mathematics

Article

Bisecting for Selecting: Using a Laplacian Eigenmaps Clustering Approach to Create the New European Football Super League

Alexander John Bond * and Clive B. Beggs

Carnegie School of Sport, Leeds Beckett University, Leeds LS6 3QQ, UK
* Correspondence: a.j.bond@leedsbeckett.ac.uk

Abstract: Ranking sports teams generally relies on supervised techniques, requiring either prior knowledge or arbitrary metrics. In this paper, we offer a purely unsupervised technique. We apply this to operational decision-making, specifically, the controversial European Super League for association football, demonstrating how this approach can select dominant teams to form the new league. We first use random forest regression to select important variables predicting goal difference, which we use to calculate the Euclidian distances between teams. Creating a Laplacian eigenmap, we bisect the Fiedler vector to identify the natural clusters in five major European football leagues. Our results show how an unsupervised approach could identify four clusters based on five basic performance metrics: shots, shots on target, shots conceded, possession, and pass success. The top two clusters identify teams that dominate their respective leagues and are the best candidates to create the most competitive elite super league.

Keywords: football; soccer; European Super League; spectral clustering; Laplacian eigenmap; Fiedler vector

MSC: 62F07; 62H30; 90B50

Citation: Bond, A.J.; Beggs, C.B. Bisecting for Selecting: Using a Laplacian Eigenmaps Clustering Approach to Create the New European Football Super League. *Mathematics* **2023**, *11*, 720. https://doi.org/10.3390/math11030720

Academic Editor: David Delgado-Gómez

Received: 7 December 2022
Revised: 19 January 2023
Accepted: 28 January 2023
Published: 31 January 2023

1. Introduction

In recent years there has been growing interest in establishing an elite European Super League (ESL) that would allow the top soccer clubs in Europe to compete in a traditional league similar to the EuroLeague established in basketball [1]. This idea came to fruition in 2021 with plans to disrupt European football operations and introduce a new elite ESL. However, due to a backlash in public opinion, the ESL was short-lived and ceased operations after clubs withdrew amid fines and sanctions [2]. Notwithstanding this, although currently paused, the commercial pressures and incentives that spawned the ESL have not disappeared, and the project may be resurrected in the future. As such, it raises the critical question of how teams might be objectively selected to participate in an elite European football league. European football generates annual revenues of approximately $28 billion per annum [3], with 60% stemming from the global demand for the five major leagues; English Premier League, Spanish La Liga, German Bundesliga, Italian Serie A, and French Ligue 1. This revenue-generating capacity is based on leagues ensuring teams are similar in resource, infrastructure, and performance [4–6]. Thus, selecting the right teams to compete in an elite football league is an important issue with substantial financial implications.

The challenge of selecting teams for inclusion in the ESL is representative of a much broader research question in sport: how can athletes, players, and teams from different contexts be objectively compared when the competition is fragmented so that participants rarely compete against each other? An important research question that has received much attention in recent years with the development of various ranking methodologies (e.g., the Elo [7,8], Colley [7,9,10], Massey [7,10,11], Keener [7,10,12], and PageRank [10,13,14] systems) developed to rate the relative strength of competitors in fragmented competition or cup tournaments [7,10]. With specific reference to European soccer, numerous techniques have been developed to [15–17] rank teams, such as the Euro Club Index [15], the ClubElo Index [16], and the UEFA club coefficient rankings [17]. These ranking systems are generally either adaptations of the Elo system utilising a who-beat-who methodology [7,8] or involve allocating points for matches in European cup competitions. For example, the UEFA club coefficient system awards 2 points for a win and 1 point for a draw, together with a complex system of bonus points, arbitrarily awarded depending on the stage of the competition and the perceived difficulty of the cup tournament [18]. Although useful, these ranking systems are limited as they either rely on: (i) the subjective allocation of points; or (ii) the competing teams playing each other on a regular basis—something that currently does not occur in the various European cup tournaments. For example, Arsenal only played Real Madrid once in 2006, while Tottenham Hotspur has never played Paris Saint-Germain.

Consequently, when teams play each other very infrequently, methodologies such as the Elo system that rank teams according to who-beat-who tend to become inaccurate because player line-ups can change significantly from season to season. Another problem with ranking systems is that they tend to use a single metric, such as who-beat-who, or the points awarded for match outcomes to construct a ranking table. Consequently, they cannot cope with multiple performance metrics, such as those recorded during match play. Furthermore, they reveal nothing about similarities in on-the-pitch performance between teams or natural groupings (clusters) that may exist within European football.

Given that successful soccer teams often display similar match performance characteristics (e.g., greater possession, more shots on target, etc.) [19–22], there is reason to believe that irrespective of the domestic league from which they come, the top teams in Europe might naturally aggregate into an elite cluster. Accordingly, we hypothesise that the top European teams would tend to cluster according to their match performance characteristics. This cluster could be used to select teams suitable for inclusion in the ESL. To this end, we developed a novel unsupervised data-driven approach that blended machine learning and graph theory to identify natural clusters of teams in Europe's five major leagues using aggregated performance data. In so doing, we aimed to demonstrate that it is possible to objectively identify the top teams in Europe without the need for supervised learning or any subjective assessment criteria.

One of the significant challenges when attempting to identify the natural clusters was how best to accommodate the multiple performance metrics. We propose a higher dimensional approach by computing the Euclidean distances between the respective soccer teams in the vector space, creating a similarity matrix that could create a network graph capturing the multi-dimensional relationships in the data [23].

While this approach enabled the closeness of the respective teams to be visualised in a network graph, it still left the problem of identifying distinct clusters within the data. Although numerous machine learning techniques exist for identifying clusters in data or classifying data according to predetermined categories, many of these necessitate a priori assumptions, such as stating the number of clusters in the data or specifying formal categories within the data. Consequently, even when unsupervised, these techniques tend to rely on subjective decisions that could compromise their objectivity. Consequently, we constructed a Laplacian eigenmap from the similarity matrix [24,25], which allowed repeated spectral graph partitioning using the Fiedler vector [26–28]. Our approach showed that it is possible to identify the natural clusters in the data without the need for any a priori assumptions—something that has never been attempted in a footballing context. By taking this approach, we developed an objective methodology that identified the top football teams in Europe purely from their match performance characteristics in the respective domestic leagues. As such, this study is the first to report the use of spectral partitioning to group football teams into natural clusters.

2. Materials and Methods

2.1. Data Acquisition

Using publicly available football performance data from footystats.com [29] and WhoScored.com [30], we used season performance data for all the teams in the Bundesliga, La Liga, Ligue 1, English Premier League, and Serie A over seven seasons between 2013/14–2019/2020. This produced a study data set comprising 686 observations from 150 football teams. The variables collected are related to an individual team's performance over the entire season. The teams are listed in Appendix A, and the variables included in the study are listed in Table 1. Each team's data from all seven seasons were aggregated into a single dataset (n = 150) to avoid pseudoreplication. This aggregated dataset was then used to perform the data analysis and to compute the descriptive statistics (mean, standard deviation (SD), median, minimum and maximum values) shown in Table 1. These were computed using R (open-source statistical computing software; R Foundation for Statistical Computing, Vienna, Austria).

Table 1. Variable description and descriptive statistics.

Variable	Description	Mean	SD	Median	Min	Max
Yellow_cards	Number of yellow cards received	75.7	17	70.79	43.7	116
Red_cards	Number of red cards received	4.06	1.76	4	0.5	9
Possession	Possession percentage	48.8	4.2	47.81	39.1	64.14
Pass_Success	Successful pass percentage	77.2	4.48	76.9	62.1	89.09
Aerials_Won	Number aerial duals won	18.2	3.79	17.66	9.8	30.65
Shots_Conceeded	Number of shots conceded per game	13.1	1.97	12.87	8.04	18.55
Tackles	Number of tackles made per game	18.2	1.64	18.33	13.3	23
Interceptions	Number of interceptions made per game	14.3	2.19	14.14	9.5	22.3
Fouls	Number of fouls conceded per game	13.3	1.7	13.51	9.34	16.9
Offsides	Number of offsides per game	2.1	0.38	2.07	1.25	3.4
Shots	Number of shots per game	12.2	1.74	11.83	8.8	17.61
Shots_OT	Number of shots on target per game	4.13	0.84	3.95	2.6	7.03
Dribbles	Number of dribbles made per game	9.12	1.73	9.09	4.75	14.1
Fouled	Number of times fouled by opposing team	12.5	1.67	12.74	7.93	17.1
GF	Goals scored	46.3	14	42.58	22	101.9
GA	Goals conceded	54.3	11.7	54.89	24.6	85
GD	Goal difference (GF-GA)	−8.06	23.4	−12.98	−51	70
Points	Total points gained	45.6	15.1	42.36	15	91.29

2.2. Data Analysis Strategy

The study aimed to develop a methodology for identifying natural groupings between teams in the various European soccer leagues, using season match data alone (excluding goals scored or conceded). We performed an exploratory analysis using basic univariate analysis on the variables used in this study before conducting a random forest regression analysis to identify the measured variables that best predicted the goal difference for the respective soccer teams. Goal difference was used because it is a better measure of team performance and less susceptible to bias than points total, which is influenced by the number of teams in the respective leagues [31].

Having identified the variables that best predicted end-of-season goal differences, we computed the Euclidean distances between the respective teams in the vector space. We used them to produce Laplacian eigenmaps of the data [24].

Laplacian eigenmaps are constructed from the eigenvectors of a graph Laplacian matrix. They are essentially an embedding algorithm that seeks to project pairwise proximity information onto a low dimensional space to preserve local structures in the data. Unlike linear dimension reduction techniques such as principal component analysis (PCA), Laplacian eigenmaps have a significant advantage in handling non-linear relationships in the data [24,25]. Therefore, by producing Laplacian eigenmaps, we succinctly visualise the relationships between the respective soccer teams and identify sub-groups within the data using spectral cluster analysis techniques. To benchmark our findings, we classified the respective teams according to their points, using 25% and 75% percentiles to reflect top and bottom-performing teams, otherwise classed as middle teams. The 25% and 75% percentiles turned out to be >56 points classified top teams, <36 points classified bottom teams, with all others classified as middle. All data and statistical analysis were performed using in-house algorithms written in R [32].

2.3. Initial Analysis

An initial univariate analysis of the aggregated data was undertaken using a one-way ANOVA, with post-hoc Bonferroni adjusted pairwise *t*-tests. This allowed a better understanding of the data and variables used in this study.

2.4. Exploratory Random Forest Analysis

An exploratory random forest regression was performed to assess the observed variables' relative importance as predictors of goal difference. Random forest analysis is an ensemble classification technique popular in machine learning that generalises classification trees [33,34]. It is a robust technique resistant to over-fitting and does not require strict distributional assumptions [34,35]. Crucially, it has the advantage of assessing variable importance, thus enabling the removal of redundant variables that do not assist in the prediction process.

Random forest models produce many regression trees that use recursive partitioning to group observations into predefined classes by binary splitting the predictor variables [36]. Bias and over-fitting are minimised by combining bootstrap bagging and utilising a random subset of predictor variables (generally the square root of the total number of predictors in the model) at each split. Each regression tree in the random forest is built using a bootstrapping algorithm, which randomly 'bags' a sample from approximately two-thirds of the data for training purposes. The remaining one-third of the cases or out-of-bag (OOB) cases are used to assess the performance of the regression tree [33,37]. For each tree, the prediction error—mean squared error (MSE) in the case of a regression tree—is computed. These are then pooled to give an overall measure of classification accuracy, thus ensuring that the assessment is unbiased [38].

We used the 'randomForest' package [39] in R [32] to perform a random forest analysis involving creating 500 random trees. Initial analysis was undertaken using all thirteen predictor variables to identify those variables that significantly influenced the outcome variable, Goal_Difference. The 13 predictor variables used to predict goal difference were shots on target, possession, shots, shots conceded, pass_success, dribbles, aerials won, offsides, tackles, yellow cards, red cards, fouls, fouled, and interceptions, as described in Table 1. The relative importance of the variables was assessed using the Gini variable importance measure (VIM), which we corrected for bias using the heuristic strategy proposed by [40,41] and implemented by [42]. For every node split in a tree, the Gini impurity criterion (which assesses the data's heterogeneity) for the two descendent nodes is less than that of the parent node [43]. Therefore, adding up the Gini decreases for each variable over all trees in the forest, it is possible to achieve a measure of variable importance. In our analysis, variables that exceeded the inflexion point's value on the Gini VIM curve were deemed to be influential and thus retained when the random forest model was refined. Having identified the key variables that best predicted goal difference, we then repeated the random forest analysis using the refined model to understand the prediction accuracy that could be achieved. Prediction of the respective teams' goal differences was then performed using the refined model and an ensemble prediction algorithm that aggregated 500 predictions. Because random forests use a self-validating MSE rate, there is no strict need for cross-validation or a separate validation test to obtain an unbiased estimate of model error [38]. However, we performed k-fold cross-validation using ten randomly sampled 'folds' of approximately equal size to demonstrate the refined random forest model's validity.

2.5. Laplacian Eigenmaps

We performed spectral cluster analysis using a Laplacian eigenmaps method to visualise relationships between the respective teams and identify natural sub-groups within the data [24]. This approach involves computing the pairwise Euclidean distances between the respective teams using the key variables identified by the random forest analysis. These were transformed into a [150 × 150] similarity matrix, Q, using a Gaussian radial basis function (rbf) kernel [44], with 1, as follows:

$$Q = exp\left(-\frac{E^2}{2 \times \sigma^2}\right) \tag{1}$$

where E is the matrix of pairwise Euclidean distances. The non-linear Gaussian function filtered the Euclidean distance matrix so that edges between close neighbours were given more weight than those between teams more distantly separated. From this, the modified similarity matrix, W, was constructed by subtracting the [150 × 150] identify matrix, I, from the similarity matrix, Q:

$$W = Q - I \tag{2}$$

This was then used to construct the degree matrix, D, as follows:

$$s = W \cdot n \tag{3}$$

where n is a [150 × 1] vector of ones and D is:

$$D_{ij} = \begin{cases} s_i & if\ i = j \\ 0 & if\ i \neq j \end{cases} \tag{4}$$

Having computed the degree matrix, D, the Laplacian, L, and normalised Laplacian, L_{norm}, matrices (both symmetric, positive semi-definite matrices) were then constructed [45–47], as follows:

$$L = D - W \tag{5}$$

$$L_{norm} = D^{-0.5} \cdot L \cdot D^{-0.5} \tag{6}$$

After this, eigendecomposition of the normalised Laplacian matrix, L_{norm}, was performed to compute the diagonal matrix of eigenvalues, Λ, and the matrix of eigenvectors, V, as follows:

$$L_{norm} = V \cdot \Lambda \cdot V^T \tag{7}$$

However, unlike PCA, where the eigenvectors corresponding to the largest eigenvalues are used to construct the principal components, Laplacian eigenmaps construct a configuration from the eigenvectors corresponding to the two or three smallest positive eigenvalues. Because the smallest eigenvalue equals zero, the eigenvector corresponding to this eigenvalue is often ignored. Instead, the eigenvectors associated with the successive two or three smallest positive eigenvalues are used to construct the map [46]. We used the last three positive eigenvectors, the fourth, third, and second (Fielder) smallest eigenvectors, to construct 3D Laplacian eigenmaps of the European football teams. We used third and Fielder vectors to construct 2D Laplacian eigenmaps.

2.6. Natural Clustering Approach

Laplacian eigenmaps are a spectral clustering technique. As such, it exhibits a critical property discovered by [48], namely that the eigenvector associated with the second smallest eigenvalue (i.e., the smallest positive eigenvalue) can be used to partition a graph. The Fiedler vector, as it is known, is widely used in spectral graph partitioning [26–28] as an unsupervised technique for bisecting graphs, enabling sub-groups (clusters) within the data to be readily identified. Multiple sub-groups can be identified by repeated bisection of the Laplacian eigenmaps using the Fiedler vector [27].

To identify how many bisections were appropriate to establish the natural clusters in the data, we ran a cluster validation using the 'clValid' package in R [49]. To do so, we used the self-organising maps algorithm [50,51] since it is an unsupervised learning technique partitioning data using artificial neural networks. To determine the suitability of 2–6 partitions of the fiedler vector, internal consistency was measured by the Dunn Index [52] and Silhouette Width [53], both of which should be maximised [54]. The Silhouette Widths were also used to inspect final cluster classifications, following the Fielder vector's bisection. We created an undirected graph network using the inverse of the Euclidean distances between the respective teams to visualise natural clustering.

3. Results

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, and the experimental conclusions that can be drawn.

3.1. Descriptive Statistics

The descriptive analysis results using the aggregated data split by benchmark percentiles (top, middle, bottom) and the one-way ANOVA are presented in Table 2. Unsurprisingly, the top teams had significantly greater possession and pass success; conceded fewer shots; made more dribbles and shots than weaker teams (all $p < 0.001$); had greater possession and pass success (both $p < 0.001$); and made more dribbles and shots (both $p < 0.001$) than the weaker teams. In addition, they made significantly fewer fouls ($p = 0.037$) but did not significantly receive fewer yellow ($p = 0.214$) and red ($p = 0.406$) cards.

Table 2. Descriptive statistical results for aggregated data (all seasons) and one-way ANOVA results.

	Bottom (*n* = 37) Mean (SD)	**Middle (*n* = 81) Mean (SD)**	**Top (*n* = 32) Mean (SD)**	**Total (*n* = 150) Mean (SD)**	**ANOVA Sig.**	**Pairwise Significant Differences (*p* = < 0.05)**
Yellow_cards	78.379 (16.887)	76.219 (17.647)	71.335 (15.232)	75.710 (17.041)	0.214	Not Sig.
Red_cards	4.157 (1.848)	4.158 (1.812)	3.685 (1.510)	4.057 (1.761)	0.406	Not Sig.
Possession	46.090 (2.575)	47.693 (2.525)	54.557 (3.815)	48.762 (4.203)	<0.001	1,2,3
Pass_Success	75.221 (4.033)	76.082 (3.356)	82.464 (3.470)	77.231 (4.482)	<0.001	1,2,3
Aerials_Won	18.209 (4.623)	18.901 (3.386)	16.223 (3.069)	18.159 (3.793)	0.003	3
Shots_Conceeded	14.958 (1.653)	13.146 (1.360)	10.850 (1.146)	13.103 (1.967)	<0.001	1,2,3
Tackles	18.481 (2.373)	18.114 (1.279)	18.233 (1.430)	18.230 (1.639)	0.532	Not Sig.
Interceptions	14.177 (2.431)	14.542 (2.124)	13.606 (2.000)	14.252 (2.195)	0.12	Not Sig.
Fouls	13.591 (1.800)	13.482 (1.655)	12.656 (1.569)	13.333 (1.701)	0.037	2,3
Offsides	2.019 (0.441)	2.048 (0.306)	2.330 (0.392)	2.101 (0.379)	<0.001	2,3
Shots	11.380 (1.231)	11.651 (0.970)	14.533 (1.791)	12.199 (1.743)	<0.001	2,3
Shots_OT	3.617 (0.480)	3.888 (0.447)	5.325 (0.809)	4.128 (0.838)	<0.001	1,2,3
Dribbles	8.666 (2.052)	8.745 (1.292)	10.586 (1.508)	9.118 (1.725)	<0.001	2,3
Fouled	12.863 (1.988)	12.348 (1.616)	12.464 (1.353)	12.500 (1.668)	0.298	Not Sig.
GF	35.137 (5.426)	43.007 (5.643)	67.335 (13.671)	46.256 (13.963)	<0.001	1,2,3
GA	66.290 (10.102)	54.662 (6.296)	39.601 (6.200)	54.317 (11.667)	<0.001	1,2,3
GD	−31.153 (9.924)	−11.655 (8.944)	27.734 (17.872)	−8.062 (23.405)	<0.001	1,2,3
Points	29.614 (5.280)	43.591 (4.988)	69.025 (10.534)	45.569 (15.056)	<0.001	1,2,3

Legend: 1. Significant after Bonferroni adjustment between Bottom and Middle. 2. Significant after Bonferroni adjustment between Bottom and Top. 3. Significant after Bonferroni adjustment between Middle and Top.

3.2. Random Forest Analysis Results

The exploratory random forest analysis incorporating all the predictor variables produced a regression model with an MSE of 115.62 and an R^2 value of 0.7875 (or 78.75% explained variance), which was used to assess variable importance (see Figure 1). From Figure 1, it can be seen that the Gini VIM values for the five variables: Shots_OT (on target); Possession; Shots_conceded; Shots; and Pass_Success, were far more than the values for the other variables, which were subsequently discarded from the refined random forest regression model. As such, this indicates that these five variables were the best predictors of end-of-season goal difference.

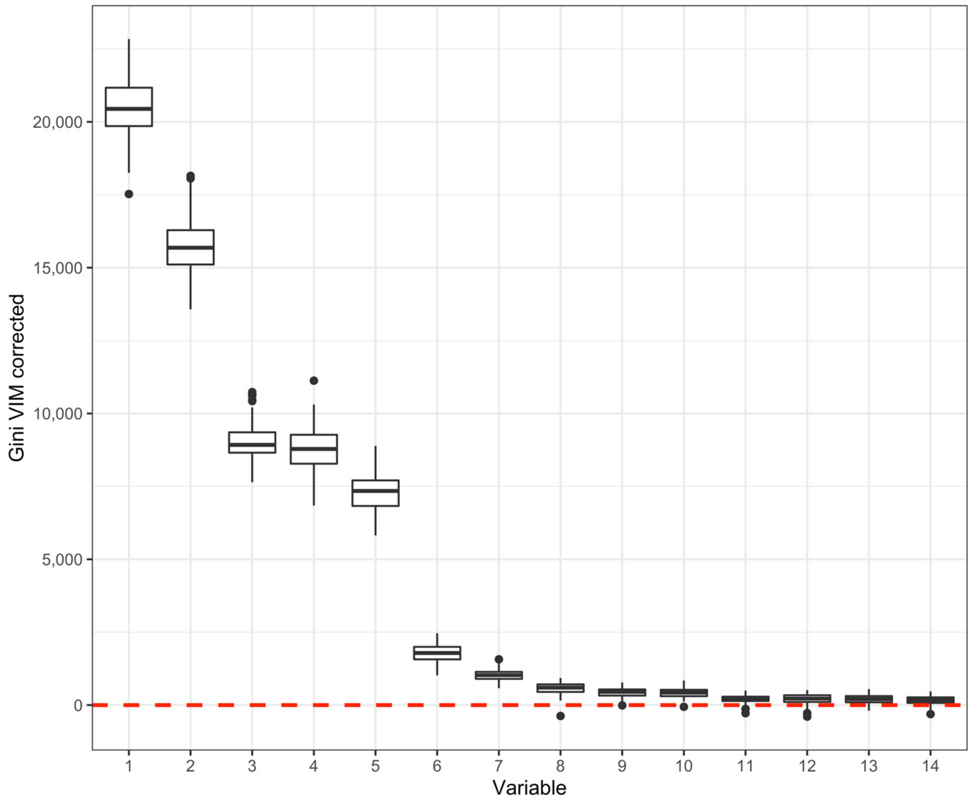


Figure 1. Random forest regression Gini corrected VIM. Legend: (1) Shots_OT; (2) Possession; (3) Shots; (4) Shots_Conceded; (5) Pass_Success; (6) Dribbles; (7); Arials_Won; (8) Offsides; (9) Tackles; (10) Yellow_Cards; (11) Red_Cards; (12) Fouled; (13) Interceptions; (14) Fouls.

The refined random forest analysis utilising only these important variables produced a regression model with an MSE of 113.84 and an R^2 value of 0.7908 (79.08% variance explained). The relationship between predicted and actual goal difference for the respective clubs is shown in Figure 2. From this, it can be seen that the refined random forest model predicted the end-of-season goal difference with a high degree of accuracy.

3.3. Laplacian Eigenmap Results

The 3D Laplacian eigenmaps of the teams are presented in Figure 3, which shows a scatter plot of the three smallest positive eigenvectors. The 3D plots demonstrate a spiral-like curve between the three dimensions, demonstrating a hierarchal structure. Figure 4 shows the 2D Laplacian eigenmap with the Fielder vector plotted against the third smallest eigenvector. Here it shows a characteristic U-shaped curve, with the teams distributed along its length. In Figure 4, the teams are classified according to the 25% and 75% percentile points benchmark groupings. From this, it is relatively clear that most top clubs plot to the Fiedler vector's right (>0.1), with a relatively clear distinction from the rest. Similarly, the bottom clubs tend to plot to the left of the Fielder vector (<0). However, middle clubs have a less clear space along the curve. Interestingly, La Palmas (Team No. 120; La Liga), who were benchmarked bottom, and Nice (Team No. 61; Ligue 1) plot closer to the top benchmarked teams >0.1 on the Fielder vector.

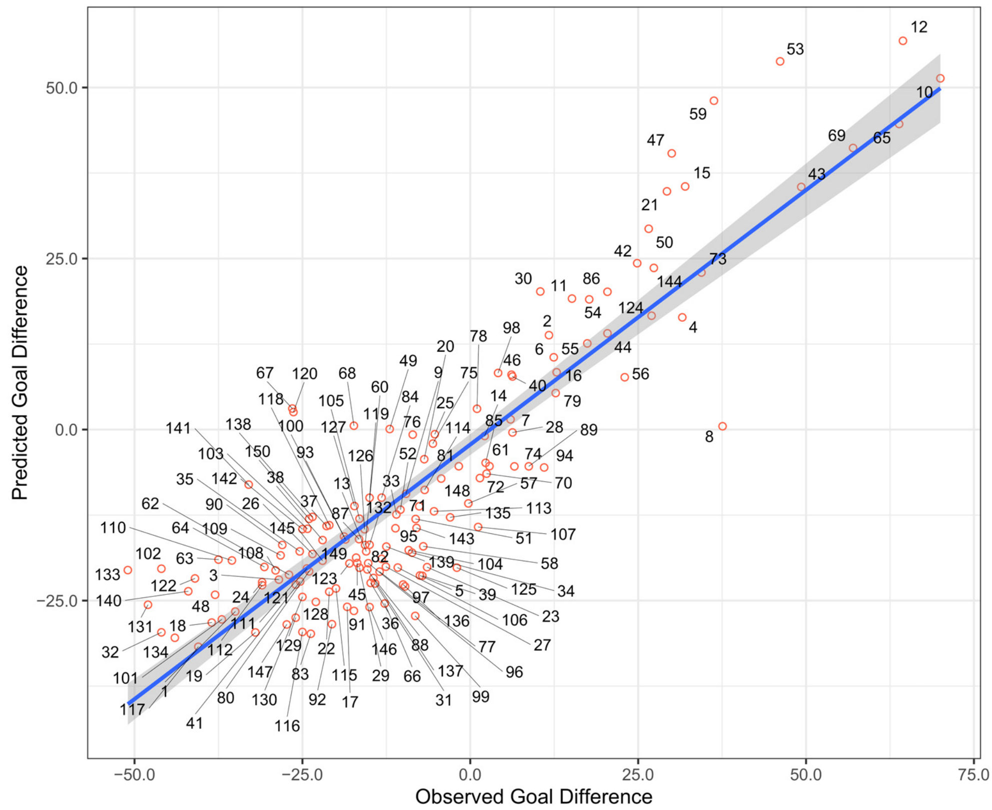
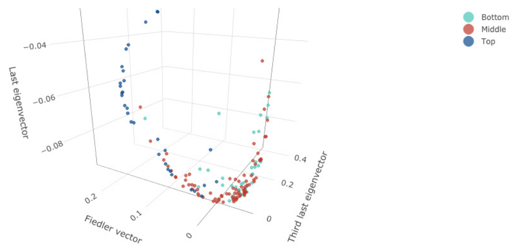
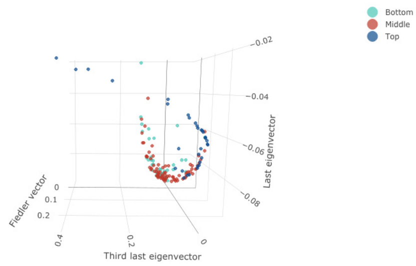


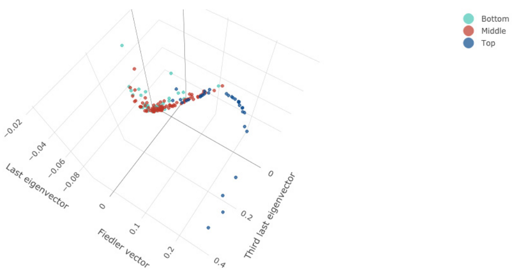
Figure 2. Scatter plot of predicted goal difference versus actual goal difference for the refined random forest regression model.



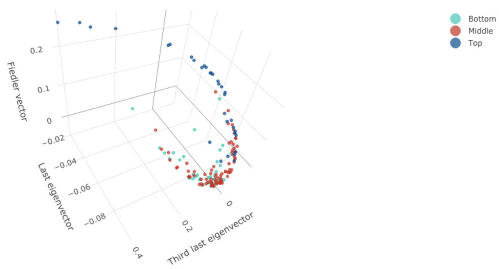
(**a**)



(**b**)



(**c**)



(**d**)

Figure 3. 3D scatter plot of Laplacian eigenmap using the three smallest positive eigenvectors (**a**–**d**) offer different angles of the 3D structure.

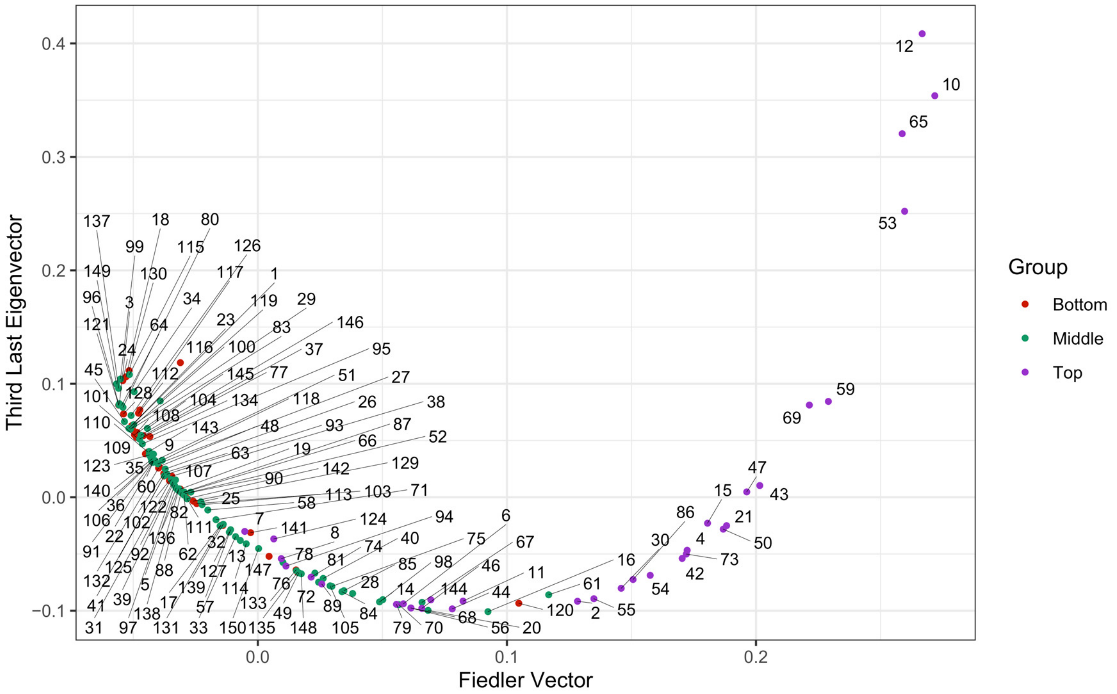


Figure 4. 2D scatter plot of Laplacian eigenmap using the Fielder vector and the third smallest eigenvector.

When the benchmarked classifications are mapped onto a network graph of the inverse Euclidean distances (Figure 5), it can be seen that although top teams cluster to the bottom right, there is considerable overlap across all top, middle and bottom teams. Indeed, the average silhouette width values for the benchmark classifications were only 0.04, indicating that classification based on the national leagues' points does not accurately reflect the natural groupings between the various European soccer clubs.

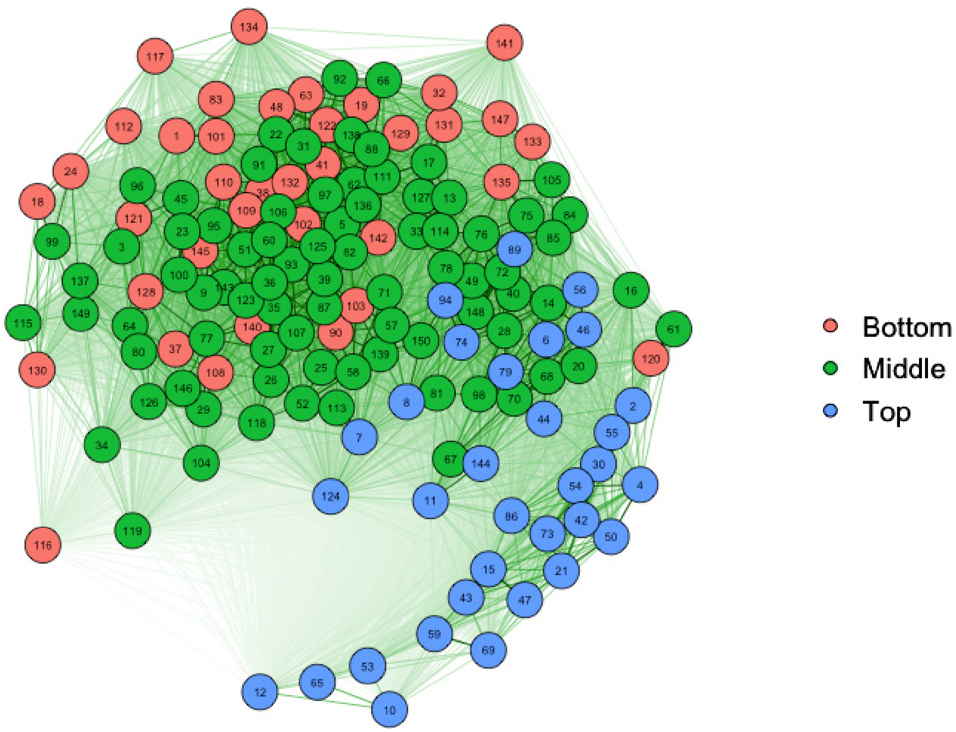


Figure 5. Network graph of the benchmark clusters.

3.4. Natural Clustering Results

The Dunn Index and Silhouette Width results for the self-organising maps cluster validation are presented in Figure 6. It is clear that the 4-cluster solution maximises both internal validation measures, requiring three bisections of the Fielder vector.

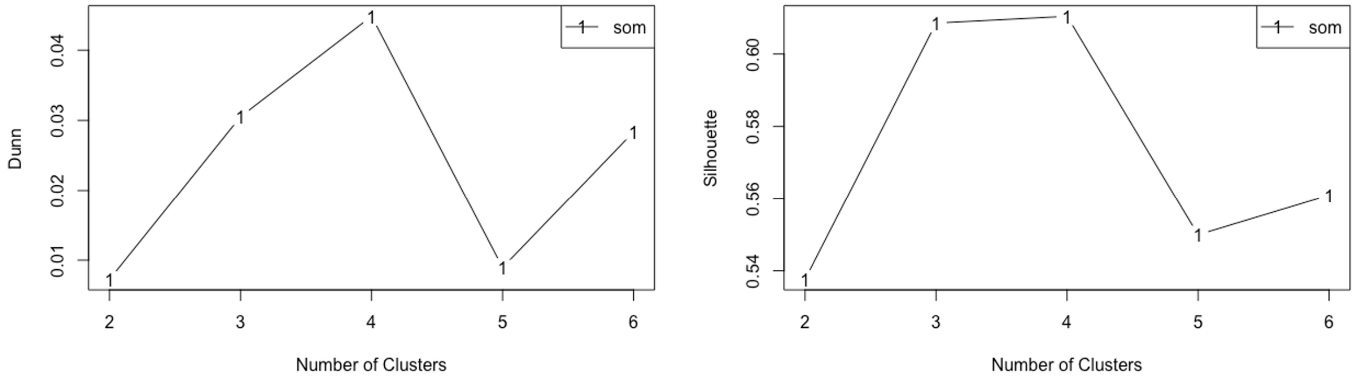


Figure 6. Dunn Index (**left**) and Silhouette Width (**right**) cluster validation for 2 to 6 clusters using self-organising maps algorithm.

The three bisections of the Fielder vector are presented in Figure 7, creating 4 clusters SC1-SC4. Here the clusters demonstrate a group of four very strong dominating teams (SC1), fifteen strong teams (SC2), thirty-seven medium-strength teams (SC3), and ninety-four weaker teams (SC4). Overall, the natural clusters identified by the Fielder vector algorithm are well defined, with an average Silhouette Width = 0.61, no cluster below 0.50 (Figure 8 and Table 3), and a Dunn Index = 0.0098. The lowest internally valid cluster is SC2 with a Silhouette Index = 0.50, suggesting this group is more heterogeneous than homogeneous. The natural clustering network graph is visualised in Figure 9, which shows how cohesive the clusters are based on the inverse Euclidian distance.

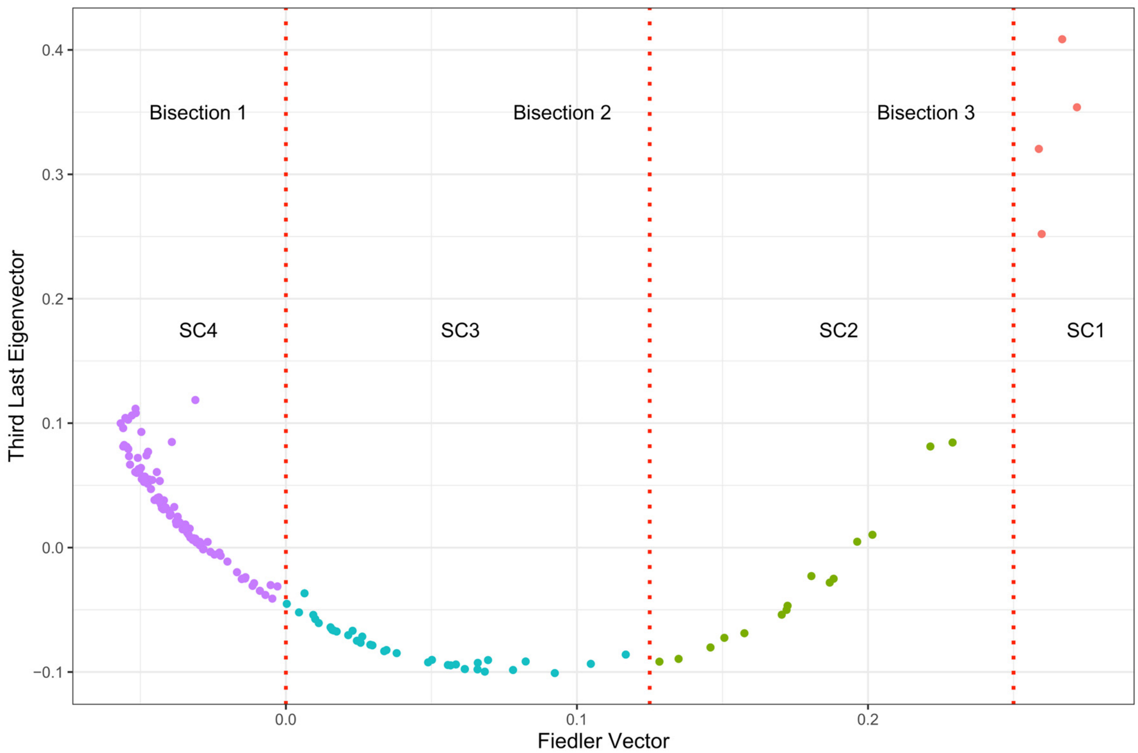


Figure 7. Natural clusters from bisections of the Fiedler vector.

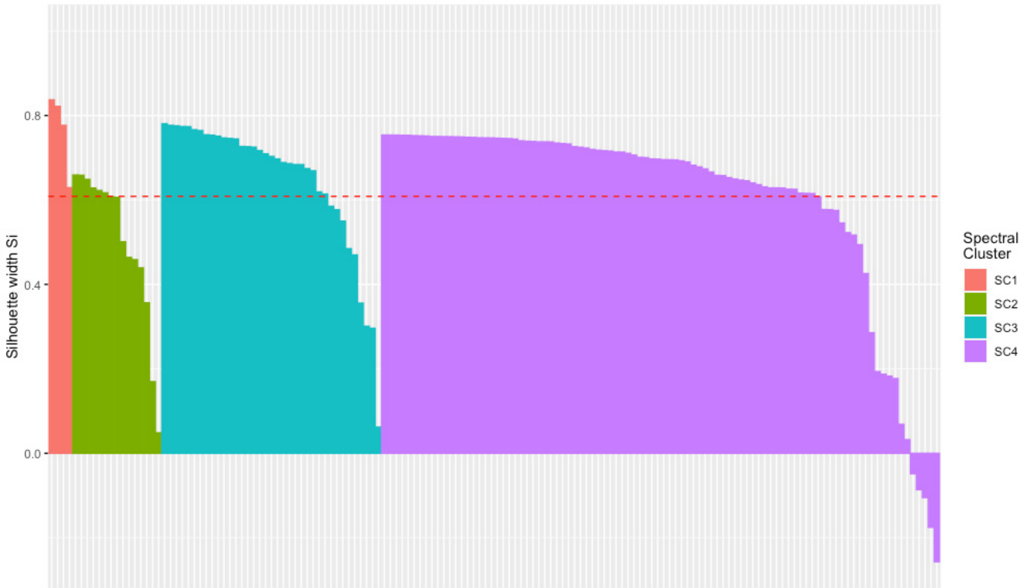


Figure 8. Silhouette Widths for each cluster.

Table 3. Number of teams in each cluster and Silhouette Index.

Cluster	Number of Teams	Team ID	Silhouette Index
SC1	4	53, 12, 65, 10	0.77
SC2	15	30, 2, 55, 50, 15, 54, 42, 86, 47, 21, 4, 73, 59, 69, 43	0.50
SC3	37	133, 147, 135, 120, 150, 105, 67, 148, 68, 49, 84, 76, 20, 75, 81, 78, 40, 98, 14, 72, 85, 70, 61, 28, 16, 46, 74, 11, 89, 6, 94, 144, 79, 44, 56, 124, 8	0.65
SC4	94	131, 102, 63, 1, 101, 24, 48, 134, 110, 122, 140, 117, 129, 32, 90, 108, 112, 116, 18, 19, 38, 142, 145, 109, 37, 130, 83, 121, 128, 141, 41, 132, 103, 3, 93, 119, 62, 118, 138, 127, 87, 64, 106, 111, 115, 22, 35, 31, 91, 126, 96, 80, 143, 13, 92, 26, 29, 17, 9, 146, 149, 95, 100, 66, 52, 123, 51, 88, 45, 60, 39, 36, 33, 5, 25, 139, 113, 77, 27, 104, 137, 71, 99, 97, 82, 58, 23, 34, 136, 57, 114, 107, 125, 7	0.60

Using the Fielder vector allows natural groupings of teams (or firms) to be created. The results show that using the Fielder vector algorithm is relatively effective in finding natural clusters within European football teams. Using unsupervised machine learning and clustering methods, we can objectively identify the dominant teams across Europe. Therefore, clusters 1 and 2 demonstrate the best teams to compete in an elite European Super League—should it be created.

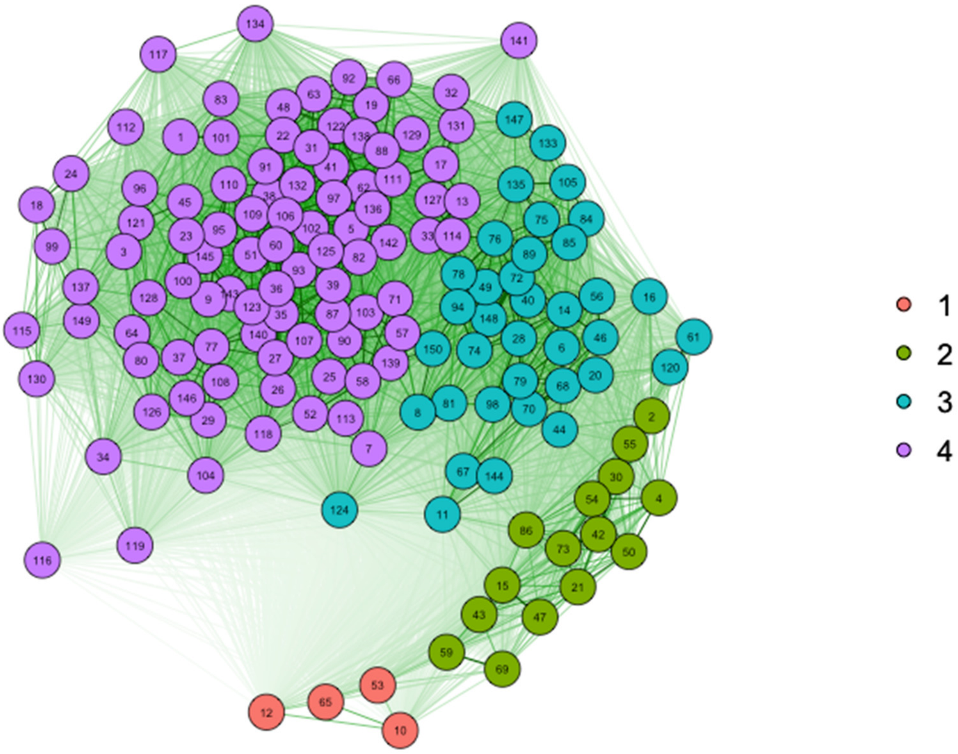


Figure 9. Network graph of the Natural Clustering in European Football.

4. Discussion

For this paper, we hypothesised that the top European teams would tend to cluster together according to their on-the-pitch performance characteristics. Accordingly, we developed an unsupervised data-driven approach that utilised a Laplacian eigenmap to identify the natural clusters of teams across the five major football leagues in Europe. We aimed to develop a robust objective methodology for selecting teams suitable for inclusion in a future ESL. To this end, we could demonstrate that the respective teams did indeed form natural clusters independent of the leagues from which they came (Figure 9) and that these could be readily identified using the Fiedler vector without any subjective input. Furthermore, concerning the question of 'who' are the top teams in European soccer, the Laplacian eigenmap methodology classified 15 out of the 16 'breakaway' ELS teams as candidates for the elite league [55]. However, our approach did not select Atlético Madrid and instead included Napoli, Tottenham Hotspur, Lyon, and Fiorentina in the elite group, which comprised 19 teams and two sub-clusters, SC1 and SC2 (Table 4). Interestingly, all 19 teams were ranked in the top 25% of their respective domestic leagues (Figure 5).

Unlike conventional ranking systems, sorting teams according to a single metric, our methodology enabled the similarities and differences between the respective football teams from disparate leagues to be mapped onto a 2D space. Using a Laplacian eigenmap of the Euclidean distance graph, we were able to project complex multivariate non-linear relationships within the match performance data onto a 2D space, making it easy to visualise the distances between the respective teams, thus identifying the natural neighbourhoods in which teams inhibit. Through the bisection of the Fiedler vector, we showed how these natural neighbourhoods created suitable clusters to categorise teams. Using the variables that best predict goal difference, we were able to show that this approach could identify the teams who dominated their respective leagues based on actual performance rather than points earned. For example, using the performance metrics of shots on target, possession, shots, shots conceded, and pass success, we were able to demonstrate that Barcelona was much closer to Paris Saint-Germain and Bayern Munich than Real Madrid, and that Arsenal, Inter Milan, and Roma were all closely related. Indeed, we were surprised by just how well these match performance metrics could cluster the top teams, even though the teams came from different leagues and points and goals were not involved in the methodology. As such, this supports the opinion that successful teams tend to share similar game style characteristics [19–22]. However, further work will be required to determine whether this is true or false.

Table 4. The Laplacian eigenvector approach to a new elite European Super League.

	Team_ID	Team	Tournament	Cluster
1	10	Barcelona	La Liga	SC1
2	12	Bayern Munich	Bundesliga	SC1
3	53	Manchester City	Premier League	SC1
4	65	Paris Saint Germain	Ligue 1	SC1
5	2	AC Milan	Serie A	SC2
6	4	Arsenal	Premier League	SC2
7	15	Borussia Dortmund	Bundesliga	SC2
8	21	Chelsea	Premier League	SC2
9	30	Fiorentina	Serie A	SC2
10	42	Inter Milan	Serie A	SC2
11	43	Juventus	Serie A	SC2
12	47	Liverpool	Premier League	SC2
13	50	Lyon	Ligue 1	SC2
14	54	Manchester United	Premier League	SC2
15	55	Marseille	Ligue 1	SC2
16	59	Napoli	Serie A	SC2
17	69	Real Madrid	La Liga	SC2
18	73	Roma	Serie A	SC2
19	86	Tottenham	Premier League	SC2

The methodology described in this paper is completely new to the field of sports analytics and could be applied to multiple applications within sports and wider fields of operational research. Within a footballing context, the approach could be applied to understanding which players naturally cluster together based on their performance metrics. This could then be used to aid decision-making regarding player acquisitions and development. Likewise, a similar approach could be used to support merger and acquisition decisions by identifying creditable target firms or help in understanding the impact of strategic choices when attempting to create a competitive advantage.

While the work reported here suggests that our approach might have wider applicability in sport than just selecting teams for inclusion in the ESL, further work will be required to refine the technique and identify suitable problems to which the methodology is well suited. However, with specific reference to the selection of teams for the ESL, one of the limitations of the present study is that we only used simple performance metrics that were open source and thus freely available. Therefore, further work is recommended to identify the performance metrics that optimise cluster identification and best describe the similarities and differences between the respective teams. Another limitation of our study is that we did not compare the teams selected by bisecting the Fiedler vector with those that might be selected using the various ranking systems. Therefore, further work should be undertaken to evaluate how our approach's results compare with those produced by the more traditional ranking systems.

5. Conclusions

In conclusion, we have shown it is possible to identify the top soccer teams in Europe using only match performance data (i.e., shots on target, possession, shots, shots conceded, and pass success) collected from their respective domestic leagues (i.e., the Bundesliga, La Liga, Ligue 1, English Premier League, and Serie A). Furthermore, using a novel unsupervised Laplacian eigenmap approach, we could visualise the similarities and differences between the respective teams in Europe and identify the natural clusters that exist without resorting to any a priori knowledge. As such, we identified 15 of the 16 top teams invited to participate in the elite European Super League in 2021. This suggests that the top teams in Europe exhibit similar playing styles that cause them to cluster into natural communities irrespective of the domestic league from which they come.

Author Contributions: Conceptualisation, A.J.B. and C.B.B.; Methodology, C.B.B.; Software, C.B.B.; Data Curation: A.J.B. and C.B.B.; Formal Analysis, A.J.B. and C.B.B.; Writing—Original Draft Preparation, A.J.B. and C.B.B.; Writing—Review & Editing, A.J.B. and C.B.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data is publicly available at https://whoscored.com accessed on 23 March 2021.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Team_ID	Team	Tournament	Team_ID	Team	Tournament
1	AC Ajaccio	Ligue 1	76	Sassuolo	Serie A
2	AC Milan	Serie A	77	SC Bastia	Ligue 1
3	Almeria	La Liga	78	Schalke 04	Bundesliga
4	Arsenal	Premier League	79	Sevilla	La Liga
5	Aston Villa	Premier League	80	Sochaux	Ligue 1
6	Atalanta	Serie A	81	Southampton	Premier League
7	Athletic Bilbao	La Liga	82	Stoke	Premier League
8	Atletico Madrid	La Liga	83	Sunderland	Premier League
9	Augsburg	Bundesliga	84	Swansea	Premier League
10	Barcelona	La Liga	85	Torino	Serie A
11	Bayer Leverkusen	Bundesliga	86	Tottenham	Premier League
12	Bayern Munich	Bundesliga	87	Toulouse	Ligue 1
13	Bologna	Serie A	88	Udinese	Serie A
14	Bordeaux	Ligue 1	89	Valencia	La Liga
15	Borussia Dortmund	Bundesliga	90	Valenciennes	Ligue 1
16	Borussia M.Gladbach	Bundesliga	91	Valladolid	La Liga
17	Cagliari	Serie A	92	Verona	Serie A
18	Cardiff	Premier League	93	VfB Stuttgart	Bundesliga
19	Catania	Serie A	94	Villarreal	La Liga
20	Celta Vigo	La Liga	95	Werder Bremen	Bundesliga
21	Chelsea	Premier League	96	West Bromwich Albion	Premier League
22	Chievo	Serie A	97	West Ham	Premier League
23	Crystal Palace	Premier League	98	Wolfsburg	Bundesliga
24	Eintracht Braunschweig	Bundesliga	99	Burnley	Premier League
25	Eintracht Frankfurt	Bundesliga	100	Caen	Ligue 1
26	Elche	La Liga	101	Cesena	Serie A
27	Espanyol	La Liga	102	Cordoba	La Liga
28	Everton	Premier League	103	Deportivo La Coruna	La Liga
29	Evian Thonon Gaillard	Ligue 1	104	Eibar	La Liga
30	Fiorentina	Serie A	105	Empoli	Serie A
31	Freiburg	Bundesliga	106	FC Koln	Bundesliga
32	Fulham	Premier League	107	Leicester	Premier League
33	Genoa	Serie A	108	Lens	Ligue 1
34	Getafe	La Liga	109	Metz	Ligue 1
35	Granada	La Liga	110	Paderborn	Bundesliga

Team_ID	Team	Tournament	Team_ID	Team	Tournament
36	Guingamp	Ligue 1	111	Palermo	Serie A
37	Hamburger SV	Bundesliga	112	Queens Park Rangers	Premier League
38	Hannover 96	Bundesliga	113	Angers	Ligue 1
39	Hertha Berlin	Bundesliga	114	Bournemouth	Premier League
40	Hoffenheim	Bundesliga	115	Carpi	Serie A
41	Hull	Premier League	116	Darmstadt	Bundesliga
42	Inter Milan	Serie A	117	Frosinone	Serie A
43	Juventus	Serie A	118	GFC Ajaccio	Ligue 1
44	Lazio	Serie A	119	Ingolstadt	Bundesliga
45	Levante	La Liga	120	Las Palmas	La Liga
46	Lille	Ligue 1	121	Sporting Gijon	La Liga
47	Liverpool	Premier League	122	Troyes	Ligue 1
48	Livorno	Serie A	123	Watford	Premier League
49	Lorient	Ligue 1	124	RasenBallsport Leipzig	Bundesliga
50	Lyon	Ligue 1	125	Alaves	La Liga
51	Mainz 05	Bundesliga	126	Leganes	La Liga
52	Malaga	La Liga	127	Dijon	Ligue 1
53	Manchester City	Premier League	128	Nancy	Ligue 1
54	Manchester United	Premier League	129	Middlesbrough	Premier League
55	Marseille	Ligue 1	130	Crotone	Serie A
56	Monaco	Ligue 1	131	Pescara	Serie A
57	Montpellier	Ligue 1	132	Amiens	Ligue 1
58	Nantes	Ligue 1	133	Benevento	Serie A
59	Napoli	Serie A	134	Brescia	Serie A
60	Newcastle United	Premier League	135	Brest	Ligue 1
61	Nice	Ligue 1	136	Brighton	Premier League
62	Norwich	Premier League	137	Deportivo Alaves	La Liga
63	Nuernberg	Bundesliga	138	Fortuna Duesseldorf	Bundesliga
64	Osasuna	La Liga	139	Girona	La Liga
65	Paris Saint Germain	Ligue 1	140	Huddersfield	Premier League
66	Parma	Serie A	141	Lecce	Serie A
67	Rayo Vallecano	La Liga	142	Mallorca	La Liga
68	Real Betis	La Liga	143	Nimes	Ligue 1
69	Real Madrid	La Liga	144	RB Leipzig	Bundesliga
70	Real Sociedad	La Liga	145	SDHuesca	La Liga
71	Reims	Ligue 1	146	Sheffield United	Premier League
72	Rennes	Ligue 1	147	SPAL	Serie A
73	Roma	Serie A	148	Strasbourg	Ligue 1
74	Saint-Etienne	Ligue 1	149	Union Berlin	Bundesliga
75	Sampdoria	Serie A	150	Wolverhampton Wanderers	Premier League

References

1. West, A. *European Super League: Will Football Follow Basketball's Lead?-BBC Sport*; BBC Sport: London, UK, 2018.
2. Marcotti, G. Super League Suspended-Why English Clubs Pulled out and What's Next for Them and UEFA. Available online: https://www.espn.co.uk/football/blog-marcottis-musings/story/4365465/super-league-suspended-why-english-clubs-pulled-outwhats-next-for-them-and-uefa (accessed on 5 July 2021).
3. Deloitte Home Truths. *Annual Review of Football Finance 2020*; Deloitte: London, UK, 2020.
4. Bond, A.J.; Addesa, F. TV Demand for the Italian Serie A: Star Power or Competitive Intensity? *Econ. Bull.* **2019**, *39*, 2110–2116.
5. Bond, A.J.; Addesa, F. Competitive Intensity, Fans' Expectations, and Match-Day Tickets Sold in the Italian Football Serie A, 2012–2015. *J. Sport. Econ.* **2020**, *21*, 20–43. [CrossRef]
6. Caruso, R.; Addesa, F.; Di Domizio, M. The Determinants of the TV Demand for Soccer: Empirical Evidence on Italian Serie A for the Period 2008-2015. *J Sports Econ* **2019**, *20*, 25–49. [CrossRef]
7. Langville, A.N.; Meyer, C.D. *Who's #1?: The Science of Rating and Ranking*; Princeton University Press: Princeton, NJ, USA, 2012.
8. Elo, A.E. Logistic probability as a rating basis. In *The Rating of Chessplayers, Past & Present*; ARCO Publishing. Inc.: New York, NY, USA, 2008.

9. Colley, W.N. Colley's Bias Free College Football Ranking Method: The Colley Matrix Explained. Available online: https://colleyrankings.com/matrate.pdf (accessed on 9 January 2023).
10. Beggs, C.B.; Shepherd, S.J.; Emmonds, S.; Jones, B. A Novel Application of PageRank and User Preference Algorithms for Assessing the Relative Performance of Track Athletes in Competition. *PLoS ONE* **2017**, *12*, e0178458. [CrossRef]
11. Massey, K. *Statistical Models Applied to the Rating of Sports Teams*; Bluefield College: Bluefield, VA, USA, 1997; p. 1077.
12. Keener, J.P. The Perron–Frobenius Theorem and the Ranking of Football Teams. *SIAM Rev.* **1993**, *35*, 80–93. [CrossRef]
13. Page, L.; Brin, S.; Motwani, R.; Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. 1999. Available online: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf (accessed on 23 March 2021).
14. Brin, S.; Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Comput. Netw. ISDN Syst.* **1998**, *30*, 107–117. [CrossRef]
15. EuroClubIndex.com. Latest Ranking-Euro Club Index. Available online: https://www.euroclubindex.com/ (accessed on 23 March 2021).
16. clubelo.com. Football Club Elo Ratings. Available online: http://clubelo.com/ (accessed on 23 March 2021).
17. UEFA Club Coefficients. Available online: https://www.uefa.com/memberassociations/uefarankings/club/#/yr/2021 (accessed on 23 March 2021).
18. UEFA.com. How the Club Coefficients Are Calculated News UEFA Coefficients UEFA.Com. Available online: https://www.uefa.com/nationalassociations/uefarankings/news/0252-0cda38714c0d-0874ab234eb6-1000--how-the-club-coefficients-are-calculated/ (accessed on 12 January 2023).
19. Kempe, M.; Vogelbein, M.; Memmert, D.; Nopp, S. Possession vs. Direct Play: Evaluating Tactical Behavior in Elite Soccer. *Int. J. Sport. Sci.* **2014**, *4*, 35–41. [CrossRef]
20. Lago-Peñas, C.; Dellal, A. Ball Possession Strategies in Elite Soccer According to the Evolution of the Match-Score: The Influence of Situational Variables. *J. Hum. Kinet.* **2010**, *25*, 93–100. [CrossRef]
21. Castellano, J.; Casamichana, D.; Lago, C. The Use of Match Statistics That Discriminate between Successful and Unsuccessful Soccer Teams. *J. Hum. Kinet.* **2012**, *31*, 139–147. [CrossRef]
22. Gómez, M.Á.; Mitrotasios, M.; Armatas, V.; Lago-Peñas, C. Analysis of Playing Styles According to Team Quality and Match Location in Greek Professional Soccer. *Int. J. Perform. Anal. Sport* **2018**, *18*, 986–997. [CrossRef]
23. Akhanli, S.E.; Hennig, C. Some Issues in Distance Construction for Football Players Performance Data. *Arch. Data Sci.* **2017**, *2*, 1–17. [CrossRef]
24. Belkin, M.; Niyogi, P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Comput.* **2003**, *15*, 1373–1396. [CrossRef]
25. Nascimento, M.C.V.; de Carvalho, A.C.P.L.F. Spectral Methods for Graph Clustering—A Survey. *Eur. J. Oper. Res.* **2011**, *211*, 221–231. [CrossRef]
26. Higham, D.J.; Kalna, G.; Kibble, M. Spectral Clustering and Its Use in Bioinformatics. *J. Comput. Appl. Math.* **2007**, *204*, 25–37. [CrossRef]
27. Naumov, M.; Moon, T. Parallel spectral graph partitioning on CUDA. In *NVIDIA Technical Report NVR-2016-001*; NVIDIA: Santa Clara, CA, USA, 2016.
28. Stone, E.A.; Griffing, A.R. On the Fiedler Vectors of Graphs That Arise from Trees by Schur Complementation of the Laplacian. *Linear Algebra Appl.* **2009**, *431*, 1869–1880. [CrossRef]
29. FootyStats.com. Complete List of Football Leagues with Stats. Available online: https://footystats.org/leagues (accessed on 23 March 2021).
30. WhoScored.com. Football Statistics. Available online: https://www.whoscored.com/Statistics (accessed on 23 March 2021).
31. Heuer, A.; Rubner, O. Fitness, Chance and Myths: An Objective View on Soccer Results. *Eur. Phys. J. B* **2009**, *67*, 445–458. [CrossRef]
32. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.
33. Boinee, P.; Angelis, A.; De Foresti, G.L. Meta Random Forests. *Int. J. Comput. Inf. Eng.* **2008**, *18*, 1148–1157. [CrossRef]
34. Breiman, L. Random Forests. *Mac. Learn.* **2001**, *45*, 5–32. [CrossRef]
35. Izenman, A.J. *Modern Multivariate Statistical Techniques*; Springer: Berlin/Heidelberg, Germany, 2013.
36. Hansen, G.J.A.; Carpenter, S.R.; Gaeta, J.W.; Hennessy, J.M.; Vander Zanden, M.J. Predicting Walleye Recruitment as a Tool for Prioritizing Management Actions. *Can. J. Fish. Aquat. Sci.* **2015**, *72*, 661–672. [CrossRef]
37. Cutler, D.R.; Edwards, T.C.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J.; Lawler, J.J. Random Forests for Classification in Ecology. *Ecology* **2007**, *88*, 2783–2792. [CrossRef]
38. Pecl, G.T.; Tracey, S.R.; Danyushevsky, L.; Wotherspoon, S.; Moltschaniwskyj, N.A. Elemental Fingerprints of Southern Calamary (Sepioteuthis Australis) Reveal Local Recruitment Sources and Allow Assessment of the Importance of Closed Areas. *Can. J. Fish. Aquat. Sci.* **2011**, *68*, 1351–1360. [CrossRef]
39. Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R News* **2002**, *2*, 18–22.
40. Sandri, M.; Zuccolotto, P. A Bias Correction Algorithm for the Gini Variable Importance Measure in Classification Trees. *J. Comput. Graph. Stat.* **2008**, *17*, 611–628. [CrossRef]
41. Sandri, M.; Zuccolotto, P. Analysis and Correction of Bias in Total Decrease in Node Impurity Measures for Tree-Based Algorithms. *Stat. Comput.* **2010**, *20*, 393–407. [CrossRef]

42. Carpita, M.; Sandri, M.; Simonetto, A.; Zuccolotto, P. Football Mining with R. In *Data Mining Applications with R*; Zhao, Y., Cen, Y.B., Eds.; Elsevier: Boston, MA, USA, 2014; pp. 397–433. ISBN 978-0-12-411511-8.
43. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232.
44. Schölkopf, B.; Tsuda, K.; Vert, J.-P. *Kernel Methods in Computational Biology*; MIT Press: Cambridge, MA, USA, 2004; ISBN 9780262256926.
45. Chung, F.R.K.; Graham, F.C. *Spectral Graph Theory*; American Mathematical Soc.: Providence, RI, USA, 1997; ISBN 0821803158.
46. Qiu, H.; Hancock, E.R. Spectral simplification of graphs. In Proceedings of the European Conference on Computer Vision, Berlin, Germany, 11–14 May 2004; pp. 114–126.
47. von Luxburg, U. A Tutorial on Spectral Clustering. *Stat. Comput.* **2007**, *17*, 395–416. [CrossRef]
48. Fiedler, M. A Property of Eigenvectors of Nonnegative Symmetric Matrices and Its Application to Graph Theory. *Czechoslov. Math. J.* **1975**, *25*, 619–633.
49. Brock, G.; Pihur, V.; Datta, S.; Datta, S. ClValid : An R Package for Cluster Validation. *J. Stat. Softw.* **2008**, *25*, 371–372. [CrossRef]
50. Kohonen, T. Self-organising maps: Ophmization approaches. In *Artificial Neural Networks*; Kohonen, T., Mäkisara, K., Simula, O., Kangas, J., Eds.; Elsevier: Espoo, Finland, 1991; pp. 981–990. ISBN 978-0-444-89178-5.
51. Kohonen, T. *Self-Organizing Maps*; Springer Science & Business Media: Berlin, Germany, 2012; Volume 30, ISBN 3642569277.
52. Dunn, J.C. Well-Separated Clusters and Optimal Fuzzy Partitions. *J. Cybern.* **1974**, *4*, 95–104. [CrossRef]
53. Rousseeuw, P.J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
54. Handl, J.; Knowles, J.; Kell, D.B. Computational Cluster Validation in Post-Genomic Data Analysis. *Bioinformatics* **2005**, *21*, 3201–3212. [CrossRef]
55. Der Spiegel Football. *Documents Show Secret Plans for Elite League of Top Clubs-Der Spiegel*; Der Spiegel: Hamburg, Germany, 2018.