Functional Clustering of Elo Ratings for Competitive Balance Analysis in Professional Soccer Leagues

Jinglin Feng[1] and Andrew Hwang[2]

[1]Pennsylvania State University, Old Main, State College, PA 16802
[2]Pennsylvania State University, Old Main, State College, PA 16802

Abstract
We assess the competitive balance for top two divisions within two professional soccer leagues: The Bundesliga and the La Liga. In order to facilitate comparisons between teams in different leagues over time, we use Elo ratings, a score assigned to each team and adjusted after each match based on various match-related variables, to represent a team's strength at a particular point in time. We treat them as functional data over time, where each team has a curve (function) of Elo ratings over an interval of time, and the collection of curves is our sample. To gain insight into the history of success of teams within these leagues, we use a clustering algorithm that accommodates the sparse, irregularly-sampled data to cluster the team curves. After our clustering analysis, we not only trace the curve of each team within each cluster but also consider team factors that characterize each cluster. Our results confirm our expectations that both La Liga and Bundesliga soccer leagues are dominated by a few teams, indicating that leagues need to review their structures and policies to facilitate greater competitive balance.

Key Words: Competitive balance, Elo ratings, Functional data analysis, Clustering algorithms, Professional Soccer League

1. Introduction

Soccer is generally considered the world's most popular sport. "The Beautiful Game," as it is known by many people, has captured the passion, imagination, and loyalty of fans across all continents. As a result, professional soccer has become one of the most lucrative entertainment industries. Given the money involved in the sport, it is little surprise that the various professional leagues are under constant scrutiny by policymakers. Fans, coaches, players, and investors alike have large stakes in the success of their teams, obliging league owners to ensure some semblance of fairness in the structure of these leagues. This idea of fairness is called competitive balance. [1]

1.1 Competitive balance

Competitive balance is primarily an economic topic concerned with a market situation where no business is too big or has an unfair advantage. It has been applied to different sport leagues, including professional soccer leagues, where fans are keenly interested in

[1] Pennsylvania State University, Old Main, State College, PA, 16802
[2] Pennsylvania State University, Old Main, State College, PA, 16802

league parity. Despite the burgeoning interest in competitive balance, metrics for quantifying and analyzing competitive balance are still relatively rudimentary.

Of interest in this study is the historic dominance of the German and Spanish soccer league systems by several teams. At the start of each season in the Bundesliga and the La Liga, fans of all teams can safely assume that teams such as FC Bayern Munich, Borussia Dortmund, Barcelona, and Real Madrid will all finish near the top of their league tables. Some see such dominance as destructive to the competitive balance of the leagues, keeping the leagues' coveted trophy in the hands of the elite few teams. We hope that by relying on the clustering tools from the realm of machine learning, we can illustrate just how pervasive this dominance is in both German and Spanish soccer systems. Specifically, we introduce a novel approach by using a clustering method to cluster teams based on Elo Ratings in the most recent five seasons to examine competitive balance in these two professional soccer leagues, with the expectation that teams mentioned above will be clustered in a class separate from most other teams based on their consistently high Elo Ratings each season.

2. Data Description

2.1 Elo Ratings

The data [1] analyzed for this project are World Football Elo Ratings based on the Elo-system. Elo Ratings are scores assigned to each club and adjusted after each match based on various match-related variables such as home field advantage, goal difference, and inter-league adjustments. The advantage of Elo lies in its simplicity, there is only one value per club for each point in time, the higher the better. We decide to use Elo Ratings rather than other rating system is because this system is better to measure how good a team is relative to its opponents over a certain time period, in particular for performance in recent time periods.

2.1.1 Data Collecting and Cleaning

We first collect Elo Ratings for the top two divisions of the Bundesliga and the La Liga. Originally, we have 36 teams in the Bundesliga and 42 teams in the La Liga, however, we find that several teams do not have complete data for the most recent five seasons (started from 2011-2012 season to 2015-2016 season), so we only retain teams that have complete data in the most recent five seasons. It turns out that we have 29 teams in the Bundesliga and 28 teams in the La Liga finally.

We treat all Elo Ratings as functional data over time for the selected 29 Bundesliga teams and 28 La Liga teams and code 1/1/2010 as 0 and sequentially, 12/31/2016 as 2191.

3. Methodology

3.1 Cluster algorithm

3.1.1 "distFPCA" method [3]

Due to the spaced game days, we have sparse and irregular data. In order to accommodate such data, we choose “distFPCA” algorithm, a clustering algorithm based on a distance measure in “funcit” function from “funcy” package in R. In this “funcit” function, we consider the number of clusters, k, is equal to 2, 3, 4, 5 and 6.

3.1.2 “Jump” method

We use “Jump” method [2], a method which is derived from within-cluster-sum-of-square (WSS) method, to determine the optimal number of k. Output from the “Funcit” function contains a value called cldist, which should be a matrix of N×2 (the number of teams by 2 columns). The first column contains the squared distance from that team’s curve to the closest cluster, while the second column contains the distance from the team’s curve to the second closest cluster. We only need column 1 (distance for each team to its assigned cluster). To calculate the WSS, we can use the formula

$$\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$$

Where K is the number of clusters, indexed by k, S_k is the set of curves in the k^{th} cluster; i indexes the curve(team); p is the number of observed time points for team i, so x_{ij} represents the observed point at j for a particular team i and $\bar{x}_{kj}$ represents the observed point j of the k^{th} cluster center (average curve of that cluster). This looks like a lot, but the “funcit” function seems to have done most of the work for us already, calculating

$$\sum_{k=1}^{K}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$$

for each team. So, all we have to do is to sum up all elements of the first column in cldist.

The “Jump” method basically looks at differences in WSS between consecutive numbers of clusters, and then exponentiates that difference. If we call the WSS a distortion and label it as $\hat{d}_k$ (the distortion for k clusters), we can calculate the “jump” as $J_k = \hat{d}_k^{-Y} - \hat{d}_{k-1}^{-Y}$, where Y > 0 is the transformation power, typically taken to be $p/2$, where p is the “effective number of dimensions in the data”. This effective dimension can be found by using Functional Principal Components Analysis. Conveniently, there is an “fpca” function within the “funcy” package, so we can just use that. The number of PCs resulting is just equal to the number of eigenvalues returned, which can be calculated as above using length.

4. Results

4.1 Finding

With the combination of “funcit” function and “Jump” method, we find that the optimal number of clusters is 4 for both leagues. The “Jump” method results for finding the optimal number of k in each league are shown in Figure 1 and plots of clustering for selected clubs in the top 2 divisions of the Bundesliga and the La Liga are shown in Figure 2.

4.1.1 “Jump” method result

In Figure 1, since the “Jump” method basically looks at differences in WSS between consecutive numbers of clusters, and then exponentiates that difference, we look for the point associated with the greatest “Jump” value. According to the “Jump” method, the optimal number of cluster is 4 for both the Bundesliga and the La Liga.

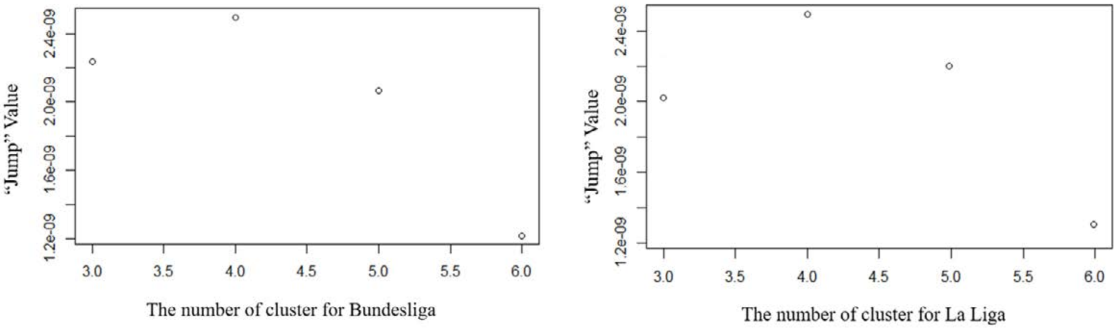
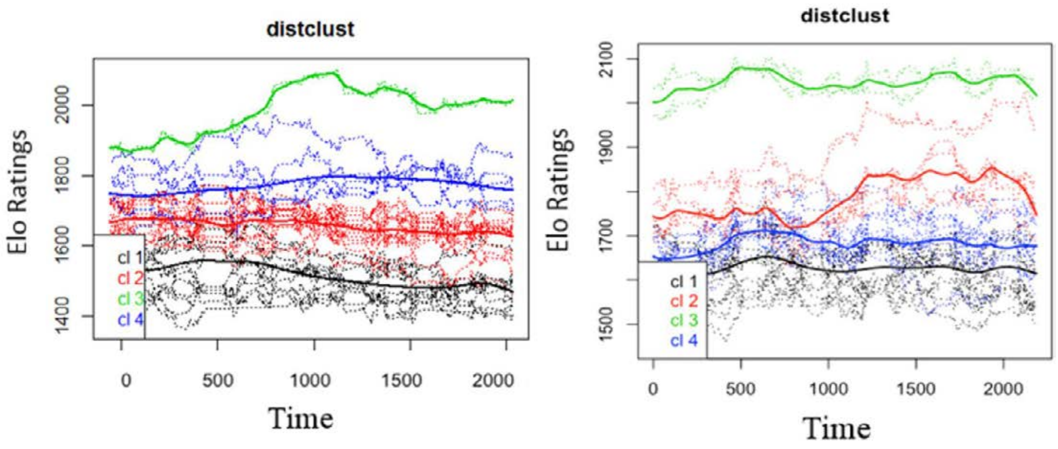


Figure 1: The optimal number of cluster is 4 for both leagues based on “Jump” method

4.1.2 Plots of clustering for teams in the top 2 divisions of the Bundesliga and La Liga

In Figure 2, each line represents a cluster. The top cluster (in green) in these 2 plots have the highest Elo ratings over the most recent 5 seasons. For the Bundesliga league on the left, Bayern Munich occupies this top cluster. For the La Liga league on the right, the top cluster is taken up by Barcelona and Real Madrid together. Such results correspond to our expectation that the Bundesliga and La Liga are dominated by a few clubs.



Bundesliga

La Liga

Figure 2: Plots of clustering for teams in the top 2 divisions of the Bundesliga and the La Liga over the most recent 5 years

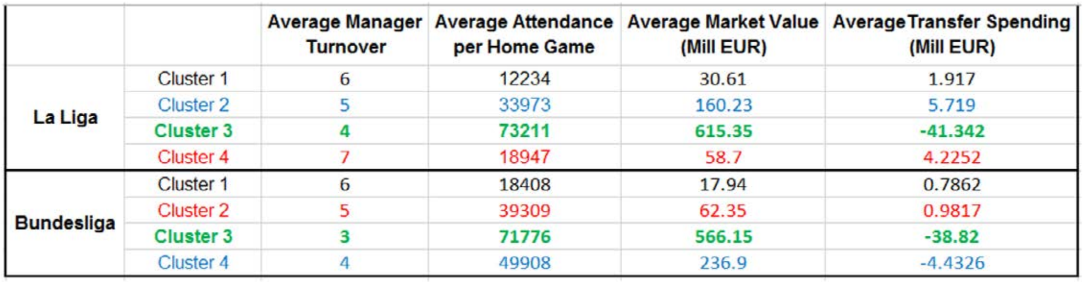
5. Discussion

5.1 Summary of Cluster Characteristics

After getting the clustering result, we explore characteristics that might explain such clusters assignment and find it from four aspects: Average manager turnover, Average attendance per home game, Average market values (in Mill Euros) and Average Transfer Spending (in Mill Euros). We highlight the top cluster by green, second cluster by blue, third cluster by red and the bottom cluster by black in order to correspond to the color of plots of clustering above. The summary table is shown in Figure 3.

In Figure 3, we can see that the top cluster (highlighted in green) in 2 leagues have the least amount of average manager turnover, highest average attendance, highest average market value, and most money spent on transfers; And the bottom cluster (highlighted in black) in 2 leagues have the highest average manager turnover for Bundesliga, second highest for La Liga, Lowest average attendance, lowest average market value, and lowest net transfer spending. Notice that the bottom cluster in the La Liga does not have the highest average manager turnover as it supposed to, so we suspect other factors involved or there exists some combination effect of these 4 aspects we analyze.

Due to the color arrangement in the plot, cluster 4 in Bundesliga corresponds to cluster 2 in La Liga and similarly, cluster 2 in Bundesliga corresponds to cluster 4 in La Liga, so that is why we highlight our summary table as the way it is.



		Average Manager Turnover	Average Attendance per Home Game	Average Market Value (Mill EUR)	Average Transfer Spending (Mill EUR)
La Liga	Cluster 1	6	12234	30.61	1.917
	Cluster 2	5	33973	160.23	5.719
	Cluster 3	**4**	**73211**	**615.35**	**-41.342**
	Cluster 4	7	18947	58.7	4.2252
Bundesliga	Cluster 1	6	18408	17.94	0.7862
	Cluster 2	5	39309	62.35	0.9817
	Cluster 3	**3**	**71776**	**566.15**	**-38.82**
	Cluster 4	4	49908	236.9	-4.4326

Figure 3: Summary of cluster characteristics of La Liga and Bundesliga over the most recent 5 years

6. Conclusion

So far, no competitive balance studies that we have seen have used a functional data approach. Treating the data as functional data allows us to group teams based on their performance over the most recent 5 seasons rather than just one year. As expected, the La Liga and the Bundesliga leagues are dominated by a few clubs, which characterized by more spending, higher attendance, higher market value and bigger transfers. Bottom cluster teams shuffle in and out of relegation. This provides useful evidence that both the La Liga and the Bundesliga soccer system are not entirely characterized by competitive balance, for better or for worse, therefore, these two professional soccer leagues should review their structure and format in order to make it more competitively balanced.

References

[1] Schiefler, L. Football Club Elo Ratings. Retrieved from http://clubelo.com/

[2] Sugar, C., & James, G. (2003). Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach. Journal of the American Statistical Association, 98(463), 750-763.

[3] Yassouridis, C. (2017). funcy: Functional Clustering Algorithm. R package version 0. 8.6. Retrieved from https://CRAN.R-project.org/package=funcy