Analysis of Spatial and Temporal Data A Systematic Approach*. s.l.:Approx. 715 p. 282 illus., 37 in colour., Hardcover, 139,05 € Springer-Verlag, December 2005, ISBN 3-540-25994-5.
Andrienko, N., Andrienko, G., Pelekis, N. & Spaccapietra, S., n.d. Basic Concepts of Movement Data. In: *Mobility, Data Mining and Privacy*. s.l.:s.n.
Bojko, A., 2009. Informative or Misleading? Heatmaps Deconstructed. *Human-Computer Interaction. New Trends,* 5610(1), pp. 30-39.
Breiman, L., 2001. *Random Forests,* Berkeley: Berkeley.
Bull, J., 2016. *How Claudio Ranieri's tactics put his rivals to shame at Leicester City*. [Online]
Available at: http://www.telegraph.co.uk/football/2016/04/30/how-claudio-ranieris-tactics-put-his-rivals-to-shame-at-leiceste/
[Accessed 22 November 2016].
Carolin Strobl, A.-L. B. Z. H., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics,* 25 January.8(25).
Charles, N., 2016. *Find me a player like Andrés Iniesta: Part 2*. [Online]
Available at: http://www.hilltop-analytics.com/football/opta-conference-2016-find-me-a-player-like-andres-iniesta-part-2/
[Accessed October 2016].
Clarke, T. V. O. a. P., 2009. *A SIMPLE CORRECTION TO REMOVE THE BIAS OF THE GINI COEFFICIENT DUE TO GROUPING ,* s.l.: Netspar.
Dawson, C., n.d. *Projects in Computing and Information Systems*. s.l.:s.n.
Duch, J., Waitzman, J. S. & Amara, L. A. N., 2010. Quantifying the Performance of Individual Players in a Team Activity. *10.1371/journal.pone.0010937,* 16.
Ducker, J., 2016. *Jose Mourinho insists Wayne Rooney won't play in midfield for Manchester United next season: 'He will never be a No 8'*. [Online]
Available at: http://www.telegraph.co.uk/football/2016/07/05/jose-mourinho-insists-wayne-rooney-wont-play-in-midfield-for-man/
[Accessed 12 November 2016].
Dunteman, G. H., 1989. *Principal Components Analysis*. s.l.:s.n.
fourfourtwo, n.d. *wenger-how-improve-your-5-side*. [Online]
Available at: http://www.fourfourtwo.com/performance/skills/wenger-how-improve-your-5-side
[Accessed 2016].
Gilles Louppe, L. W. A. S. P. G., n.d. *Understanding variable importances in forests of randomized trees,* Belgium: s.n.
github, n.d. *ggbiplot*. [Online]
Available at: https://github.com/vqv/ggbiplot
Google, n.d. *Google Cloud Platform*. [Online]
Available at: https://console.cloud.google.com/
Gudmundsson, J. & Horton, M., n.d. *Spatio-Temporal Analysis of Team Sports – A Survey*. [Online]
Available at: http://arxiv.org/pdf/1602.06994v1.pdf
Ingle, S., 2016. *Danny Willett's extra club: how new analytics company aided Masters winner.* [Online]
Available at: http://www.theguardian.com/sport/2016/apr/13/danny-willett-golf-analytics-masters-

champion-15th-club
[Accessed 20 April 2016].
Jolliffe, I., 2002. *Principal Component Analysis, Second Edition*. Second ed. Aberdeen: Springer.
Lewis, M., 2003. *Moneyball*. s.l.:W. W. Norton & Company.
Lewis, M., 2003. *Moneyball: The Art of Winning an Unfair Game*. s.l.:Norton.
Likasa, A., Vlassisb, N. & Jakob J. Verbeekb, 2003. The global k-means clustering algorithm. *Pattern Recognition*, February, 36(2), p. 451–461.
Lucey, P. et al., 2013. Assessing Team Strategy using Spatiotemporal Data.
Lucey, P. et al., n.d. Assessing Team Strategy using Spatiotemporal Data.
Malone, J., McGarry, K., Wermter, S. & Bowerman, C., 2005. Data mining using rule extraction from Kohonen self-organising maps. *Neural Comput & Applic (2005) 15: 9–17 DOI 10.1007/s00521-005-0002-1*.
Mike Hughes, I. F., 2005. Analysis of passing sequences, shots and goals in soccer. *Journal of Sports Sciences,* 23(5), pp. 509-514.
Oliver, D., 2005. *Basketball on Paper: Rules and Tools for Performance Analysis*. Washington: Potomac Books, Inc..
OptaPro, n.d. *Opta Pro*. [Online]
Available at: http://www.optasportspro.com/
[Accessed 30 October 2016].
Prozone, n.d. [Online]
Available at: http://prozonesports.stats.com
Reep, B. B. a. C., 1968. Skill and Chance in Association Football. *Journal of the Royal Statistical Society*. *Series A (General),* 131(4), pp. 581-585.
Reep, R. P. a. C., 1997. Measuring the Effectiveness of Playing Strategies at Soccer. *Journal of the Royal Statistical Society*. *Series D (The Statistician),* 46(4), pp. 541-550.
Sacha, D. et al., n.d. Knowledge Generation Model for Visual Analytics. *IEEE*.
Salakhutdinov, R., Mnih, A. & Hinton, G., n.d. Restricted Boltzmann Machines for Collaborative Filtering.
Sally, C. A. a. D., n.d. *The Numbers Game: Why Everything You Know About Football is Wrong*. s.l.:Penguin.
Smith, R., 2015. *Bespoke data gives Arsene Wenger the tactical edge at Arsenal*. [Online]
Available at: http://www.thetimes.co.uk/tto/sport/football/article4638099.ece
[Accessed November 2016].
Sports, O., n.d. [Online]
Available at: http://www.optasports.com/services/media/data-feeds/opta-data-feeds-overview.aspx
Sports, S., 2016. *Sky Sports News*. [Online]
Available at: http://www.skysports.com/football/news/11667/10528626/paul-pogba-joins-manchester-united-for-world-record-89m
[Accessed 2016].
Statistical properties of position-dependent ball-passing networks in football games Takuma Narizuka*1, K. Y. a. Y. Y., n.d. [Online]
Available at: http://arxiv.org/pdf/1311.0641.pdf
STATS, n.d. *Sports Data Company*. [Online]
Available at: www.stats.com
T. Taki, J.-i. H. a. T. F., 1996. Development of motion analysis system for quantitative evaluation of teamwork in soccer games.. *In Proceedings of 3rd IEEE International Conference on Image Processing*, September, Volume 3, pp. 815-818.
T. Warren Liao, 2005. Clustering of time series data—a survey. *Pattern Recognition,* November, 38(11), p. 1857–1874.
Taki, T., Hasegawa, J.-i. & Fukumura, T., n.d. Development of motion analysis system for quantitative evaluation of teamwork in soccer games.. *In Proceedings of 3rd IEEE International*

Conference on Image Processing, volume 3, pages 815–818, Lausanne, sep 1996. IEEE. ISBN 0-7803-3259-8. doi: 10.1109/ICIP.1996.560865..
Taylor, D., 2016. *Wayne Rooney says Allardyce left him 'battered' with carte blanche remark.* [Online]
Available at: https://www.theguardian.com/football/2016/oct/04/wayne-rooney-sam-allardyce-england-left-me-battered
[Accessed 2016].
Website, B. M., n.d. [Online]
Available at: https://www.fcbayern.de/us/news/news/2015/bayern-head-of-match-analysis-michael-niemeyer-on-soccer-analytics-at-mit-sloan-sports-conference.php
Wenger, A., n.d. [Online]
Available at:
https://www.reddit.com/r/Gunners/comments/3wcrf5/bespoke_data_gives_ars%C3%A8ne_wenger_a_tactical_edge
Wikipedia, n.d. *2013-2014 Premier League*. [Online]
Available at: https://en.wikipedia.org/wiki/2013%E2%80%9314_Premier_League
Zuccolotto, M. S. a. P., 2008. *A bias correction algorithm for the Gini variable importance measure in classification trees,* Brescia: s.n.

Ethics Review Form: Bsc, Msc and MA Projects
Computer Science Research Ethics Committee (CSREC)

Undergraduate and postgraduate students undertaking their final project in the Department of Computer Science are required to consider the ethics of their project work and to ensure that it complies with research ethics guidelines. In some cases, a project will need approval from an ethics committee before it can proceed. Usually, but not always, this will be because the student is involving other people ("participants") in the project.
In order to ensure that appropriate consideration is given to ethical issues, all students must complete this form and attach it to their project proposal document. There are two parts:
Part A: Ethics Checklist. All students must complete this part. The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.
Part B: Ethics Proportionate Review Form. Students who have answered "no" to questions 1 – 18 and "yes" to question 19 in the ethics checklist must complete this part. The project supervisor has delegated authority to provide approval in this case. The approval may be provisional: the student may need to seek additional approval from the supervisor as the project progresses.

A.1 If your answer to any of the following questions (1 – 3) is YES, you must apply to an appropriate external ethics committee for approval.		*Delete as appropriate*
1.	Does your project require approval from the National Research Ethics Service (NRES)? For example, because you are recruiting current NHS patients or staff? If you are unsure, please check at http://www.hra.nhs.uk/research-community/before-you-apply/determine-which-review-body-approvals-are-required/.	**No**
2.	Does your project involve participants who are covered by the Mental Capacity Act? If so, you will need approval from an external ethics	**No**

	committee such as NRES or the Social Care Research Ethics Committee http://www.scie.org.uk/research/ethics-committee/.	
3.	Does your project involve participants who are currently under the auspices of the Criminal Justice System? For example, but not limited to, people on remand, prisoners and those on probation? If so, you will need approval from the ethics approval system of the National Offender Management Service.	**No**
A.2 If your answer to any of the following questions (4 – 11) is YES, you must apply to the City University Senate Research Ethics Committee (SREC) for approval (unless you are applying to an external ethics committee).		*Delete as appropriate*
4.	Does your project involve participants who are unable to give informed consent? For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf?	**No**
5.	Is there a risk that your project might lead to disclosures from participants concerning their involvement in illegal activities?	**No**
6.	Is there a risk that obscene and or illegal material may need to be accessed for your project (including online content and other material)?	**No**
7.	Does your project involve participants disclosing information about sensitive subjects? For example, but not limited to, health status, sexual behaviour, political behaviour, domestic violence.	**No**
8.	Does your project involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning? (See http://www.fco.gov.uk/en/)	**No**
9.	Does your project involve physically invasive or intrusive procedures? For example, these may include, but are not limited to, electrical stimulation, heat, cold or bruising.	**No**
10.	Does your project involve animals?	**No**
11.	Does your project involve the administration of drugs, placebos or other substances to study participants?	**No**
A.3 If your answer to any of the following questions (12 – 18) is YES, you must submit a full application to the Computer Science Research Ethics Committee (CSREC) for approval (unless you are applying to an external ethics committee or the Senate Research Ethics Committee). Your application may be referred to the Senate Research Ethics Committee.		
12.	Does your project involve participants who are under the age of 18?	**No**
13.	Does your project involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.	**No**

14.	Does your project involve participants who are recruited because they are staff or students of City University London? For example, students studying on a specific course or module. (If yes, approval is also required from the Head of Department or Programme Director.)	**No**
15.	Does your project involve intentional deception of participants?	**No**
16.	Does your project involve participants taking part without their informed consent?	**No**
17.	Does your project pose a risk to participants or other individuals greater than that in normal working life?	**No**
18.	Does your project pose a risk to you, the researcher, greater than that in normal working life?	**No**
A.4 If your answer to the following question (19) is YES and your answer to all questions 1 – 18 is NO, you must complete part B of this form.		
19.	Does your project involve human participants or their identifiable personal data? For example, as interviewees, respondents to a survey or participants in testing.	**No**

Appendix B – Team Results

Player Name	Player Position	31:Crystal Palace
Mile Jedinak	Central Midfielder	51.10
Marouane Chamakh	Striker	51.03
Joel Ward	Full Back	48.59
Jason Puncheon	Attacking Midfielder	45.19
Yannick Bolasie	Winger	39.41
Cameron Jerome	Striker	37.66
Kagisho Dikgacoi	Central Midfielder	37.25
Damien Delaney	Central Defender	30.85
Barry Bannan	Central Midfielder	29.55
Adrian Mariappa	Central Defender	24.22
Dwight Gayle	Striker	23.17
Dean Moxey	Full Back	21.25
Daniel Gabbidon	Central Defender	21.14
Glenn Murray	Striker	18.14
JosIe CampaI±a	Central Midfielder	17.44
Stuart O'Keefe	Central Midfielder	16.26
Jonathan Parr	Full Back	15.31
AdlIne GuIedioura	Central Midfielder	14.39
Jerome Thomas	Winger	13.38
Julian Speroni	Goalkeeper	12.52
Jonathan Williams	Attacking Midfielder	8.40
Jimmy KIebIe	Winger	5.29
Owen Garvan	Attacking Midfielder	3.99
Aaron Wilbraham	Striker	3.48
Kevin Phillips	Striker	1.18
Patrick McCarthy	Central Defender	0.52

Player Name	Player Position	54:Fulham
Steve Sidwell	Central Midfielder	52.43
Pajtim Kasami	Central Midfielder	45.78
Scott Parker	Central Midfielder	45.17
Dimitar Berbatov	Striker	36.93
Lewis Holtby	Attacking Midfielder	33.13
Ashkan Dejagah	Attacking Midfielder	31.28
Bryan Ruiz	Second Striker	31.14
Brede Hangeland	Central Defender	30.77
Kieran Richardson	Central Midfielder	30.56
Giorgos Karagounis	Attacking Midfielder	29.58
Alexander Kacaniklic	Winger	28.47
Sascha Riether	Full Back	27.12
Fernando Amorebieta	Central Defender	26.58
Darren Bent	Striker	26.07
Adel Taarabt	Attacking Midfielder	23.35
John Arne Riise	Full Back	23.23
Aaron Hughes	Central Defender	22.03
Johnny Heitinga	Central Defender	20.87
Hugo Rodallega	Striker	20.55
Philippe Senderos	Central Defender	17.96
Damien Duff	Winger	16.68
Elsad Zverotic	Central Defender	10.88
David Stockdale	Goalkeeper	8.59
Maarten Stekelenburg	Goalkeeper	7.69
Derek Boateng	Defensive Midfielder	7.10
Chris David	Attacking Midfielder	4.06
Matthew Briggs	Full Back	3.86
Muamer Tankovic	Striker	1.28
Mesca	Winger	0.16

Player Name	Player Position	88:Hull City
Tom Huddlestone	Central Midfielder	53.04
Jake Livermore	Central Midfielder	49.82
David Meyler	Central Midfielder	46.72
Curtis Davies	Central Defender	40.44
Liam Rosenior	Full Back	39.11
Yannick Sagbo	Striker	37.22
Ahmed Elmohamady	Winger	32.95
Nikica Jelavic	Striker	32.84
Robert Koren	Central Midfielder	32.46
Maynor Figueroa	Full Back	31.88
Sone Aluko	Second Striker	29.66
George Boyd	Winger	29.26
Shane Long	Striker	28.55
James Chester	Central Defender	27.82
Danny Graham	Striker	25.45
Alex Bruce	Central Defender	24.35
Robbie Brady	Attacking Midfielder	23.79
Stephen Quinn	Central Midfielder	23.33
Paul McShane	Central Defender	19.90
Allan McGregor	Goalkeeper	11.03
Steve Harper	Goalkeeper	7.49
Abdoulaye Faye	Central Defender	5.78
Eldin Jakupovic	Goalkeeper	1.82
Gedo	Attacking Midfielder	0.86
Nick Proschwitz	Striker	0.79
Aaron McLean	Striker	0.00

Player Name	Player Position	21:West Ham United
Mark Noble	Central Midfielder	52.57
Mohamed DiamI	Defensive Midfielder	45.84
Stewart Downing	Winger	42.06
Kevin Nolan	Attacking Midfielder	39.11
Matthew Taylor	Winger	36.48
James Tomkins	Central Defender	35.39
Andy Carroll	Striker	34.55
Ravel Morrison	Attacking Midfielder	32.47
George McCartney	Full Back	31.79
Joey O'Brien	Full Back	30.70
Modibo Maiga	Striker	30.15
Matthew Jarvis	Winger	29.04
James Collins	Central Defender	28.32
Guy Demel	Full Back	28.28
Carlton Cole	Striker	27.75
Winston Reid	Central Defender	25.08
Joe Cole	Attacking Midfielder	22.23
Razvan Rat	Full Back	19.61
Jack Collison	Central Midfielder	18.05
Ricardo Vaz Te	Second Striker	11.19
AdriIAn	Goalkeeper	10.23
Jussi JISISskellSinen	Goalkeeper	7.67
Alou Diarra	Defensive Midfielder	2.60
Mladen Petric	Striker	1.32

Player Name	Player Position	56:Sunderland
Sebastian Larsson	Central Midfielder	43.46
Phillip Bardsley	Full Back	42.81
Jack Colback	Central Midfielder	42.45
Ki Sung-Yueng	Central Midfielder	42.07
Adam Johnson	Winger	39.36
John O'Shea	Central Defender	38.56
Fabio Borini	Striker	38.26
Jozy Altidore	Striker	37.84
Lee Cattermole	Defensive Midfielder	37.00
Emanuele Giaccherini	Winger	35.67
Steven Fletcher	Striker	30.23
Wes Brown	Central Defender	28.54
Ondrej Celustka	Central Defender	26.35
Craig Gardner	Central Midfielder	24.84
Valentin Roberge	Central Defender	18.31
Modibo DiakitIe	Central Defender	17.56
Andrea Dossena	Full Back	13.49
Vito Mannone	Goalkeeper	11.23
Carlos CuIellar	Central Defender	9.04
Ji Dong-Won	Striker	8.63
Keiren Westwood	Goalkeeper	6.34
Charalampos Mavrias	Winger	5.69
Cabral	Defensive Midfielder	4.99
StIphane Sessegnon	Attacking Midfielder	2.44
El-Hadji Ba	Defensive Midfielder	0.42

Player Name	Player Position	80:Swansea City
Jonathan De GuzmIAn	Attacking Midfielder	50.94
Jonjo Shelvey	Attacking Midfielder	49.78
Wayne Routledge	Winger	46.02
Leon Britton	Attacking Midfielder	43.56
Pablo HernIAndez	Winger	42.94
Wilfried Bony	Striker	41.56
JosIe CaI±as	Central Midfielder	39.12
Chico	Central Defender	37.64
Ashley Williams	Central Defender	36.44
Angel Rangel	Full Back	34.76
Ben Davies	Full Back	34.72
Ilex Pozuelo	Winger	33.09
Nathan Dyer	Winger	32.76
Jordi Amat	Central Defender	30.09
Michu	Second Striker	28.05
Dwight Tiendalli	Full Back	20.91
Roland Lamah	Winger	20.09
Neil Taylor	Full Back	15.60
Ilvaro	Striker	11.36
Michel Vorm	Goalkeeper	8.10
Gerhard Tremmel	Goalkeeper	6.16
Ki Sung-Yueng	Central Midfielder	0.54

Player Name	Player Position	8:Chelsea
Oscar	Attacking Midfielder	47.98
Ramires	Central Midfielder	47.70
Eden Hazard	Second Striker	47.21
Willian	Winger	42.00
Frank Lampard	Central Midfielder	41.78
CIesar Azpilicueta	Full Back	40.96
Branislav Ivanovic	Central Defender	39.25
Fernando Torres	Striker	35.25
David Luiz	Central Defender	34.96
John Terry	Central Defender	34.23
Andre SchISrrle	Second Striker	33.77
Gary Cahill	Central Defender	33.20
Juan Mata	Attacking Midfielder	31.57
John Obi Mikel	Defensive Midfielder	31.08
Ashley Cole	Full Back	26.26
Samuel Eto'o	Striker	25.07
Demba Ba	Striker	20.57
Michael Essien	Defensive Midfielder	11.29
Petr Cech	Goalkeeper	9.66
Kevin De Bruyne	Attacking Midfielder	5.77
Ryan Bertrand	Full Back	5.71
Mark Schwarzer	Goalkeeper	3.49
Tomas Kalas	Central Defender	2.71
Romelu Lukaku	Striker	0.68
Nathan AkIe	Central Defender	0.34
Marco van Ginkel	Central Midfielder	0.12

Player Name	Player Position	1:Manchester United
Wayne Rooney	Second Striker	47.87
Michael Carrick	Central Midfielder	45.98
Tom Cleverley	Winger	42.28
Phil Jones	Central Defender	42.20
Marouane Fellaini	Attacking Midfielder	41.82
Adnan Januzaj	Attacking Midfielder	38.78
Juan Mata	Attacking Midfielder	37.62
Nemanja Vidic	Central Defender	35.18
Shinji Kagawa	Attacking Midfielder	34.68
Danny Welbeck	Striker	32.77
Chris Smalling	Central Defender	31.79
Ashley Young	Attacking Midfielder	31.30
Robin van Persie	Striker	30.62
Darren Fletcher	Central Midfielder	30.54
Patrice Evra	Full Back	30.53
Jonny Evans	Central Defender	29.55
Antonio Valencia	Winger	27.58
Rafael	Full Back	24.95
Ryan Giggs	Central Midfielder	23.96
Nani	Winger	23.60
Javier HernIAndez	Striker	20.72
Rio Ferdinand	Central Defender	19.78
Alexander BISttner	Full Back	15.02
Anderson	Central Midfielder	9.62
David De Gea	Goalkeeper	9.52
Fabio	Full Back	4.50
Wilfried Zaha	Attacking Midfielder	2.21
Anders Lindegaard	Goalkeeper	1.93

Player Name	Player Position	35:West Bromwich Albion
Youssuf Mulumbu	Defensive Midfielder	49.47
James Morrison	Winger	49.21
Chris Brunt	Attacking Midfielder	41.95
StIphane Sessegnon	Attacking Midfielder	41.88
Claudio Yacob	Defensive Midfielder	37.74
Morgan Amalfitano	Central Midfielder	36.11
Saido Berahino	Striker	34.69
Gareth McAuley	Central Defender	33.88
Jonas Olsson	Central Defender	33.09
Graham Dorrans	Central Midfielder	32.73
Liam Ridgewell	Central Defender	30.68
Victor Anichebe	Striker	30.30
Shane Long	Striker	28.74
Billy Jones	Full Back	28.52
Nicolas Anelka	Striker	25.84
Craig Dawson	Central Defender	22.95
Steven Reid	Full Back	21.88
ZoltIAn Gera	Winger	20.87
Matej Vydra	Striker	16.48
Diego Lugano	Central Defender	14.54
Ben Foster	Goalkeeper	12.73
Boaz Myhill	Goalkeeper	8.44
Scott Sinclair	Winger	6.66
Goran Popov	Full Back	4.98
Markus Rosenberg	Striker	1.72
Luke Daniels	Goalkeeper	0.43

Player Name	Player Position	7:Aston Villa
Fabian Delph	Central Midfielder	47.64
Christian Benteke	Striker	45.23
Ashley Westwood	Central Midfielder	43.73
Leandro Bacuna	Attacking Midfielder	41.45
Gabriel Agbonlahor	Striker	39.24
Andreas Weimann	Striker	38.47
Karim El Ahmadi	Defensive Midfielder	37.09
Ciaran Clark	Central Defender	34.72
Ron Vlaar	Central Defender	30.69
Nathan Baker	Central Defender	30.18
Libor KozIAk	Striker	25.55
Matthew Lowton	Full Back	24.34
Ryan Bertrand	Full Back	23.22
Yacouba Sylla	Central Midfielder	22.63
Antonio Luna	Full Back	21.09
Aleksandar Tonev	Winger	17.10
Brad Guzan	Goalkeeper	14.83
Jordan Bowery	Striker	13.22
Joe Bennett	Full Back	8.83
Chris Herd	Central Midfielder	7.94
Jores Okore	Central Defender	7.50
Callum Robinson	Unknown	1.83
Nicklas Helenius	Striker	1.28

Player Name	Player Position	14:Liverpool
Jordan Henderson	Central Midfielder	54.06
Steven Gerrard	Attacking Midfielder	53.20
Philippe Coutinho	Attacking Midfielder	47.69
Luis Suarez	Second Striker	46.54
Raheem Sterling	Winger	44.36
Lucas Leiva	Central Midfielder	43.97
Glen Johnson	Full Back	41.98
Joe Allen	Central Midfielder	38.86
Martin Skrtel	Central Defender	37.52
Daniel Sturridge	Striker	37.41
Jon Flanagan	Full Back	37.09
Kolo Toure	Central Defender	32.52
Mamadou Sakho	Central Defender	26.81
Victor Moses	Attacking Midfielder	25.78
Daniel Agger	Central Defender	25.01
Aly Cissokho	Full Back	19.16
JosIe Enrique	Full Back	15.18
Iago Aspas	Striker	15.12
Luis Alberto	Attacking Midfielder	9.94
Simon Mignolet	Goalkeeper	9.83
Martin Kelly	Full Back	5.58
Jordon Ibe	Winger	0.30

Player Name	Player Position	11:Everton
Gareth Barry	Central Midfielder	51.77
Leon Osman	Attacking Midfielder	47.64
James McCarthy	Central Midfielder	46.98
Ross Barkley	Central Midfielder	45.07
Romelu Lukaku	Striker	43.67
Kevin Mirallas	Attacking Midfielder	39.16
Steven Naismith	Striker	37.84
Phil Jagielka	Central Defender	35.66
Steven Pienaar	Attacking Midfielder	33.34
Sylvain Distin	Central Defender	32.90
Leighton Baines	Full Back	32.74
Seamus Coleman	Full Back	31.30
John Stones	Full Back	30.47
Gerard Deulofeu	Second Striker	24.19
Bryan Oviedo	Full Back	18.92
Antolin Alcaraz	Central Defender	17.80
Marouane Fellaini	Attacking Midfielder	17.67
Nikica Jelavic	Striker	12.36
Tim Howard	Goalkeeper	10.12
Joel Robles	Goalkeeper	2.56
Arouna KonIe	Striker	1.69
Darron Gibson	Central Midfielder	0.76
Johnny Heitinga	Central Defender	0.45
Tony Hibbert	Full Back	0.10

Player Name	Player Position	4:Newcastle United
Moussa Sissoko	Attacking Midfielder	50.84
Cheick TiotIe	Defensive Midfielder	50.28
Yoan Gouffran	Second Striker	47.24
Vurnon Anita	Defensive Midfielder	46.18
Yohan Cabaye	Central Midfielder	40.36
Davide Santon	Full Back	37.70
Shola Ameobi	Striker	36.80
LoIÀc Remy	Striker	36.52
Hatem Ben Arfa	Attacking Midfielder	35.51
Michael Williamson	Central Defender	34.27
Papiss Demba CissIe	Striker	33.01
Mathieu Debuchy	Full Back	30.96
Mapou Yanga-Mbiwa	Central Defender	30.53
Fabricio Coloccini	Central Defender	28.76
Paul Dummett	Full Back	22.37
Tim Krul	Goalkeeper	17.97
Dan Gosling	Central Midfielder	17.05
Massadio Haidara	Full Back	14.29
Sammy Ameobi	Striker	14.17
Steven Taylor	Central Defender	13.38
Sylvain Marveaux	Winger	12.99
Robert Elliot	Goalkeeper	3.50
JonIAs GutiIerrez	Attacking Midfielder	2.29
Gabriel Obertan	Winger	1.53

Player Name	Player Position	45:Norwich City
Leroy Fer	Central Midfielder	47.16
Bradley Johnson	Central Midfielder	45.03
Robert Snodgrass	Attacking Midfielder	44.34
Jonathan Howson	Central Midfielder	42.37
Nathan Redmond	Winger	38.84
Johan Elmander	Striker	36.82
Alexander Tettey	Defensive Midfielder	35.94
Russell Martin	Full Back	35.22
Gary Hooper	Striker	30.98
Martin Olsson	Full Back	29.49
Wes Hoolahan	Central Midfielder	29.47
Ricky van Wolfswinkel	Striker	28.49
Steven Whittaker	Full Back	28.20
SIebastien Bassong	Central Defender	27.09
Michael Turner	Central Defender	26.64
Anthony Pilkington	Winger	23.63
Ryan Bennett	Central Defender	21.89
John Ruddy	Goalkeeper	12.33
Javier Garrido	Full Back	11.49
JonIAs GutiIerrez	Attacking Midfielder	8.67
Josh Murphy	Winger	4.80
Elliott Bennett	Winger	4.05
Luciano Becchio	Striker	2.15

Player Name	Player Position	6:Tottenham Hotspur
Paulinho	Central Midfielder	46.33
Christian Eriksen	Attacking Midfielder	43.77
Mousa DembIeIle	Central Midfielder	43.20
Emmanuel Adebayor	Striker	40.64
Kyle Naughton	Full Back	40.26
Nacer Chadli	Winger	40.05
Michael Dawson	Central Defender	38.25
Kyle Walker	Full Back	36.96
Aaron Lennon	Winger	36.49
Gylfi Sigurdsson	Attacking Midfielder	34.04
Roberto Soldado	Striker	33.74
Jan Vertonghen	Central Defender	32.32
Nabil Bentaleb	Unknown	30.31
Andros Townsend	Attacking Midfielder	29.61
Vlad Chiriches	Central Defender	28.86
Younes Kaboul	Central Defender	28.09
Sandro	Defensive Midfielder	27.92
Lewis Holtby	Attacking Midfielder	25.85
Danny Rose	Full Back	25.80
Etienne Capoue	Defensive Midfielder	25.48
Harry Kane	Striker	17.68
Erik Lamela	Attacking Midfielder	17.11
Ezekiel Fryers	Full Back	14.87
Jermain Defoe	Striker	11.35
Hugo Lloris	Goalkeeper	10.75
Brad Friedel	Goalkeeper	1.90

Player Name	Player Position	3:Arsenal
Santiago Cazorla	Winger	51.98
Olivier Giroud	Striker	48.23
Mesut ozil	Attacking Midfielder	48.02
Aaron Ramsey	Central Midfielder	47.80
Mikel Arteta	Central Midfielder	47.22
Jack Wilshere	Attacking Midfielder	44.28
Mathieu Flamini	Defensive Midfielder	39.94
Bacary Sagna	Full Back	39.68
Tomas Rosicky	Attacking Midfielder	38.06
Per Mertesacker	Central Defender	34.70
Laurent Koscielny	Central Defender	30.49
Kieran Gibbs	Full Back	29.75
Lukas Podolski	Striker	26.05
Alex Oxlade-Chamberlain	Attacking Midfielder	23.43
Nacho Monreal	Full Back	23.12
Thomas Vermaelen	Central Defender	20.66
Theo Walcott	Winger	17.88
Carl Jenkinson	Full Back	17.12
Serge Gnabry	Attacking Midfielder	15.94
Wojciech Szczesny	Goalkeeper	10.55
Nicklas Bendtner	Second Striker	6.89
Yaya Sanogo	Striker	3.64
Lukasz Fabianski	Goalkeeper	2.24
Ryo Miyaichi	Attacking Midfielder	0.69
Vassiriki Abou Diaby	Central Midfielder	0.39
Chuba Akpom	Striker	0.00

Player Name	Player Position	20:Southampton
Adam Lallana	Attacking Midfielder	53.18
Steven Davis	Central Midfielder	50.86
Morgan Schneiderlin	Defensive Midfielder	50.77
Rickie Lambert	Striker	44.27
James Ward-Prowse	Central Midfielder	44.25
Jack Cork	Central Midfielder	42.48
Jay Rodriguez	Striker	41.77
Victor Wanyama	Defensive Midfielder	38.50
Dejan Lovren	Full Back	36.92
Jose Fonte	Central Defender	35.97
Nathaniel Clyne	Full Back	33.03
Luke Shaw	Full Back	30.27
Calum Chambers	Winger	26.14
GastISn RamISrez	Attacking Midfielder	25.59
Pablo Osvaldo	Striker	23.37
Maya Yoshida	Central Defender	17.28
Artur Boruc	Goalkeeper	12.61
Jos Hooiveld	Central Defender	10.39
Daniel Fox	Full Back	8.63
Paulo Gazzaniga	Goalkeeper	5.44
Guly	Attacking Midfielder	3.50
Kelvin Davis	Goalkeeper	2.89
Harrison Reed	Unknown	1.65

Player Name	Player Position	110:Stoke City
Steven N'Zonzi	Defensive Midfielder	53.13
Peter Crouch	Striker	46.82
Charlie Adam	Central Midfielder	45.41
Glenn Whelan	Central Midfielder	44.79
Marko Arnautovic	Second Striker	41.25
Jonathan Walters	Winger	40.32
Marc Wilson	Full Back	40.15
Stephen Ireland	Central Midfielder	37.11
Ryan Shawcross	Central Defender	35.63
Geoff Cameron	Central Defender	34.68
Erik Pieters	Full Back	31.87
Peter Odemwingie	Striker	28.33
Marc Muniesa	Central Defender	21.98
Oussama Assaidi	Attacking Midfielder	19.89
Robert Huth	Central Defender	19.46
Wilson Palacios	Defensive Midfielder	18.82
Kenwyne Jones	Striker	14.72
Matthew Etherington	Winger	13.47
Asmir Begovic	Goalkeeper	11.69
Jermaine Pennant	Winger	8.81
Andy Wilkinson	Full Back	8.57
Thomas SIurensen	Goalkeeper	3.57
John Guidetti	Striker	1.83
Brek Shea	Winger	0.20
Cameron Jerome	Striker	0.13

Appendix C – R Code for Running Player Dashboard

The full code for this project consisted of over 50 files, each of which have hundreds of lines of code. Therefore, all the code will be submitted as a supplementary zip file. Appendix C lists the code needed to run the dashboard and includes some examples of key functions used throughout the project. Such as get_player_data and player_plot_fn_app.

To run the code:

Save all files in msc_shiny_app folder on local drive e.g. setwd('~/msc_shiny_app')

```
  library(shinydashboard); library(readr);  library(dplyr); library(stringr);
library(shiny); library(ggplot2);

################
  ## data
  ################
  #mdi
  merge_imp2 = read_csv('pi_data.csv', col_types = "ciicdddd")
  #mdi wavg
  merge_imp2_wavg = read_csv('pi_data_wavg.csv', col_types = "ciicddd")
  #mda
  merge_imp2_mda = read_csv('pi_data_mda.csv',col_types = "ciicdddd")
  #mda wavg
  merge_imp2_mda_wavg = read_csv('pi_data_wavg_mda.csv',col_types= "ciicddd")

  df_for_app = read_csv('df_for_app.csv', col_types = "ddiiTiiicccc")
  team_id_index = read_csv('team_name_id.csv')
  player_info_for_app = read_csv('player_info_for_app.csv')
  event_qid_labels = read_csv('final_event_q_id_labels.csv') %>% filter(!evet
id %in% c('49','5','15','16',NA))
  event_type_list = c(1,2, 3, 4,5,6,7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 41,
42, 44,45,49,50, 51, 52, 54,58,60,61)
  q_pass_list = c(1:8, 106, 107, 155:157, 168,195,196,210,210,223:225)

  #decode characters in names
  player_info_for_app$name = str_replace_all(player_info_for_app$name, "[^[:a
lnum:]]", " ")

  get_player_data = function(team_id, merge_df,game_id){
    format_df = df_for_app %>% dplyr::filter(Event_team_id %in% team_id & !is
.na(Event_player_id), Event_type_id %in% event_type_list, Game_id %in% game_i
d) %>%
      select(Game_id, Event_x, Event_y, Event_team_id, Event_player_id, Event
_timestamp, Event_type_id, Q_qualifier_id) %>%
      group_by(Game_id, Event_x, Event_y,Event_team_id, Event_player_id, Even
t_timestamp, Event_type_id) %>%
      dplyr::summarise(concat_q_id=paste(Q_qualifier_id, collapse=",")) %>% u
```

```
ngroup() %>%
      dplyr::filter(!is.na(Event_type_id)) %>%
      dplyr::mutate(q_pass_only = ifelse(Event_type_id ==1, paste0(Event_type
_id,'_',q_pass_list[q_pass_list %in% unlist(strsplit(concat_q_id,","))]),Even
t_type_id)) %>%
      dplyr::mutate(coord_group_x = plyr::round_any(as.numeric(Event_x), 10,
f = ceiling),
                    coord_group_y = plyr::round_any(as.numeric(Event_y), 10,
f = ceiling)) %>%
      dplyr::mutate(coord_group_x = ifelse(coord_group_x ==0, 10, coord_group
_x)) %>%
      dplyr::mutate(coord_group_y = ifelse(coord_group_y ==0, 10, coord_group
_y))

    # how many times does each player do all events in each coord
    total_passes_per_coord = format_df %>% select(Event_team_id, Event_player
_id, q_pass_only, coord_group_x, coord_group_y) %>%
      dplyr::group_by(coord_group_x, coord_group_y, Event_player_id,Event_tea
m_id) %>%
      dplyr::summarise(count_q_pass = n())

    # how many time player does each event / total events in that coord
    total_passes_per_event_per_coord = format_df %>% select(Event_team_id, Ev
ent_player_id, q_pass_only, coord_group_x, coord_group_y) %>%
      dplyr::group_by(coord_group_x, coord_group_y, q_pass_only, Event_player
_id,Event_team_id) %>%
      dplyr::summarise(each_event_per_coord = n()) %>%
      dplyr::group_by(Event_player_id,coord_group_x, coord_group_y ) %>%
      dplyr::mutate(total_event_per_coord = sum(each_event_per_coord)) %>%
      ungroup() %>%
      #dplyr::mutate(percent_of_pass1 = (each_event_per_coord/total_event_per
_coord)) %>%
      dplyr::mutate(percent_of_pass = (each_event_per_coord/total_event_per_c
oord)) %>%
      merge(merge_df, by.x =c('coord_group_x','coord_group_y','q_pass_only'),
by.y = c('coord_group_x','coord_group_y','event_id')) %>%
      dplyr::mutate(pi = percent_of_pass * norm * (log(total_event_per_coord)
/4)) %>%
      dplyr::group_by(coord_group_x, coord_group_y, Event_player_id, Event_te
am_id) %>%
      dplyr::summarise(sum_pi = sum(pi))


    merge_all = merge(total_passes_per_event_per_coord, total_passes_per_coor
d, by.x=c('coord_group_x','coord_group_y','Event_player_id'),by.y = c('coord_
group_x','coord_group_y','Event_player_id')) %>%
      mutate(total_pi = (sum_pi ))

    sum_all_players = merge_all %>%
      merge(player_info_for_app[,c('uid','name_decoded','real_position', 'rea
```

```
l_position_side')], by.x = c('Event_player_id'), by.y =c('uid'))

    return(sum_all_players)
  }
  # plot function
  player_plot_fn_app = function(i,team_df, team_name){

    player_data = team_df %>% dplyr::filter(Event_player_id== i) %>%
      select(coord_group_x, coord_group_y, count_q_pass, total_pi, name_decod
ed, real_position) %>%
      distinct() %>%
      mutate(coord_group_x = factor(coord_group_x, levels = c(10,20,30,40,50,
60,70,80,90,100))) %>%
      mutate(coord_group_y = factor(coord_group_y, levels = c(10,20,30,40,50,
60,70,80,90,100)))

    player_name =paste0(unique(player_data$name_decoded),'_',team_name,'_',pl
ayer_data$real_position)


    g1 = ggplot(player_data, aes(as.factor(coord_group_x), coord_group_y, gro
up=coord_group_y)) +
      ggtitle(paste0(player_name,'\n',' Number of Events Completed'))+
      geom_tile(aes(fill = count_q_pass)) +
      geom_text(aes(fill = player_data$count_q_pass, label = round(player_dat
a$count_q_pass, 1))) +
      xlab('x_coord\n') +
      ylab("y_coord\n") +
      scale_fill_gradientn(colours = c('green','yellow','orange','red'))+
      theme(panel.background = element_blank(),legend.title = element_blank()
)

    g2 = ggplot(player_data, aes(as.factor(coord_group_x), coord_group_y, gro
up=coord_group_y)) +
      ggtitle('Performance Metric')+
      geom_tile(aes(fill = total_pi)) +     xlab('x_coord\n') +
      ylab("y_coord\n") +
      geom_text(aes(fill = player_data$total_pi, label = round(player_data$to
tal_pi, 1))) +
      scale_fill_gradientn(colours = c('green','yellow','orange','red'))+
      theme(panel.background = element_blank(),legend.title = element_blank()
)


    grid.arrange(g1,g2, ncol =1)

  }
```

```
  ui <- dashboardPage(

    dashboardHeader(title = "Player Output 3060 Model"),
    dashboardSidebar(
        sidebarMenu(
          menuItem("Players", tabName = "player", icon = icon("user")),
          menuItem("Performance Indicator", tabName = "pi", icon = icon("bar-
chart")),
          #menuItem("All Teams Summary", tabName = "allTeam"),
          selectInput("wavg",'Calculation Type', multiple = FALSE,
                               choices = c('Standard','Weighted Average')),
          radioButtons("radio", label = h3("Importance Variable"),
                               choices = list("Mean Decrease Gini" = 1, "Mean
Decrease Accuracy" = 2),selected = 1)
        )
      ),

    dashboardBody(tabItems(
      tabItem(tabName = "player",
              h2("Player level data"),

              column(3,selectInput("teamInput", "Team Name", multiple= FALSE,
selected = "Arsenal",
                           choices = c(paste0(team_id_index$team_id, ':',team_
id_index$team_name)))),
                             # c('Hull City','Fulham','Liverpool','Chelsea','C
ardiff City','Crystal Palace','Manchester United','West Ham United','West Bro
mwich Albion','Aston Villa','Sunderland','Newcastle United','Manchester City'
,'Norwich City','Arsenal','Southampton','Swansea City','Tottenham Hotspur','E
verton','Stoke City')),
              column(4,uiOutput("playerOutput")),
              uiOutput("gameOutput"),
              plotOutput("playerHeatmaps"),
              dataTableOutput("teamTable")


      ),

      tabItem(tabName = "pi",
              h2("Performance Indicator Overview"),
              selectInput("eventInput",'Event Name', multiple = FALSE,
                           choices = paste0(event_qid_labels$evetid, ':',event
_qid_labels$name)),
              plotOutput('piHeatmaps')

      )

    ))
  )
```

```
  server <- function(input, output) {

    output$playerOutput <- renderUI({
      selectInput("playerInput", "Player Name",
                  player_info_for_app %>%
                  select(uid, name, team_name) %>%
                  filter_(paste0("team_name =='",gsub('.*:','',input$teamIn
put),"'")) %>% distinct(uid, name) %>%
                  mutate(str_name = paste0(uid, ':', name)) %>% select(str_
name))

    })

    output$gameOutput <- renderUI({
      if(is.null(input$teamInput))
        return()
      team_id = gsub(':.*','',input$teamInput)
      games_incl_chosen_team = df_for_app %>% filter(Event_team_id == team_id
) %>% group_by(Game_id,Game_home_team_name, Game_away_team_name) %>% dplyr::s
ummarise() %>%
          mutate(game_str = paste0(Game_id,':',Game_home_team_name, ' Vs ', G
ame_away_team_name))
      selectizeInput("gameInput", "Match",multiple = TRUE,options = list(plac
eholder = 'choose match'),
                     choices = games_incl_chosen_team$game_str)

    })


    output$playerHeatmaps <- renderPlot({
      if(is.null(input$playerInput) | is.null(input$teamInput))
        return()
      if(is.null(merge_imp2)|is.null(merge_imp2_wavg)|is.null(merge_imp2_mda)
|is.null(merge_imp2_mda_wavg))
        return()

      if(input$wavg == 'Standard' & input$radio =='1'){
        merge_df = merge_imp2}
      else if(input$wavg == 'Weighted Average' & input$radio =='1'){
        merge_df = merge_imp2_wavg
      }
      else if(input$wavg == 'Standard' & input$radio =='2'){
        merge_df = merge_imp2_mda
      } else if(input$wavg == 'Weighted Average' & input$radio =='2'){
        merge_df=merge_imp2_mda_wavg
      }


      if(is.null(input$gameInput)){
        game_id = c(unique(df_for_app$Game_id))
```

```
      } else {
        game_id = c(gsub(':.*','',input$gameInput))
      }


      player_id = gsub(':.*','',input$playerInput)
      team_id = gsub(':.*','',input$teamInput)
      team_df = get_player_data(team_id, merge_df,game_id)
      team_name =gsub('.*:','',input$teamInput)
      hm1 = player_plot_fn_app(player_id,team_df,team_name)
      hm1
    })

    output$piHeatmaps <- renderPlot({
      if(is.null(input$playerInput) | is.null(input$teamInput))
        return()
  if(is.null(merge_imp2)|is.null(merge_imp2_wavg)|is.null(merge_imp2_mda)|is.
null(merge_imp2_mda_wavg))
    return()

      if(input$wavg == 'Standard' & input$radio =='1'){
        merge_df = merge_imp2}
      else if(input$wavg == 'Weighted Average' & input$radio =='1'){
        merge_df = merge_imp2_wavg
      }
      else if(input$wavg == 'Standard' & input$radio =='2'){
        merge_df = merge_imp2_mda
      } else if(input$wavg == 'Weighted Average' & input$radio =='2'){
        merge_df=merge_imp2_mda_wavg
      }


      pi_data = merge_df %>% filter_(paste0("event_id =='",gsub(':.*','',inpu
t$eventInput),"'")) %>%
        select(event_id, coord_group_x, coord_group_y, norm) %>%
        distinct() %>%
        mutate(coord_group_x = factor(coord_group_x, levels = c(10,20,30,40,5
0,60,70,80,90,100))) %>%
        mutate(coord_group_y = factor(coord_group_y, levels = c(10,20,30,40,5
0,60,70,80,90,100)))


      g1 = ggplot(pi_data, aes(as.factor(coord_group_x), coord_group_y, group
=coord_group_y)) +
        ggtitle(paste0('Performance Indicator - ball recovery and shots on ta
rget\n',pi_data$event_id))+
        geom_tile(aes(fill = norm)) +
        geom_text(aes(fill = pi_data$norm, label = round(pi_data$norm, 2))) +
        xlab('x_coord\n') +
        ylab("y_coord\n") +
```

```
        scale_fill_gradientn(colours = c('green','yellow','orange','red'))+
        labs(fill="\nActual\n\n")+
        theme(panel.background = element_blank(),legend.title = element_blank
())
      g1
    })

    output$teamTable = renderDataTable({

      if(is.null(input$playerInput) | is.null(input$teamInput))
        return()
      if(is.null(merge_imp2)|is.null(merge_imp2_wavg)|is.null(merge_imp2_mda)
|is.null(merge_imp2_mda_wavg))
        return()

      if(input$wavg == 'Standard' & input$radio =='1'){
        merge_df = merge_imp2}
      else if(input$wavg == 'Weighted Average' & input$radio =='1'){
        merge_df = merge_imp2_wavg
      }
      else if(input$wavg == 'Standard' & input$radio =='2'){
        merge_df = merge_imp2_mda
      } else if(input$wavg == 'Weighted Average' & input$radio =='2'){
        merge_df=merge_imp2_mda_wavg
      }


      if(is.null(input$gameInput)){
        game_id = c(unique(df_for_app$Game_id))
      } else {
        game_id = c(gsub(':.*','',input$gameInput))
      }


      team_id = gsub(':.*','',input$teamInput)
      team_df = get_player_data(team_id, merge_df,game_id)
      total_player = team_df %>% distinct() %>% group_by(Event_player_id, nam
e_decoded, real_position)  %>% dplyr::summarise(sum_player = sum(total_pi, na
.rm=TRUE)) %>% arrange(-sum_player)
      return(total_player)
    })



  }
  shinyApp(ui, server)
```

Shiny applications not supported in static R Markdown documents