Clustering match-level statistics to determine playing styles in soccer

João Gabriel Gelli[1]
joao.gelli@tecgraf.puc-rio.br

Fernanda Araujo Baião[1]
fbaiao@puc-rio.br

[1]Pontifícia Universidade Católica (PUC-Rio)
R. Marquês de São Vicente, 225 - Gávea, Rio de Janeiro - RJ, 22451-900

ABSTRACT

Soccer is one of the most popular sports in the world and is increasingly welcoming data analysis and quantitative approaches to search for advantages in the soccer pitch. In this direction, the analysis of different playing styles is promising. Accordingly, this paper proposes a non-traditional application of clustering techniques on data from soccer matches in worldwide championships. It uses the data to describe how each team behaves in terms of defense, progressions, laterality, and efficiency in different matches. The approach revealed distinct and recurring playing styles which helped to understand how the teams match with one another, and their chances of obtaining positive results depending on the combination of their playing styles. Our experimental results pointed to the importance of the efficiency in the fundamentals of the sport and show that this approach can bring benefits in terms of team strategy.

KEYWORDS. Soccer, Data Analytics, Clustering

Paper topics: Big Data and Analytics

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

1. Introduction

Soccer presents itself as the sport that attracts the most viewers, with around 43% of the world population having a strong interest in it, which beats basketball by 7% [Nielsen Sports 2018]. Researches also show that the number of athletes had already surpassed a hundred million over fifteen years ago [Dvorak et al. 2004].

When it comes to market value, a valuation of the five biggest Soccer leagues in Europe showed that they were worth more than 25 billion euros [Deloitte 2018]. With such high amounts, it is necessary that clubs employ an efficient management system, with the help of the right tools in search of financial and sporting results [Kelly 2017].

The analysis of the sport itself has been mostly qualitative, focused in the experience of scouts and coaches. However, with the financial impact brought by a win or a loss, teams have recently started to look for further advantages, particularly from historical data collected over the years. Therefore, a more quantitative side of analysis is spreading in soccer [Arriaza and Zuniga 2016].

The application of Data Analytics approaches on sports data, and specifically on soccer, has only been made possible after a technological breakthrough, with data collection taking place based on videos [Hernandez-Mendo et al. 2002] and tracking devices [Coutts and Dufeld 2008]. This way, data analysis in soccer has evolved and focused on performance by means of a tactical and technical lens [Hughes and Bartlett 2002], in addition to its emotional and intuitive counterparts. In fact, this is considered one of the major factors in helping teams with less economic power to compete against bigger opponents [Alamar 2013]. From a technical perspective, the combination of a huge volume of data collected on a significant number of matches with the multiple alternative team playing styles and strategies that may be followed during a soccer match makes for a promising field to apply mining techniques [Carpita et al. 2014].

One of the ways to extract knowledge from this data is to study playing styles, which may refer to an analysis of how the team operates on the field, how its opponent plays or how both impact each other and the result. Literature presents some applications of clustering techniques applied to soccer player data for multiple purposes, described as follows.

To analyze players, [Ven 2018] categorizes them based on their ratings in the FIFA videogame. The author combined k-means with SPCA algorithms to find the best combinations of kinds of players and their contribution to the team results. These combinations were then used to predict results.

In terms of team performance, [Wibowo 2016] used data from the champions of the five largest European leagues between 1993 and 2014 to define a new performance indicator named Situational Score Line. This indicator considers the team results, its performance as an away team and the quality of opposing team. The author applied k-means algorithm to group teams according to this indicator.

[Moura et al. 2014] used data from the 2006 World Cup to perform a PCA and a cluster analysis to identify aspects that were mutual to winning teams, to losing teams and to tying teams. They applied k-means and assessed the clusters using the Silhouette Index.

Finally, [Yi et al. 2019] split the teams from the 2014 World Cup between those that opted for a possession-based style and those which were more direct in their actions. Teams were described by features related to the amount of shots after counter attacks and ball possession. Then, the clusters were found using k-means, and subject to a Poisson regression to compare the performance of the teams from each style and how they matched with each other, based on a series of match statistics that represented both the technical and the physical sides.

The present paper aims to advance the discussions in this growing field of Data Analytics and Machine Learning techniques applied to soccer data by grouping teams according to their playing styles, which in turn are defined based on their match-level statistics. We report the results of an experiment applying k-means cluster analysis on top of feature groups describing different aspects of each team on a specific match (defense, efficiency in fundamentals, ability in progressions and the side that the team prefers to attack). These aspects were chosen due to their capability to portray multiple

playing styles. Finally, the specified playing styles are subject to a matchup analysis, which revealed that a higher efficiency in fundamentals leads to a better percentage in points won.

2. Methodology

2.1. Dataset and Preprocessing

Data was obtained through the API from the InStat Sports Performance Analysis and Scouting platform [InStat 2019]. Originally, the dataset had 616,305 registers and 133 features, where each row represented a team in a match, while the features represented match-level statistics for each team. More specifically, the dataset features comprised metadata from the match (identifications, home and away teams, date, result, goals scored and goals against), from the teams involved in each match (names in different languages and identifications) and from the tournament (identifications, country, round, season), as well as a set of statistics, each describing an action performed on the field by one of the teams during a match. These statistics included counting measures (such as the number of left side attacks) or performance indicators (such as team efficiency in left side attacks). After all the treatments described in this section, these features were presented to the K-Means algorithm and represented the information of interest available from the domain.

Table 1 illustrates a short view of the dataset, describing one match (ID = 1) that occurred in one of the rounds (ID = 6) from a particular tournament (ID = 3), between two teams (ID = A and ID = B). Each row describes the same match from the perspective of one of the playing teams. The 5 first columns represent the identifications, followed by the number of goals scored and against by each team (A scored 1 goal, while B scored 2 goals in this match), and finally followed by a set of N fields representing the match statistics, which are further detailed.

Table 1. Basic shape of the database

Match ID	Team ID	Opponent ID	Tournament ID	Round ID	Goals Scored	Goals Against	Stat 1	Stat N
1	A	2	3	6	1	2	0,75	6
1	B	1	3	6	2	1	0,47	12

The original dataset was first cleaned to replace missing values and to remove registers that were not coherent (negative values in features that should only accept positive ones, such as ball possession or total actions in the game). The features were then normalized using the z-scores to remove outliers. Every register that had any feature with a z-score greater than 3 or lesser than -3 was considered an outlier and removed. After that, a final treatment was to keep only matches that still had both teams in the database. These data cleaning and transformation actions were performed using the R language in RStudio, through the package “tidyverse” [Wickham et al. 2019].

The dataset was reduced so it only had specific tournaments from Europe and Americas, because they are the most representative continents in soccer. Therefore, the final dataset only contained data from matches played in the national championships from Spain, Italy, Netherlands, Portugal, Germany, France, England, the USA, Brazil, Argentina, Uruguay, Mexico, Venezuela, Chile, Colombia, Ecuador, Paraguay and Peru.

As the amount of real chances of scoring were only registered starting from 2017, all rows from a match before that were removed. Then, with the available data and the help from an expert in the domain, nine new features were created. They reflect the teams’ abilities in terms of defense (pressure, quality), laterality (which is the preferred side of attacks and its efficiency), progressions (how much quality the teams has in advancing on the field) and efficiency in fundamentals (shooting, passing, dribbling) and are represented in Table 2.

The next step was to select the features of interest for the analysis. These features were then grouped into different abilities: defense, laterality, progressions, and efficiency in fundamentals. The teams, competitions, rounds, and results were kept in the database, but were not used in the clustering.

Table 3 presents the chosen features and their respective groupings. At the end of all the preprocessing stage, the dataset had 22822 rows and 31 features.

Table 2. Created features description

Feature	Formula	Meaning
PPDA	(Opponent passes) / (Defensive challenges + Tackles)	Number of passes that the opponent makes per defensive action
Recovery Height	(Ball recoveries in the attacking field) / (Ball recoveries)	A higher number represents that ball recoveries usually were closer to the opponent side of the field.
Attack Speed	(Counter attacks) / (Positional attacks)	Represents the tendency to be more direct or cautious in the attacks.
Progressions to Opponent Side	(Entrances in the attacking field) / (Ball possessions)	Indicates the efficiency in advancing to the attacking field.
Progressions to the Final Third	(Entrances in the final third) / (Entrances in the attacking field)	Indicates the efficiency in advancing to the final third after getting to the attacking field.
Progressions to the Box	(Entrances in the box) / (Entrances in the final third)	Indicates the efficiency in advancing to the box after getting to the final third.
xG per Real Chance	(Expected goals) / (Real chances)	How much each real chance of scoring represents in expected goals.
Real Chances per Time of Possession	(Real chances) / (Time of possession)	How a team converts its control in real chances.
Real Chances per Possession	(Real chances) / (Amount of possessions)	Represents the effectiveness in converting possessions in real chances.

Table 3. Selected features and their respective groupings

Feature	Group
PPDA	Defense
Efficiency in Defensive Challenges	
Opponent Real Chances	
Recovery Height	
Amount of Pressures	
Pressing Efficiency	
Attacks Left	Laterality
Attacks Center	
Attacks Right	
Attacks Left - Efficiency	
Attacks Center - Efficiency	
Attacks Right - Efficiency	
Average Time of Possession	Progressions
Attack Speed	
Progressions to Opponent Side	
Progressions to the Final Third	
Progressions to the Box	
Passing Accuracy	Efficiency
Shooting Accuracy	
Dribbling Accuracy	

Key Passes Accuracy
xG per Real Chance
Real Chances per Time of Possession
Real Chances per Possession

2.2. K-Means Clustering

K-Means was applied to determine the clusters for each group of features. It aims to split a sample of points into k groups to reduce the variance inside each cluster [MacQueen 1967].

The k-means algorithm starts by randomly designating k points as centroids of the initial clusters. After each iteration, the points in each cluster may change to another cluster, to minimize the sum of the squares of the distances to the centroids of each cluster. Then, the centroids are reset, and a new sweep begins until there is no gain possible [Hartigan and Wong 1979].

The analysis of the resulting clusters can be made through a series of indicators. Among the most recommended are the Silhouette Index and the Davies-Bouldin Index [Arbelaitz et al. 2013]. This paper also used the Negative Silhouette Ratio as a validation indicator. The Silhouette Index takes the lowest intercluster distance difference for a certain point and subtracts the average intracluster distance difference for that same point. Then, divides by the highest between these intercluster and intracluster values. It results in a value between -1 and 1 and the Silhouette Index is the average of the Silhouette values from every cluster. It can be interpreted as the more positive the value, the better the point fits in the designated cluster [Rousseeuw 1987]. The Negative Silhouette Ratio is the proportion of points with negative values of Silhouette in the clustering configuration.

The Davies-Bouldin (DB) Index measures how compact the clusters are and is calculated as follows:

$$DB(C) = \frac{1}{K}\sum_{i=1}^{K} \max_{i \neq j}\{\frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i\,,\, C_j)}\}$$

In this formula, $\Delta(C_i)$ represents the intracluster difference and $\delta(C_i\,,\, C_j)$ is the intercluster distance between clusters i and j. A lower value in the DB Index represents a better clustering [Davies and Bouldin 1979].

3. Determining styles of play

Once the data was preprocessed, the cluster analysis was performed to designate different groupings related to defense, progressions, efficiency, and laterality. The idea was to determine a cluster for each of these facets of the game, to describe how a team played in a game regarding each facet. Then, they could be further combined into a complete description of the playing style of team in that match.

The clustering algorithm chosen was k-means, varying k from 2 to the number of features that the group has. Every run had 100 random starts. The clusters found were validated using the Silhouette Index, the Negative Silhouette Ratio, and the Davies-Bouldin Index. This was executed using the R language in RStudio through the packages “cluster” [Maechler et al. 2018], “clusterSim” [Walesiak and Dudek 2019] and the function “kmeans”. Table 4 has a summary of the experiment plan.

Table 4. Experiment plan

Group of Features	Values of K	Random Starts	Evaluation Metrics
Defense	2 to 6	100	Silhouette Index, Negative Silhouette Ratio and Davies-Bouldin Index
Progressions	2 to 5		
Laterality	2 to 6		
Efficiency	2 to 7		

The following tables present the evaluation metrics for each value of k tested for each group of features. The Silhouette Index, the Davies-Bouldin Index and the Negative Silhouette Ratio were the chosen metrics.

3.1. Defense

Table 5 presents the results of the evaluation metrics for the k-means performed on the defensive features.

Through Table 5, it can be noted that the Silhouette Index with two clusters is 40% better than the second-best value in this indicator (0.248 as opposed to 0.176). In addition, the scenario with k=2 is the one with the smallest proportion of negative Silhouette. Furthermore, its DB Index is only about 3% worse than the best one. Therefore, the best grouping comes when k is two.

Table 5. Evaluation metrics for the defense features

K	Silhouette Index	DB Index	Negative Silhouette Ratio
2	0,248	1,720	0,016
3	0,175	1,781	0,067
4	0,155	1,738	0,112
5	0,176	1,739	0,090
6	0,172	1,673	0,088

With two clusters as the best setting for the defensive features, Table 6 brings the median, first quartile and third quartile for each feature, and the number of registers in each cluster. Each cluster is named according to the interpretation of the results.

Table 6. Features' statistics in the defense clusters. Median [First quartile - Third quartile]

Feature	Cluster 1 ("No Pressing")	Cluster 2 ("Pressing")
Registers	17452	5370
Efficiency Defensive Challenges	0.56 [0.51 – 0.6]	0.56 [0.51 – 0.6]
Opponent Real Chances	5 [3 - 7]	5 [3 - 7]
PPDA	3.59 [2.98 – 4.4]	3.91 [3.22 – 4.75]
Pressing Efficiency	0 [0 - 0]	0.5 [0.33 - 1]
Amount of Pressures	0 [0 - 1]	2 [1 - 4]
Recovery Height	0.15 [0.11 – 0.2]	0.18 [0.14 – 0.24]

A quality interpretation of the results reveals that the main difference between clusters 1 and 2 are relative to pressuring. Teams in cluster 1 do not pressure often and are not highly effective when they do ("No Pressing"), whereas the ones in cluster 2 are effective and do it at a higher rate ("Pressing"). However, it is more common that teams do not use pressure as part of their styles.

3.2. Progressions

Table 7 presents the results of the evaluation metrics for the k-means performed on the progression features.

Table 7. Evaluation metrics for the progression features

K	Silhouette Index	DB Index	Negative Silhouette Ratio
2	0,222	1,714	0,019
3	0,191	1,740	0,020
4	0,177	1,656	0,022
5	0,161	1,653	0,045

Based on the values from Table 7, the definition of the number of clusters for the progression variables has a quick exclusion, as the case with k equal to five is worse than the others in terms of

proportion of negative Silhouettes. Between the other three options, the one with two clusters brings the better combination between a higher Silhouette Index and a lower DB Index.

With two clusters as the best setting for the progression features, Table 8 brings the median, first quartile and third quartile for each feature, and the number of registers in each cluster. Each cluster is named according to the interpretation of the results.

Table 8. Features' statistics in the progression clusters.
Median [First quartile - Third quartile]

Feature	Cluster 1 ("Counter Attacks")	Cluster 2 ("Positional Attacks")
Registers	12568	10254
Attack Speed	0.29 [0.23 – 0.36]	0.22 [0.17 – 0.27]
Average Time of Possession	12.09 [10.83 – 13.42]	15.93 [14.38 – 17.82]
Progressions to Opponent Side	0.33 [0.27 – 0.39]	0.37 [0.31 – 0.43]
Progressions to the Final Third	0.59 [0.53 – 0.64]	0.67 [0.62 – 0.72]
Progressions to the Box	0.53 [0.48 – 0.57]	0.63 [0.59 – 0.68]

A quality interpretation of the results reveals that cluster 1 presents itself as the one that includes teams that attack in a faster and direct way, with a tendency to prefer counter attacks and tougher passes, which explains the lower values in the metrics in general ("Counter Attacks"). Meanwhile, cluster 2 represents teams that are more cautious with the ball and opt to control the possession, progressing slowly across the field ("Positional Attacks").

3.3. Laterality

Table 9 presents the results of the evaluation metrics for the k-means performed on the laterality features.

Table 9. Evaluation metrics for the laterality features

K	Silhouette Index	DB Index	Negative Silhouette Ratio
2	0,131	2,522	0,029
3	0,123	2,183	0,069
4	0,122	1,980	0,056
5	0,124	1,854	0,060
6	0,128	1,755	0,047

In the laterality features showed in Table 9, the Silhouette Index are distributed in intervals of less than 0,01, which does not bring conclusive results. Nevertheless, the setting when k is 6 has the second-best value in this indicator and the best in the DB Index, making it the choice.

With six clusters as the best setting for the laterality features, Table 10 brings the median, first quartile and third quartile for each feature, and the number of registers in each cluster. Each cluster is named according to the interpretation of the results.

Table 10. Features' statistics in the laterality clusters. Median [First quartile - Third quartile]

Feature	Cluster 1 ("Center")	Cluster 2 ("Right")	Cluster 3 ("Left Efficiency")	Cluster 4 ("Right Efficiency")	Cluster 5 ("Left")	Cluster 6 ("Center Efficiency")
Registers	4867	4272	3065	3186	4249	3183
Attacks Center – Efficiency	0.12 [0.07 – 0.16]	0.09 [0.05 – 0.13]	0.12 [0.07 – 0.17]	0.12 [0.07 – 0.16]	0.09 [0.05 – 0.14]	0.24 [0.21 – 0.28]
Attacks Center	29 [26 - 33]	20 [17 - 23]	24 [20 - 27]	22 [19 - 26]	20 [17 - 24]	21 [17 - 24]

Attacks Left – Efficiency	0.05 [0.03 - 0.08]	0.06 [0.04 - 0.1]	0.17 [0.15 - 0.2]	0.09 [0.06 - 0.12]	0.07 [0.04 - 0.1]	0.09 [0.06 - 0.12]
Attacks Left	25 [21 - 29]	27 [23 - 31]	26 [22 - 30]	29 [25 - 33]	36 [32 - 40]	31 [27 - 35]
Attacks Right – Efficiency	0.06 [0.03 - 0.09]	0.07 [0.05 - 0.1]	0.08 [0.05 - 0.11]	0.17 [0.15 - 0.2]	0.05 [0.03 - 0.09]	0.09 [0.06 - 0.12]
Attacks Right	27 [23 - 31]	38 [35 - 42]	30 [25 - 34]	28 [24 - 32]	28 [24 - 31]	32 [28 - 37]

A quality interpretation of the results reveals that each cluster represents a team that stands out in one of the metrics. Cluster 1 encompasses teams that attack through the middle ("Center"), two represents the ones that prefer advancing their possessions on the right side ("Right") and five groups the ones that opt to build their offenses mostly through the left side ("Left"). Meanwhile, cluster 3 brings teams that efficient through the left ("Left Efficiency"), four the ones that are efficient through the right ("Right Efficiency") and six shows the ones that are efficient through the middle ("Center Efficiency").

3.4. Efficiency in fundamentals

Table 11 presents the results of the evaluation metrics for the k-means performed on the efficiency features.

Table 11. Evaluation metrics for the efficiency features

K	Silhouette Index	DB Index	Negative Silhouette Ratio
2	0,186	1,947	0,033
3	0,145	2,158	0,036
4	0,140	1,959	0,060
5	0,132	1,882	0,071
6	0,133	1,762	0,080
7	0,136	1,728	0,072

Among the efficiency features presented in Table 11, the grouping with two clusters was the best. This happens because when k is two or three it makes for the best proportions of negative Silhouette. Between these options, the first one has a better value in both the Silhouette Index and DB Index.

With two clusters as the best setting for the efficiency features, Table 12 brings the median, first quartile and third quartile for each feature, and the number of registers in each cluster. Each cluster is named according to the interpretation of the results.

Table 12. Features' statistics in the efficiency clusters.
Median [First quartile - Third quartile]

Feature	Cluster 1 ("Less Efficient in Fundamentals")	Cluster 2 ("More Efficient in Fundamentals")
Registers	12168	10654
Real Chances per Possession	0.03 [0.02 - 0.04]	0.06 [0.05 - 0.08]
Real Chances per Time of Possession	0 [0 - 0]	0 [0 - 0.01]
Dribbling Accuracy	0.56 [0.48 - 0.63]	0.58 [0.5 - 0.65]
Key Passes Accuracy	0.4 [0.22 - 0.53]	0.56 [0.45 - 0.67]
Passing Accuracy	0.81 [0.77 - 0.84]	0.82 [0.78 - 0.85]
Shooting Accuracy	0.31 [0.22 - 0.42]	0.43 [0.33 - 0.53]

xG per Real Chance	0.36 [0.24 - 0.56]	0.27 [0.2 - 0.37]

A quality interpretation of the results reveals that teams in cluster 1 tend to be less efficient, in general, but their chances have a greater expected goal value ("Less Efficient in Fundamentals"). Whilst, squads in cluster 2 are more efficient in terms of generating chances in their possessions, with a greater precision in key passes and the shots that come from them ("More Efficient in Fundamentals").

4. Playing styles matchups

To compare how these styles affect the results when they face each other, the following heat maps show every matchup between each group of features. Figure 1 shows the percentage of points accrued by a team of style X when it faces ones from style Y inside the same group of features. The percentage of points is the proportion between points scored (3 for a win and 1 for a tie) and total points available. A black tile represents that the team from style X gets all the points in the matchup against team from style Y and a light grey one represents that it gets no points.

Figure 1. Point winning percentage in playing styles matchups inside the same group of features. A darker value represents a better percentage.

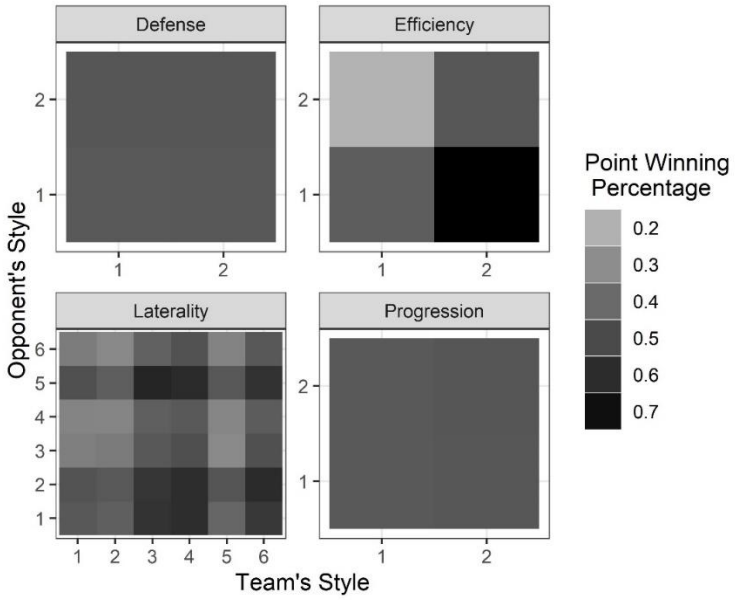


Figure 1 illustrates that defensive and progression styles do not have much impact in winning percentage, that is, when a pressing team faces against a non-pressing one, this does not affect the result. However, in terms of laterality, teams that play in styles 3, 4 and 6 (i.e., efficient teams) tend to win more. Also, there is a clear advantage for teams that are more precise in fundamentals when compared to the ones that are less effective.

Beyond this analysis, it is possible to unify the playing styles across the group of features and determine a complete style. We represent each complete style as a 4-digit code "ABCD", in which: A (digit 1) represents the defensive cluster, B (digit 2) represents the progressions, C (digit 3) represents the efficiency, and D (digit 4) represents laterality.

By using this code, for example, a team with a complete style "1213" is one that: (i) doesn't tend to pressure, (ii) uses positional attacks, (iii) is less efficient in fundamentals and (iv) has a higher efficiency building their offenses through the left. This way, there are 48 (2 * 2 * 6 * 2) possible complete styles, based on the number of clusters previously assigned to each group of features. Table

13 shows the 10 complete playing styles with the best point winning percentage, along with the total matches played and its ranking among all the styles.

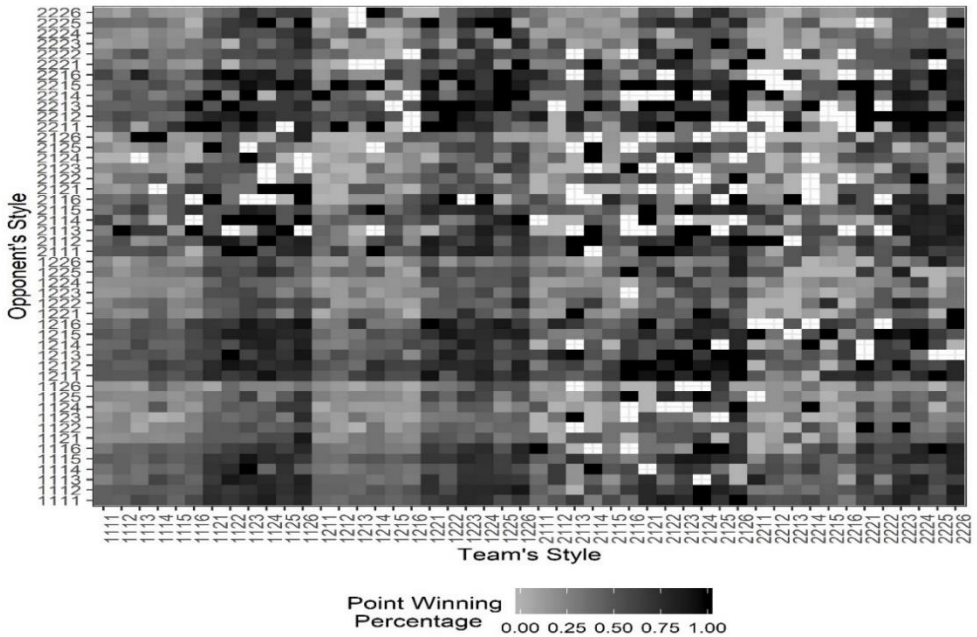
Table 13 shows that the style in which the team doesn't pressure (1), prefers counter attacks (1), is efficient in fundamentals (2) and attacks through the middle with efficiency (6) is the best in terms of point winning percentage. Only four styles ("1123", "1224", "1223" and "1226") are among both the ten best in point winning percentage and the most frequent, which represents that they are efficient and highly used.

Table 13. Ten best complete styles in terms of point winning percentage

Complete Style	Point Winning Percentage	Matches Played (Ranking)
1126	69.18%	465 (21)
1123	67.26%	617 (10)
2124	65.77%	185 (37)
2126	65.63%	161 (43)
1124	65.58%	611 (12)
1224	63.32%	798 (7)
1223	61.71%	700 (9)
2223	61.68%	321 (29)
1226	61.56%	980 (4)
2224	61.54%	351 (28)

As in Figure 1, Figure 2 presents a heat map of point winning percentage among style matchups, but this time focused in the complete styles. White matchups are the ones in which a game has not been played. From Figure 2, we notice that a better point winning percentage comes when the team is among the ones with a higher efficiency in fundamentals.

Figure 2. Point winning percentage for the complete style matchups. A darker value represents a better percentage.



5. Conclusion

Soccer presents itself as the biggest sport in the world and is increasingly accepting data science to get advantages in the game. Clustering techniques have been used in different ways regarding players and teams and could also be applied to group teams according to their performances in different characteristics of the game. This would be useful to better understand the team and their opponents.

This paper created clusters of teams based on their playing styles in multiple characteristics. First, the teams were divided regarding how they approached the game in terms of defense, progressions, laterality, and efficiency in fundamentals. Then, with a defined playing style for each of these facets, a complete playing style was determined, which characterized the team according to multiple facets, explaining how it defended, which side it preferred to build their attacks, how efficient they were and how they advanced through the field.

The experimental setup comprised data from 18 tournaments from Europe and the Americas. We applied k-means to determine clusters of teams in matches based on statistics from defense, progressions, laterality, and efficiency. This resulted in multiple playing styles that were analyzed in terms of their impact to winning and advancing the knowledge in understanding style matchups.

The interpretation of the results revealed interesting insights, such as that efficiency in fundamentals is a high priority aspect of the game, since it brought better point winning percentages. Thus, focusing in training the quality of passing, shooting and the ability to create shots is paramount. Also, this study can be used to identify a specific style for an opponent and help to prepare a game plan to better react to it.

Future investigations can be such as checking how a team's style varies over time what is the impact in its performance. Another possible study is to analyze the differences in styles adopted in Europe and in the Americas.

References

Alamar, B. C. (2013). Sports Analytics: A Guide for Coaches, Managers, and Other Decision Makers. New York, Columbia University Press.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., Perona, I. (2013). An extensive comparative study of cluster validity indices. Pattern Recognition, 46(1):243--256.

Arriaza, E. J., Zuniga, M. D. (2016). Soccer as a Study Case for Analytic Trends in Collective Sports Training: A Survey. International Journal of Performance Analysis in Sport, v. 16, n. 1, pp. 171--190. ISI.

Carpita, M., Sandri, M., Simonetto, A., Zuccolotto, P. (2014). Football Mining with R. Data Mining Applications with R, 397–433. doi:10.1016/b978-0-12-411511-8.00015-3

Coutts, A.J., Dufeld, R. (2008). Comparison of global positioning and computer-based tracking systems for measuring player movement distance during Australian football. Journal of Sciences and Medicine in Sport 13(1): 133–135.

Davies, D. L., Bouldin, D. W. (1979). A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, (2):224--227.

Deloitte. (2018). Rising stars Football Money League. https://www2.deloitte.com/uk/en/pages/sports-businessgroup/articles/annual-review-of-football-finance.html. Last accessed on 2019-12-10

Dvorak, J., Junge, A., Graf-Baumann, T., Peterson, L. (2004). Editorial. The American Journal of Sports Medicine, 32(1_suppl), 3–4. doi:10.1177/0363546503262283

FIFA. (2018) Global Transfer Market Report 2018 - A review of all international football transfers in 2017. https://www.fifatms.com/wpcontent/uploads/dlmuploads/2018/01/GTM 2018.pdf. Last accessed on 2019-12-11

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1):100--108.

Hernández-Mendo, A., Anguera, M.T., Bermúdez-Rivera, M.A. (2000). Software for recording observational files. Behavior Research Methods, Instruments and Computers 32(3): 436–445.
Hughes, M., Bartlett, R. (2002). The use of performance indicators in performance analysis. Journal of Sports Sciences, v. 20, n. 10, pp. 739—754
InStat (2019). InStat API. Retrieved from https://instat.tv/
Kelly, S. (2017). The role of the professional football manager. https://books.google.co.uk/books?id=LCAlDwAAQBAJ. Last accessed on 2019-12-10
Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2018). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.7-1.
Moura, F.A., Martins, L.E.B., Cunha, S.A. (2014) Analysis of football game-related statistics using multivariate techniques, Journal of Sports Sciences, 32:20, 1881-1887, DOI: 10.1080/02640414.2013.853130
Nielsen Sports. (2018). World Football Report. https://nielsensports.com/reports/world-football-2018/. Last accessed on 2019-10-28
Rowe, D. (2011). Global Media Sport: Flows, Forms and Futures. London, Bloomsbury.
Rousseeuw, P.J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics. 20: 53–65. doi:10.1016/0377-0427(87)90125-7.
Ven, E.V. (2018). Clustering soccer players to find the drivers of soccer team performance. Master Thesis. Erasmus University Rotterdam. Retrieved from http://hdl.handle.net/2105/43857.
Walesiak, M., Dudek, A. (2019). clusterSim: Searching for Optimal Clustering Procedure for a Data Set. R package version 0.48-1.
Wibowo, C. (2016). Clustering seasonal performances of soccer teams based on situational score line. Communications in Science and Technology. 1. 10.21924/cst.1.1.2016.11.
Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
Yi, Q., Gómez, M. A., Wang, L., Huang, G., Zhang, H., Liu, H. (2019). Technical and physical match performance of teams in the 2018 FIFA World Cup: Effects of two different playing styles. Journal of Sports Sciences. doi:10.1080/02640414.2019.164812