Defining Soccer Playing Styles through a Data-Driven Approach

Undergraduate Thesis Interim Report

Author: Mohammad Mustafa Arif
Supervisor: Professor Timothy Chan

Student Number: 1003116736
Engineering Science: Machine Intelligence
University of Toronto
February 1st, 2021

Contents

1 Introduction **2**

2 Literature Review **3**

3 Data **6**

4 Progress to Date **7**

4.1 Data Preprocessing 7

4.2 Exploratory Data Analysis 8

4.3 Zonal Networks 10

4.4 Player Networks 14

4.5 Defense Analysis 16

4.6 Team Clustering 18

5 Future Work **21**

6 Appendix: List of Derived Metrics **22**

7 References **23**

1 Introduction

Soccer is the most popular sport in the world. Today, more than 240 million people play soccer regularly, and a combined 3.5 billion viewers tuned in to the 2018 FIFA World Cup broadcast [1][2]. Despite its rich history, the first record of analytics in soccer is from 1996, when Opta Sports, a British sports analytics company, began recording match statistics for the English Premier League [3]. Since then, the field has rapidly evolved and teams have begun using data-driven approaches for tactical decision making. With increased data collection, novel methodologies in statistics, computer science, and network theory have been employed to study teams and exploit strengths and weaknesses. Recently, Liverpool FC, a first division English soccer club, announced the development of "pitch control", a probabilistic model used to quantify space creation during matches in real time [4].

Traditional methods in soccer analytics involve utilizing high level match statistics such as possession percentages, numbers of shots, passes, crosses, fouls, and offsides. However, these methods are prone to oversimplifying 90+ minute interactions between two teams to a set of global statistics. An increasingly popular approach to add nuance to these methods is the use of expected goals (xG), a metric to measure the quality of shots taken by their probability of resulting in a goal [5]. In addition, modern approaches in literature have attempted to study team passing profiles through the distribution of passes amongst the players. However, these approaches are also prone to oversimplification; they do not distinguish between different types of passes or the distribution of passes across different regions of the pitch. Furthermore, they do not explore defensive activity with respect to team performance.

The first objective of this project is to expand on existing literature characterizing team passing distributions and fill in the aforementioned gaps by:

1. Augmenting the passing distributions with descriptive metrics measuring deeper passing statistics using network theory.

2. Incorporating defensive metrics: the distribution of tackles, interceptions, and fouls to represent team performances.

The appropriate features are derived from the Wyscout spatio-temporal soccer event logs, a public dataset with reports of soccer match events marked by time, spatial coordinates, players involved, and a standardized taxonomy of event tags [6]. The dataset covers Europe's top five domestic leagues for the 2017/2018 season, as well international fixtures

in the 2016 Euros and the 2018 World Cup. Some features are engineered though direct manipulation of the logs, while others require a deeper analysis through the construction of passing networks across players and pitch-zones. After engineering the features describing each performance, feature importance will be measured via calculating correlation coefficients and mutual information with respect to some measure of performance. The purpose of this exercise is to gauge which features contain the most descriptive power about a team's performance.

Next, features for each team will be aggregated for the entire season, and teams will be categorized into distinct playing styles. For this step, unsupervised clustering methods such as K-Means and hierarchical clustering, along with probabilistic soft-clustering algorithms such as Gaussian mixture models will be used. Once cluster assignments have been established, they will be analysed from a fundamental perspective to attempt an intuitive interpretation of the established playing styles. Finally, performances will be analysed with respect to playing styles study if there are certain playing styles that tend to overperform/underperform in comparison to others.

Thus, the desired outcome of this project is to arrive at classification of teams in terms of a feature space spanned by custom-defined metrics. The project aims to study these playing styles from an intuitive perspective and analyze any discrepancies in effectiveness/performance across these styles.

2 Literature Review

Increasingly available data has facilitated recent developments in the use of analytics in soccer. Soccer-logs capturing all events within a match have commonly been used to study many aspects of the game [6]. In particular, there exists literature investigating certain squads and explaining unexpected performance patterns through a data-driven lens. There have also been notable research attempts to derive representations and insights via statistical methods for teams as well as individual players. While this project is framed at segmenting teams into characteristic playing styles and investigating performance discrepancies across these playing styles, the following literature survey explores all the aforementioned sub-domains as the methodology discussed can be reapplied in the context of this project.

One key question individual player analyses address is: What makes a player "good" or "bad"? A popular method to quantify player quality using statistics is to use information from in-game events for extracting information. For example, Pappalardo et al.

[7] use soccer-logs to design and implement PlayeRank, a role-aware, multi-dimensional evaluation framework for individual soccer players, tuned through team performance outcomes. They validate the performance by testing the algorithm against ground truth ratings from professional soccer scouts, achieving a significant level of agreement. They also demonstrate the use of PlayerRank to distinguish top players from others, and measure player versatility. As such, Pappalardo et al. addressed issues with the time's simple data-driven player performance metrics by:

1. Designing a "role" aware system: The evaluation function is dynamic with respect to a player's role on a team.

2. Evaluating performance across many dimensions instead of reporting a scalar score combining all aspects of player performance.

However, a major drawback of this method is its limited scope at individual player level. Analysing squads of 11 players (with possible substitutions and injuries) becomes an increasingly challenging task with this framework. In addition, the input features for PlayeRank are solely derived from individual player activity and fail to measure interactions between players or "team chemistry". Nevertheless, the use of unsupervised clustering with the K-Means algorithm on spatial features for role detection is a robust method, and can be generalized, with the help of involved feature engineering, to segment teams into playing styles.

To address the intrinsic drawbacks of individual player analysis, attempts have been made to identify team performance metrics from a higher level of abstraction. For instance, Cintia et al. [8] extract five separate team metrics summarizing passing behaviour amongst players and across zones of the pitch, eventually combining them via a weighted harmonic mean into a single indicator, the 'H-Statistic'. They report a significant correlation between the H-Statistic and performance measures and explain game outcomes based on the H-Statistic difference across the teams. Further, they demonstrate out-of-sample predictive power in the H-Statistic by utilizing various classifiers to predict match outcomes through a season based exclusively on past H-Statistic values and find high agreement between real and simulated rankings. The study successfully managed to quantify passing behaviour across teams, but one limitation is that only five simple statistics are used to describe passing: mean and variance of passes across players, mean and variance of passes across zones, and total passing volume. These shortcomings present a requirement for more expressive features to describe team passing, and other aspects of the game such as defense.

Another popular way to study teams and player-player interactions is through network science. Originally proposed by Gould and Gatrell in 1979 [9], the idea of constructing passing networks from team performances gained popularity after Duch et al. demonstrated in 2010 [10] that player flow centrality in passing networks, a measure of the frequency of the player being involved in paths resulting in a shot, can be used to quantify the contribution of individual players in team performances. More recently, Clemente et al. [11] employed the use of network science to study team performances in the 2014 World Cup and demonstrated to a statistically significant degree that large connectivity between teammates is associated with better overall team performance. In their 2019 article, Buldu et al. [12] also utilized metrics engineered using network science to compare Pep Guardiola's FC Barcelona team from the 2009/2010 season, considered one of history's best soccer teams, with the rest of the teams in the Spanish first division. They found a significant difference in the advance ratio, a measure of passing directness, between FC Barcelona and the rest of the teams in the league, indicating that FC Barcelona's squad employed a relatively indirect approach to passing. Moreover, they ran graph algorithms to compute the clustering coefficient (a measure of a network's local robustness), average topological shortest path, largest eigenvalue of the connectivity matrices, and algebraic connectivity (a quantification of the team's integration or fault tolerance). Across all these metrics, they found statistically significant differences between FC Barcelona and the rest of the teams in the Spanish league. Thus, Buldu et al. [12] defined a network science framework for studying the playing style for a specific team against its rivals. For this project, their methodology can certainly be augmented to incorporate features measuring other aspects of the game beyond passing, and generalized to compute numeric representations for more teams across multiple leagues to find commonalities/differences and discern team playing styles.

Each of the works discussed approached soccer analysis through a statistical framework exploiting soccer-logs data-sets. However, they do not apply their methods to general team playing style identification. Furthermore, relatively little exploration is done beyond passing, especially for defensive attributes. This project will therefore attempt to expand on methodology described above to fill in these gaps.

Finally, note that the use of soccer-logs is just one of many different approaches for utilizing analytics in soccer. There have also been notable attempts to exploit video footage, GPS tracking, and player physiological signals [13][14]. These approaches, however, are fundamentally different in scope to the problem addressed and the data used in this project. Thus, they are not discussed comprehensively here.

3 Data

For this project, the data being used is sourced from a public, spatio-temporal dataset of soccer-logs spanning Europe's top 5 domestic leagues for the 2017/2018 season: Spanish first division, Italian first division, English first division, German first division, and French first division [6]. In addition, international games from the 2018 World Cup and the 2016 European Cup are also covered. The data is provided by Wyscout, a leading company in the soccer industry, and is collected through a 3 step process.

1. Expert video analysts set team formation at the beginning of each game. This includes mapping the on-field players to their positions as well as the list of available players on the bench.

2. For each touch on the ball, the analysts, using a propriety tagger software, select one player, and create a corresponding event on the timeline. The event description involves specifying an event type and sub-type, along with the spatial coordinates, and other special tags to specify additional attributes.

3. The logs are quality controlled, both algorithmically and through manual cross comparisons.

The exact taxonomy of the events, sub-events, and tags can be found summarised in Table 1[1]. Altogether, the dataset covers 1941 games, 3,252,294 events, and 4,299 players.

Table 1: Dataset Event Taxonomy

Event	**Sub-Events**	**Tags**
pass	cross, simple pass	accurate, not accurate, key pass, opportunity, assist, goal
foul	-	no card, yellow, red, 2nd yellow
shot	-	accurate, not accurate, block, opportunity, assist, goal
duel	air duel, dribbles, tackles, ground loose ball	accurate, not accurate
free kick	corner, shot, goal kick, throw in, penalty, simple kick	accurate, not accurate, key pass, opportunity, assist, goal
offside touch	acceleration, clearance, simple touch	counterattack, dangerous ball lost, missed ball, interception, opportunity, assist, goal

[1]Table taken from [6].

4 Progress to Date

This section will summarize the methodology, implementation challenges, and early phase results, keeping in mind the overarching goal to eventually arrive at a classification of distinct playing styles for soccer teams. In broad-strokes, the progress to date can be broken down into six categories:

1. Data Preprocessing
2. Exploratory Data Analysis
3. Zonal Networks
4. Player Networks
5. Defense Analysis
6. First Round of Unsupervised Clustering

The following sub-sections will elaborate more on work done under each of these categories.

4.1 Data Preprocessing

The soccer-logs dataset used in this project was released under the CC BY 4.0 License, and available on figshare[2]. While the dataset was provided in a standard JSON format, it comprised of files with large memory footprints. Therefore it was not feasible to open and explore the dataset on local PCs, and consequently, cloud services were used to clean the data and filter out fields and tables not required for the project. For this task, the Standard D8s v3 virtual machine from Azure was chosen as it provided a large enough RAM of 32 GiB to accommodate the dataset. In addition to filtering out unnecessary information, the original dataset was decomposed into smaller files, and translated into a more user-friendly, tabular format. The process is summarised in Figure 1:

The information removed at the preprocessing stage included details regarding referees and coaches, as well as events pertaining to international games in the 2018 World Cup and the 2016 European Championship. Note that these international games were

[2]The dataset is publicly posted at https://figshare.com/collections/Soccer_match_event_dataset/4415000

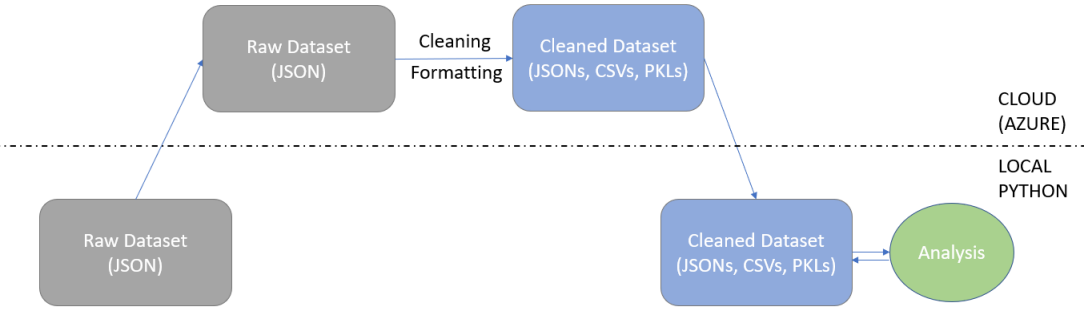


Figure 1: Data Prepossessing

intentionally excluded from the scope of the project because the volume of data obtained per team was severely limited in comparison to the five domestic leagues for the 2017/2018 season. This is a consequence of the longer format of the domestic leagues in which each team plays 19 games, as opposed to elimination style international tournaments. The scope of this project was thus narrowed down to classify playing styles exclusively over long-form domestic leagues.

4.2 Exploratory Data Analysis

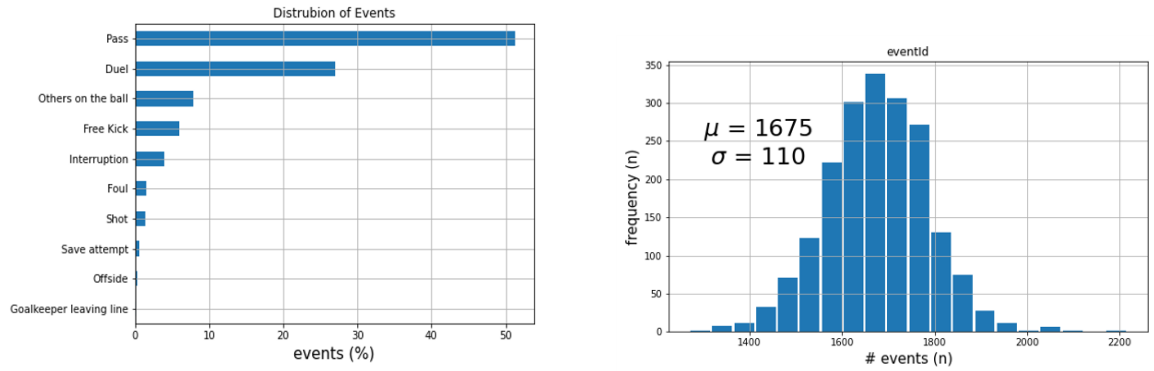
After the dataset was prepossessed and made feasible to open locally, an exploratory data analysis was conducted to gauge the information available within the dataset. First, the data coverage was verified to note the number of games, events, and players from each of the domestic leagues. The findings are summarised in Table 2:

Table 2: Matches, Events, and Player Counts

League	**Matches**	**Events**	**Players**
La Liga (Spain)	380	628,659	619
Premier League (England)	380	643,150	603
Serie A (Italy)	380	647,372	686
Ligue 1 (France)	380	632,807	629
Bundesliga (Germany)	306	519,407	537

Note that with the exception of the German Bundesliga, all the leagues comprised of exactly 380 games. The reason for the discrepancy is that while all the other leagues have 20 teams, each playing 19 games a season, the Bundesliga is structured to have 18 teams, each playing 17 games per season. A breakdown of event and subs-events counts was conducted, as shown in Figure 2a. By far, the most common event was a pass, followed

by a duel. The distribution of the number of events per game was analysed as shown in Figure 2b. It appears to be fairly symmetric around a mean value of 1675 events per game.

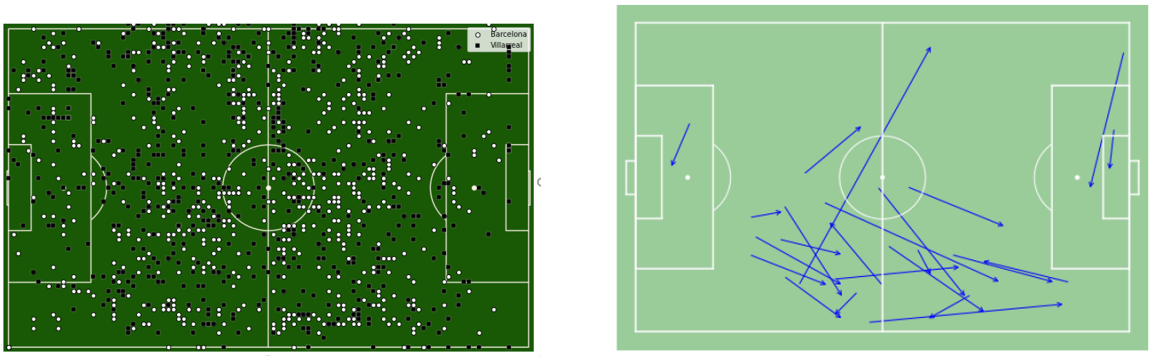


(a) Event Percentages

(b) Events Per Game

Figure 2: Event Breakdowns and Event Counts per Game

In the exploratory phase, the spatial aspect of the events was also studied. As noted previously, each event is marked by a set of coordinates. In the case of passes, the origin and destination points are defined separately. All coordinates are two dimensional and provided in units of f.u. (field units)[3], ranging from 0 to 100 in both the x and y dimensions. Using these coordinates, visualizations for events in a game as well as passes conducted by single players were derived, as shown in Figure 3. Note that the process of constructing these visualizations was standardized in the exploratory data analysis phase in case metrics derived from spatial information in the later phases of the project needed to be analysed visually at a low level.



(a) All Events: FC Barcelona 5 - 1 Villareal

(b) Passes From a Central Midfielder

Figure 3: Spatial Data Visualization with respect to the Pitch

[3]The x measures nearness to the opponents goal, with $x = 100$ indicating the opponent's goal line. Similarly the y coordinate represents the nearness to the right side of the pitch. Field units conveniently provide a standardization for spatial coordinates as not all pitches share exactly the same dimensions.

Finally, visualizations of the temporal nature of the games were also constructed. Because this project is framed at analysing team performance per game (as opposed to breaking the performance down by smaller time intervals), much of the intra-game temporal data is not thoroughly studied except for the construction of passing networks. Nevertheless, some visualizations were made to gauge how the distribution of goals vary with respect to time. As seen in Figure 4, the second half of the game is more eventful in terms of goals across all the domestic leagues.

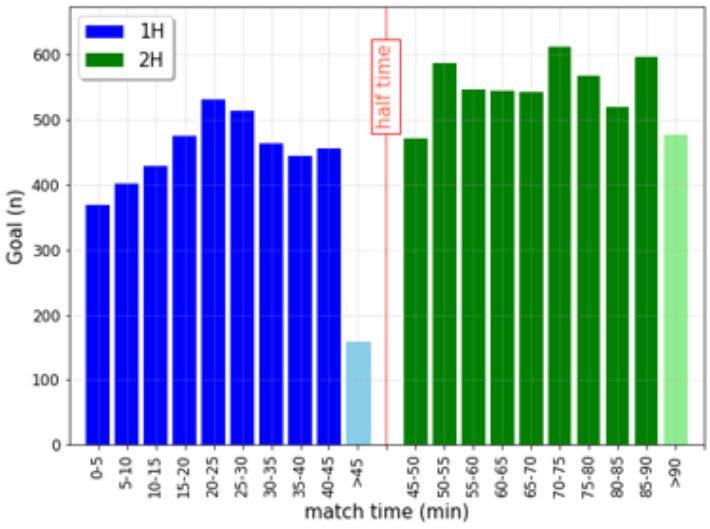


Figure 4: Distribution of Goals over Time

With the data prepossessed and analysed at a high level, and frameworks for visualizations in place, the exploratory data analysis phase concluded and paved way for more involved analysis.

4.3 Zonal Networks

This section describes the spatial analysis of team passing activity. The study involved constructing metrics from zonal network graphs to investigate the inter-zone passing activity, and concluded with the derivation of multi-dimensional representations to describe, from a spatial lens, the passing profiles per team, per performance.

The study began by discretizing the soccer pitch evenly into 9 non-overlapping zones, as illustrated in Figure 5a. Choosing to discretize the pitch this way provided an intuitive mapping between the zones and soccer positions: defense, midfield, and forward on the left, central, and right sides of the pitch. However, a drawback of this method is that the large zone sizes conceal the activity within the zones, as analysis was conducted at the inter-zone level, not the intra-zone level.

(a) 9 Discretized Zones for Inter-Zone Passing Analysis

(b) Team Zonal Network Graph for a Single Performance

Figure 5: Inter-zone Analysis

After defining the zones, zonal network graphs were produced for each game, for each team. The nodes in these graphs represented the 9 zones while the edges corresponded to the inter-zone passing. The graphs were directed and weighed, with weights being set by the frequency of passes along the corresponding edge. One example of the zonal network graph can be seen in Figure 5b. Here, the boldness of the edges indicates the frequency of passes along the paths. It is visually apparent that R5-R8 was a key passing lane (by passing volume) in this performance.

Building these network graphs enabled reconstructing some of the elemental building blocks of the H-Statistic. Using the methodology from Cintia et al. [8], each zone was attributed the number of passes it "handled" by summing up the in-degree and out-degree for the corresponding node on its network graph. Then, the mean and standard deviation of the passes handled were calculated across the zones. Both these quantities were normalized by the total passing volume of the team in that performance. Equations 1-3 show illustrate these definitions for team t in a certain performance.

$$x_i^t = InDegree_i^t + OutDegree_i^t \quad \forall\ i \in [0, 8] \tag{1}$$

$$\mu_{zone}^t = \frac{1}{\#passes^t} \times mean_i(x_i^t) \tag{2}$$

$$\sigma_{zone}^t = \frac{1}{\#passes^t} \times stdev_i(x_i^t) \tag{3}$$

Note that a high value of σ_{zone}^t implies the coexistence of "hot" zones with high passing activity and "cold" zones with lower passing activity for team t in the performance being analysed. In contrast, a low σ_{zone}^t indicates a more uniform distribution of passing

across the zones. As shown in Figure 6, σ^t_{zone}, macro-averaged over all performances by team t in the 2017/2018 season, shows a positive correlation with end-of-season points, with $\rho = 0.59$. The correlation seems to be consistent across all 5 domestic leagues.

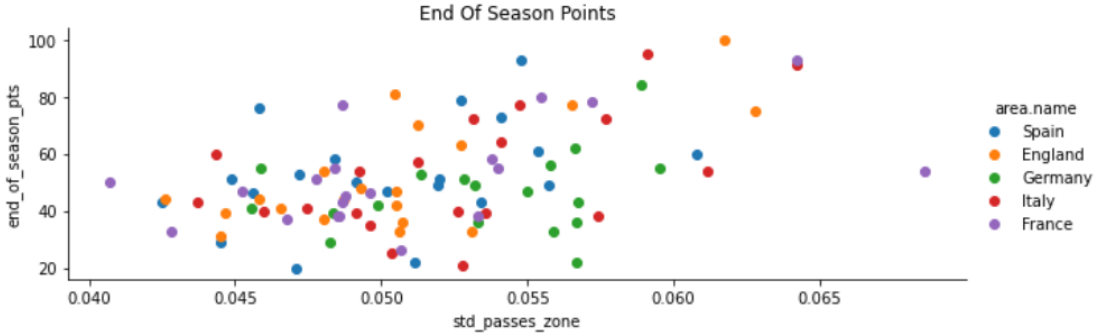


Figure 6: End of Season Points vs Macro-Averaged σ^t

Zonal network graphs also enabled the identification of key passing lanes. For instance, from Figure 5b it is evident that R5-R8 is indeed a key passing lane (the boldness of the edge signifies the frequency of passing). This begged the question: Is it possible to segment performances through passing lane activity? If so, are there certain passing lanes that are linked to higher performance outcomes? To answer these questions, the original zonal networks were first transformed into an array of features indicating the edge weights. This yielded a 36 dimensional vector, which was then normalized by the L1 norm[4] to ensure that the weights add to 1. Equations 4-6 formally describe this process. Note that $\mathbf{w} \in \mathbb{R}^{36}$ is a 36 dimensional column vector, with element $w_{i,j}$ corresponding to the passing lane from zone i to zone j.

$$w_{i,j} = ZoneNetworkEdge(i, j) \tag{4}$$

$$\mathbf{w}_{raw} = [w_{i,j}] \ \forall \ i, j \in [0, 8] \ s.t \ i \neq j \tag{5}$$

$$\mathbf{w} = \frac{\mathbf{w}_{raw}}{||\mathbf{w}_{raw}||_1} \tag{6}$$

Thus, for each performance, each team had a corresponding $\mathbf{w}$ vector with the 36 elements representing the 36 passing lane intensities. The set of $\mathbf{w}$ vectors for all teams across all performances of the season was then clustered in vector space using the K-Means algorithm with the Euclidean distance metric to determine distinct passing styles from passing lane intensity [15]. Setting K=2 through the elbow method, two distinct clusters were obtained.

[4]The L1 norm was chosen to make it such that the elements in the final vector indicate the percentage of all inter-zone passes that went through the corresponding lane

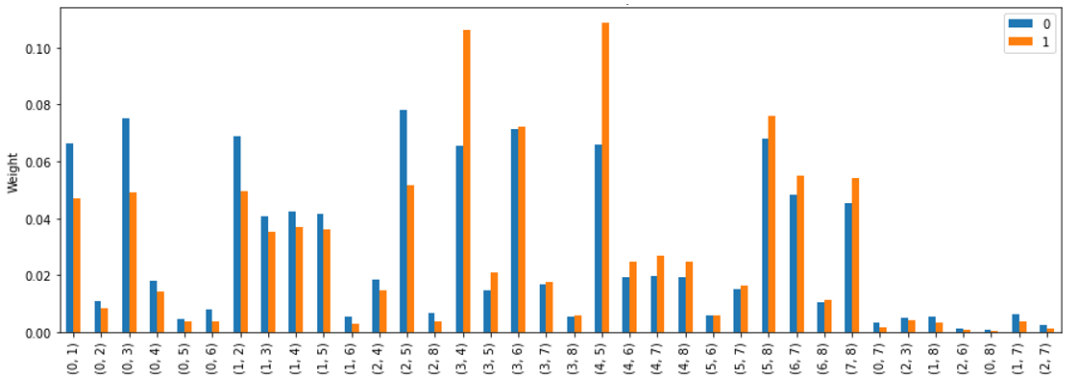
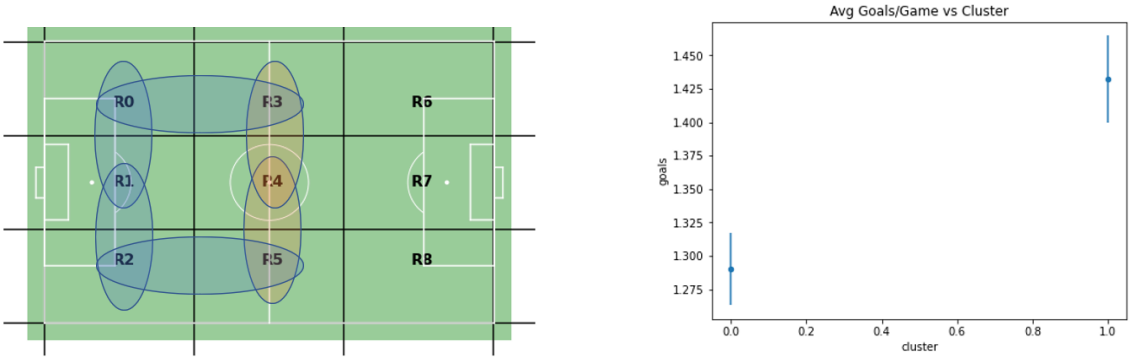


Figure 7: Centroids for the 2 Clusters from Passing Lanes

Figure 7 displays a component-wise comparison for the 2 cluster centroids with regards to the relative weight/intensity along each of the components. It is evident that R3-R4 and R4-R5 stand out as key passing lanes in cluster 1 (orange) while R0-R1, R0-R3, R1-R2, and R2-R5 stand out as key passing lanes in cluster 0 (blue). These key lanes are illustrated in terms of the pitch itself in Figure 8a. The performance discrepancy across performances belonging to the clusters is also apparent, as seen in Figure 8b. In this figure, the average goals per game, along with the standard error bars can be seen for all performances binned in cluster 0 and cluster 1 respectively. Clearly, cluster 1 is associated with more goals per game, on average, than cluster 0. From a fundamental perspective, this corresponds to one popular philosophy in soccer that high lateral passing in the midfield is a favorable quality for teams to have.



(a) Key Passing Lanes for Cluster 0 (Blue) and Cluster 1 (Orange)

(b) Performance Discrepancy across Clusters

Figure 8: 2 Cluster Passing Profile Contrast

To verify the correlation of increased lateral passing with performance, a finer grain analysis was done. Moving away from the 9-zone discretization and zonal network graphs, each pass was isolated in terms of its origin and destination coordiantes, and its advance ratio was computed. The advance ratio refers to the pass's lateral (left/right) trajectory as a ratio of its vertical (towards/away from goal) trajectory. Equation 7 provides a

mathematical definition of the mean advance ratio for team t in a given performance.

$$MeanAdvanceRatio^t = \frac{1}{\#passes_t} \sum_{i \in passes_t} \frac{|\Delta X|_i}{|\Delta Y|_i} \tag{7}$$

Upon macro-averaging the mean advance ratio per team for all the performances by that team in the 2017/2018 season, a positive correlation with $\rho = 0.68$ was obtained with the end of season points. The relationship is illustrated in Figure 9

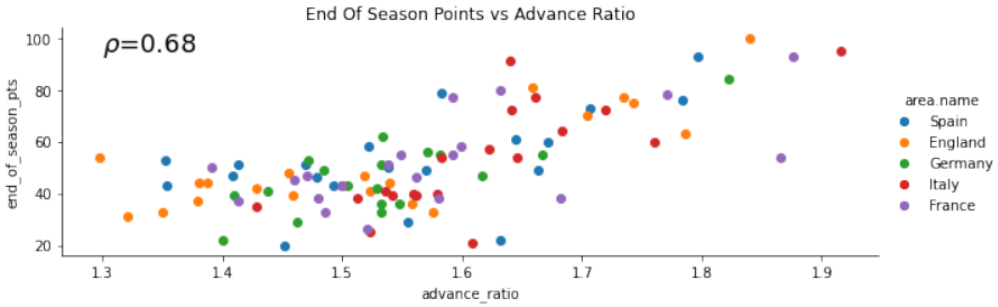


Figure 9: Macro-Averaged Mean Advance Ratio vs End Of Season Points

Thus, zonal network graph analyses yielded 3 metrics do describe passing activity in terms of spatial coordinates: μ^t_{zone}, σ^t_{zone}, and $MeanAdvanceRatio^t$. In addition, the passing lane study yielded 2 distinct clusters in terms of lane-intensities with a performance discrepancy between them. In the larger context, these metrics will be used as components describing team t's performance in a specific game so that playing style segmentation can be done across teams.

4.4 Player Networks

Analogous to zonal networks discussed in the previous section, player passing networks were constructed by defining each player as a node, and passes amongst the players as edges. Note that the soccer-logs dataset [6] did not provide a player ID for the receiver of the pass. Therefore, an assumption had to be made that if a certain player i passes the ball and the next chronological event is from another player j on the same team, then it is assumed that the pass was from i to j. This assumption was necessary to construct player passing networks. A possible way to avoid this assumption in future studies is to use ball-tracking and player-tracking data in conjunction with soccer-logs.

Just as with zonal networks, methodology from Cintia et al. [8] was adopted to attribute to each player the number of passes he "handled" by summing up the in-degree and out-degree for the corresponding node on its network graph. Using equations 8-10,

the mean and standard deviations of passes being handled by individual players was evaluated. These definitions hold a one to one correspondence with the analogous definitions for zonal networks in equations 1-3, the only difference being that nodes corresponded to players instead of zones.

$$x_i^t = InDegree_i^t + OutDegree_i^t \quad \forall\ i \in players(t) \tag{8}$$

$$\mu_{player}^t = \frac{1}{\#passes^t} \times mean_i(x_i^t) \tag{9}$$

$$\sigma_{player}^t = \frac{1}{\#passes^t} \times stdev_i(x_i^t) \tag{10}$$

Again, a high value of σ_{player}^t implied the coexistence of "active" players with high passing activity and "inactive" players with lower passing activity for team t (in the performance being analysed). As before, investigating the player networks yielded a high correlation ($\rho = 0.56$) between the macro-averaged σ_{player}^t values for all performances of team t across the entire season, and the end-of-season points.

Player networks, however, were also analysed using algorithms from graph theory. Borrowing methodology from [12], three metrics were derived from each player network graph[5]:

The **clustering coefficient** of a node u is a measure of the network's local robustness with respect to that node. Intuitively, it counts the number of triangles centered around the node as a fraction of the number of triangles there could have been. In weighed networks, it is weighed by the geometric mean of the weights of the three edges in the local triangle as shown in Equation 11:

$$c_u = \frac{1}{(deg(u))(deg(u)-1)} \sum_{v,w} (w_{uv} w_{uw} w_{vw})^{\frac{1}{3}} \tag{11}$$

where w_{ij} is the weight in the edge from node i to node j. To characterize entire performance, the clustering coefficient over all nodes (players) in the player network was averaged to arrive at a single metric per game, per team.

The **algebraic connectivity** is a measure of integration/segregation of nodes in the network, with a value of 0 indicating complete separation amongst communities. Mathematically, it is defined as the second smallest eigenvalue of the Laplacian matrix

[5]Recall that there is one player network graph per match, per team

of the network. The Laplacian matrix(L) is defined as in Equation 9:

$$L = D - A \tag{12}$$

where $D = diag(d_1, d_2, ...)$ is a diagonal degree matrix formed by the degrees of vertices d_i and A is the adjacency matrix. A high algebraic connectivity can be interpreted as the network being tolerant to faults. In the context of soccer player networks, this means that if a player is marked (by an opponent's defender) or off the field (due to a red card or injury), the network as a whole is tolerant highly to it in terms of ball movement.

The **average all pairs topological shortest path** is a measure of how frequently the ball was passed between pairs of nodes on average. To compute this, the edge weight between nodes i and j was set to $\frac{1}{frequency\ of\ passes\ from\ i\ to\ j}$. A low value indicates movement of the ball between 2 random nodes, on average, was performed frequently.

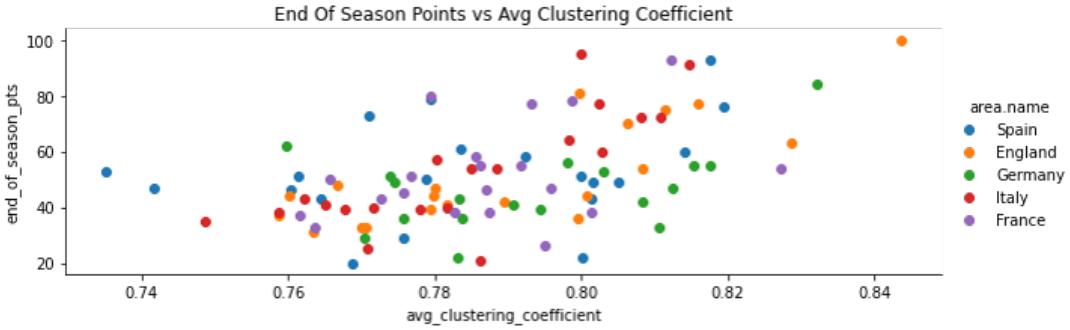
All three of these metrics were computed to describe ball movement across the players. The average clustering coefficient, the algebraic connectivity, and the average all pairs topological shortest path, when macro-averaged for the entire season, yielded fairly linear trends in team performance in terms of end of season points, as shown in Figure 11. The correlation values were $\rho = 0.52$, $\rho = 0.56$, and $\rho = -0.73$ respectively.

To summarize, player network graphs yielded 2 analogous metrics to zonal graphs: μ^t_{player} and σ^t_{player}. In addition, running graph algorithms on these networks yielded the average clustering coefficient, the algebraic connectivity, and the average all pairs topological path. These five metrics provide a way to represent the passing activity amongst players on a team across five dimensions, and eventually these will constitute components in a team's performance representation in order to perform playing style clustering.

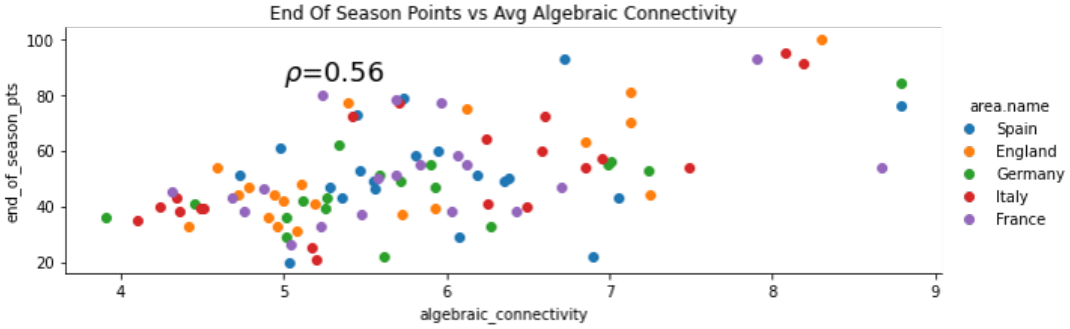
4.5 Defense Analysis

In addition through passing, which was explored via player and zonal network graphs, defensive aspects of games were also explored. Although a graph theory analysis was not conducted to engineer defensive features, a shallower analysis was done and the event tags from the soccer-logs dataset[6] were used to enumerate and aggregate event counts. Table 3 summarizes the events along with their correlation to end-of-season points when macro averaged.

Note that the fraction of slide tackles and interceptions inside the box yielded an overall negative correlation. The next natural questions were: does this pattern continue

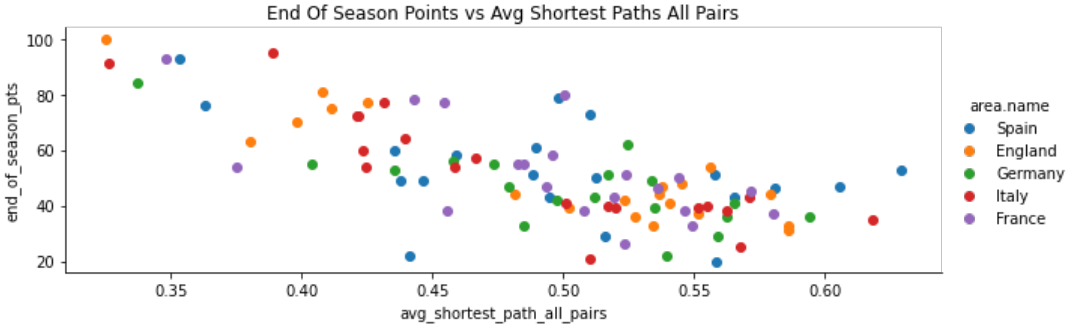


(a) Average Clustering Coefficient Lanes



(b) Algebraic Connectivity Clusters

Figure 10: 2 Cluster Passing Profile Contrast



(a) Average All Pairs Topological Shortest Paths Clusters

Figure 11: Player Network Metrics vs End of Season Points

Table 3: Defensive Events & Correlation to End Of Season Points

Event	**Correlation To EOS Points**
Number of Yellow Cards / Game	-0.43
Dangerous Balls Lost	-0.30
Interceptions Inside Box	-0.66
Total Interceptions	-0.67
Fraction of Slide Tackles in Box	-0.48
Slide Tackles Inside Box	-0.57
Total Slide Tackles	-0.55
Fraction of Slide Tackles in Box	-0.25

further down the pitch? Does the position of defensive events (slide-tackles/interceptions) correlate with performance in general or is it unique to the penalty box? To answer these, the spatial X and Y coordinates were referenced again. More specifically, the centroids of slide tackles and interceptions were computed per game for each team, and the macro-averaged results analysed against the end of season performance (like all the other analyses so far). The macro-averaged X and Y centroids had correlation coefficients of 0.64 and -0.02 respectively with the end of season points. Note the correlation is only apparent in the forward/backward direction. The left/right coordinate is immaterial as far as correlation with performance is concerned. This is in line with expectations as stopping opponent attacks further down the pitch (a higher X coordinate) is less risky because set pieces become more difficult, and the likelihood of conceding a goal from a set-peace decreases.

With these defensive metrics, the overall features space comprised of 20 derived features as summarized in Table 4 [6]. Therefore, with the feature engineering concluded, each performance for each team could be characterized across these 20 dimensions as a 20 dimensional vector. The next section will describe early events to use these 20 dimensional vectors for clustering playing styles.

4.6 Team Clustering

Using the 20 features engineered, a first attempt was made at clustering teams in 20 dimensional space. Note that all the features had initially been tabulated at the per game, per team level. To represent teams as a single set of 20 features, all the performances for

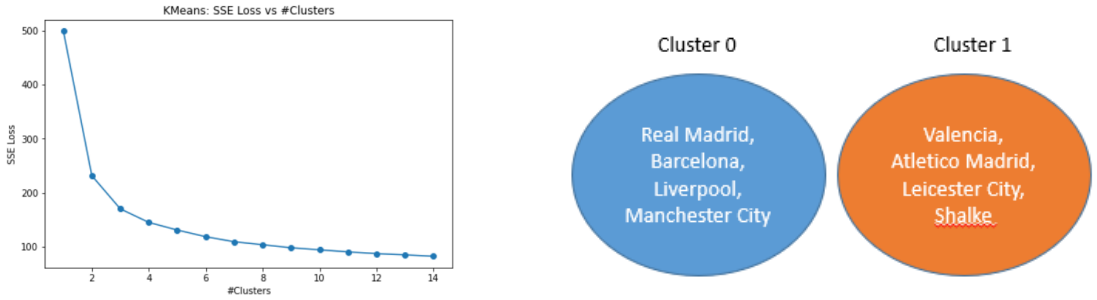
[6]See Appendix for a detailed description for each of the 20 derived metrics

Table 4: A Summary of All Engineered Features

Type of Analysis	**Number of Features Derived**
Exploratory Data Analysis	1
Zonal Networks	3
Passing Lane Clustering	1 cluster label or 36 individual features
Player Networks	5
Count-based Defense Metrics	10

each team were averaged. After each team had been mapped to a 20 dimensional vector, each of the 20 features were standardized across all teams by transforming them into a Z-Score[7]. Next, a K-Means clustering approach with a Euclidean distance metric was explored. It is worth noting that since a standardization on each feature across all teams was performed, an implicit assumption was made that all features are equally important. In future attempts, this assumption may be revisited.

Implementing a similar framework to one used for clustering in section 4.3, the appropriate K for the K-Means algorithm was first found using the elbow method, as shown in Figure 12a. Visually, the elbow was determined to be at K=2.



(a) Sum of Squared Errors vs #Clusters

(b) Sample Cluster Assignments Teams

Figure 12: Team Clustering

Proceeding with 2 clusters, the team-cluster assignments were studied in more detail to find that cluster 1 was two times as large as cluster 0. Sample cluster assignments can be seen in Figure 12b. To ensure that clustering was not purely a function of skill difference, individual performances were also clustered using the centroids found in this investigation, and it was qualitatively verified that team-cluster assignments are consistent across most performances irrespective of the skill difference between the two teams[8]. An interpretation of the clusters was also made by analysing the two cluster centroids

[7]This involved subtracting the average value across all teams and then scaling down by the standard deviation of each feature across teams.

[8]Skill difference was measured by the difference in end of season points. Two teams were said to be

with respect to the 20 dimensional feature space. The centroids differ as shown in Figure 13.

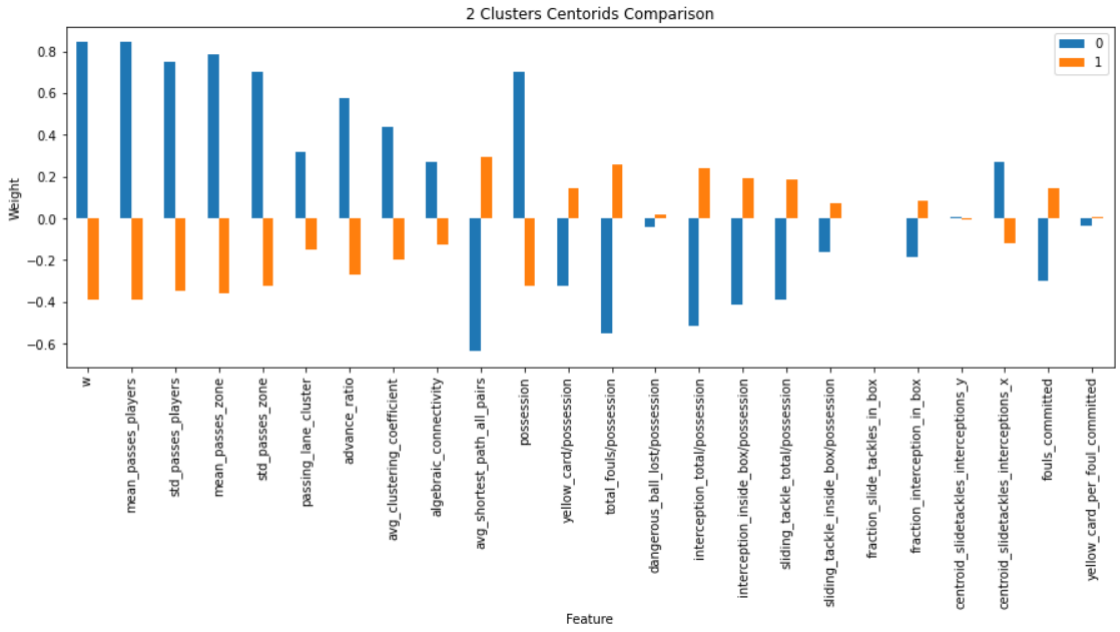


Figure 13: 2 Clusters Segmenting Team Playing Styles

From Figure 13 a clear discrepancy between an attacking and a defensive playing styles can be observed. The attacking style cluster centroid (cluster 0, indicated in blue) shows above average scores on attacking metrics such as advance ratio, possession, algebraic connectivity, clustering coefficient, etc... In contrast, the defensive cluster centroid (cluster 1, indicated in orange) shows below average scores on all these features and an above average score on defensive metrics such as total fouls, slide tackles and interceptions.

This exercise was a first attempt at clustering, and moving on, the playing style segmentation will be repeated with different clustering techniques and possibly feature selection/re-weighing. Performance discrepancy across the clusters will also be studied. Finally, attempts will be made to arrive at intuitively interpretable cluster assignments for K values beyond K=2 so that more nuanced playing styles can be extracted.

at the same skill level if their end-of-season points were within 10 points of each other. Otherwise the skill imbalance was noted.

5 Future Work

As a recap, the desired outcome of this project is to arrive at a meaningful classification of team playing styles in terms of a feature space spanned by custom-defined metrics. As an immediate next step, the two clusters discussed on section 4.6 need to be interpreted more rigorously and analysed with respect to performance i.e. answering if one of the clusters, on average, perform better. If a significant performance discrepancy is found, an attempt will be made to explain any anomalies that may exist. The study could also be generalized to a higher K to derive more nuanced playing styles.

Over the next 3 months, the 20 derived features would be studied from an information gain perspective. More specifically, mutual information between the features and some performance measure would need to be computed. These mutual information values could be used to eliminate noise features (features that do not contain substantial information about performance). They could also be used to weigh the 20 features by their mutual information in geometric space, essentially assigning a higher feature importance to features that contain more information about performance in the K-Means clustering algorithm [15]. Other clustering methods such as hierarchical clustering could be employed to independently cluster passing and defensive attributes [16].

Going beyond K-Means, other probabilistic clustering approaches such as Gaussian Mixture Models and soft K-Means models will be explored [17][15]. The key advantage with these methods is that they do not assign a team to one particular cluster; their probabilistic nature allows for a more continuous probabilistic assignments to multiple clusters, which can make interpretation and anomaly explanation clearer. Finally, the clustering and regression with integer optimization framework by Bertsimas and Shioda [18] can be used to cluster teams based on how the performance outcomes for the season change with changes in the feature space. This would allow feature by feature analysis, and under certain assumptions, making recommendations to teams belonging to different clusters to maximize end of season performance.

If at any stage feature re-engineering is required, or the current 20 dimensional feature space is deemed too inexpressive for playing style segmentation, more features would need to be derived to describe more soccer performances with greater expressivity. This could potentially involve changing the original 9 zone discretization approach to a finer grain one, using more graph algorithms on player and zonal graphs, and developing a deeper framework to analyze defense.

6 Appendix: List of Derived Metrics

Metric	Description
Passing Volume	Number of passes (successful/unsuccessful) executed by a team in a given match.
μ^t_{zone}	Mean number of passes "handled" by each of the 9 zones
σ^t_{zone}	Standard deviation of passes "handled" across the 9 zones
$MeanAdvanceRatio^t$	The average ratio of the lateral trajectory to the vertical trajectory for all passes in a game
Passing Lane Cluster	A binary value indicating which cluster the passing lane profiles belong to according to Section 4.3
$\mathbf{w}$	A 36 dimensional vector measuring passing lane intensities
μ^t_{player}	Mean number of passes "handled" by each of the player
σ^t_{player}	Standard deviation of passes "handled" across the players on a team
Average Clustering Coefficient	A measure of a player network's local robustness
Algebraic Connectivity	A measure of a player's network's segregation and fault tolerance
Average All-Pairs Topological Shortest Path	A measure of how frequently the ball was passed between two players, on average
Number of Yellow Cards / Game	Aggregated per match, per team
Dangerous Balls Lost	Aggregated per match, per team
Interceptions Inside Box	Aggregated per match, per team
Total Interceptions	Aggregated per match, per team
Fraction of Slide Tackles in Box	Aggregated per match, per team
Slide Tackles Inside Box	Aggregated per match, per team
Total Slide Tackles	Aggregated per match, per team
Fraction of Slide Tackles in Box	Aggregated per match, per team
$\mathbf{X}$	X centroid of slide tackles and interceptions
$\mathbf{Y}$	Y centroid of slide tackles and interceptions

7 References

[1] FIFA.com, *Who we are - news - fifa survey: Approximately 250 million footballers worldwide*. [Online]. Available: https://www.fifa.com/who-we-are/news/fifa-survey-approximately-250-million-footballers-worldwide-88048.

[2] Hermesauto, *Football: 2018 world cup watched by record 3.5 billion people, says fifa*, Dec. 2018. [Online]. Available: https://www.straitstimes.com/sport/football/football-2018-world-cup-watched-by-record-35-billion-people-fifa.

[3] J. Prince-Wright, *How opta altered the premier league, and soccer, forever - prosoccertalk: Nbc sports*, Oct. 2013. [Online]. Available: https://soccer.nbcsports.com/2013/10/02/how-opta-has-helped-alter-the-premier-league-and-soccer-forever-part-i/.

[4] J. Williams, *Liverpool are using incredible data science, and match effects are extraordinary*, Jan. 2020. [Online]. Available: https://www.liverpool.com/liverpool-fc-news/features/liverpool-transfer-news-jurgen-klopp-17569689.

[5] *Advanced metrics*. [Online]. Available: https://www.optasports.com/services/analytics/advanced-metrics/.

[6] L. Pappalardo, P. Cintia, A. Rossi, E. Massucco, P. Ferragina, D. Pedreschi, and F. Giannotti, "A public data set of spatio-temporal match events in soccer competitions," *Scientific Data*, vol. 6, no. 1, 2019. DOI: 10.1038/s41597-019-0247-7.

[7] L. Pappalardo, P. Cintia, P. Ferragina, E. Massucco, D. Pedreschi, and F. Giannotti, "Playerank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 5, Sep. 2019, ISSN: 2157-6904. DOI: 10.1145/3343172. [Online]. Available: https://doi.org/10.1145/3343172.

[8] P. Cintia, F. Giannotti, L. Pappalardo, D. Pedreschi, and M. Malvaldi, "The harsh rule of the goals: Data-driven performance indicators for football teams," in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2015, pp. 1–10. DOI: 10.1109/DSAA.2015.7344823.

[9] P. Gould and A. Gatrell, "A structural analysis of a game: The liverpool v manchester united cup final of 1977," *Social Networks*, vol. 2, no. 3, pp. 253–273, 1979. DOI: 10.1016/0378-8733(79)90017-0.

[10] J. Duch, J. S. Waitzman, and L. A. N. Amaral, "Quantifying the performance of individual players in a team activity," *PLoS ONE*, vol. 5, no. 6, 2010. DOI: 10.1371/journal.pone.0010937.

[11] F. M. Clemente, F. M. L. Martins, D. Kalamaras, P. D. Wong, and R. S. Mendes, "General network analysis of national soccer teams in fifa world cup 2014," *International Journal of Performance Analysis in Sport*, vol. 15, no. 1, pp. 80–96, 2015. DOI: 10.1080/24748668.2015.11868778.

[12] J. Buldú, J. Busquets, I. Echegoyen, and F. Seirul Lo, "Defining a historic football team: Using network science to analyze guardiola's f.c. barcelona," *Scientific reports*, vol. 9, no. 1, p. 13 602, Sep. 2019, ISSN: 2045-2322. DOI: 10.1038/s41598-019-49969-2. [Online]. Available: https://europepmc.org/articles/PMC6753100.

[13] S. A. Pettersen, D. Johansen, H. Johansen, V. Berg-Johansen, V. R. Gaddam, A. Mortensen, R. Langseth, C. Griwodz, H. K. Stensland, and P. Halvorsen, "Soccer video and player position dataset," in *Proceedings of the 5th ACM Multimedia Systems Conference*, ser. MMSys '14, Singapore, Singapore: Association for Computing Machinery, 2014, pp. 18–23, ISBN: 9781450327053. DOI: 10.1145/2557642.2563677. [Online]. Available: https://doi.org/10.1145/2557642.2563677.

[14] T. Stolen, K. Chamari, C. Castagna, and U. Wisl??ff, "Physiology of soccer," *Sports Medicine*, vol. 35, no. 6, pp. 501–536, 2005. DOI: 10.2165/00007256-200535060-00004.

[15] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognition*, vol. 36, no. 2, pp. 451–461, 2003, Biometrics, ISSN: 0031-3203. DOI: https://doi.org/10.1016/S0031-3203(02)00060-2. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320302000602.

[16] J. Joe H. Ward and M. E. Hook, "Application of an hierarchical grouping procedure to a problem of grouping profiles," *Educational and Psychological Measurement*, vol. 23, no. 1, pp. 69–81, 1963. DOI: 10.1177/001316446302300107. eprint: https://doi.org/10.1177/001316446302300107. [Online]. Available: https://doi.org/10.1177/001316446302300107.

[17] A. W. Moore, "Very fast em-based mixture model clustering using multiresolution kd-trees," *Advances in Neural information processing systems*, pp. 543–549, 1999.

[18] D. Bertsimas and R. Shioda, "Classification and regression via integer optimization," *Operations Research*, vol. 55, no. 2, pp. 252–271, 2007.