Annals of Operations Research (2023) 325:85–113
https://doi.org/10.1007/s10479-022-04609-3

ORIGINAL RESEARCH

A rank-size approach to analyse soccer competitions and teams: the case of the Italian football league “Serie A"

Valerio Ficcadenti[2] · Roy Cerqueti[1,2,4] · Ciro Hosseini Varde'i[3]

Accepted: 14 February 2022 / Published online: 15 March 2022

Abstract
In this paper, we present a data-analysis rank-size approach to assess the features of soccer competitions and competitors. We investigate the championships rankings and the teams' final scores in the most relevant Italian league, the “Serie A”, between 1930 and 2020. We use the final rankings and the teams' scores to explore the presence of rank-size regimes in the various yearly championships. Besides, we analyse the teams one by one, ranking their performance over the years and using the rank-size law's parameters to compare their performances across the tournaments. We chose to do so via the Discrete Generalised Beta Distribution, a three-parameter rank-size function. We offer a cluster analysis of the rank-size law parameters based on a *k*-means algorithm to provide additional insights and capture similarities and deviations among championships and teams. Concluding, we propose a measure of competitiveness within championships and per team. The best fit results are statistically outstanding, and the cluster analysis presents two main clusters capturing teams' performances and years in which they have competed in the “Serie A”. The competitiveness analysis shows that the teams at the bottom of the championships ranking have obtained decreasing scores in recent years.

Keywords Italian football league · Rank-size analysis · K-means clustering · Competitiveness indicator

✉ Roy Cerqueti
roy.cerqueti@uniroma1.it

Valerio Ficcadenti
ficcadv2@lsbu.ac.uk

Ciro Hosseini Varde'i
chosseinivardei@gmail.com

1 Department of Social and Economic Sciences, Sapienza University of Rome, Piazzale Aldo Moro, 5, 00185 Rome, Italy

2 Business School, London South Bank University, Borough Road, 103, SE1 0AA, London, UK

3 School of Health and Sport Science, University of Urbino Carlo Bo, Via Aurelio Saffi, 2, 61029 Urbino, Italy

4 GRANEM, University of Angers, Rue de Rennes, 40, Angers 49100, France

1 Introduction

Soccer is undoubtedly one of the most popular sports competitions. In several regional contexts, it presents not only entertainment features but also socio-economic implications (see, e.g. Galariotis et al., 2018 for joint treatment of sport, business and financial performance of soccer teams, or the older contribution Neale, 1964). Such a relevance justifies the increasing data scientists attention and efforts in analysing and creating data sets related to soccer. A few examples are here discussed. In Hughes et al. (2012), the authors identify data-driven performance indicators for the different roles taken by the soccer players. An interesting review of the promising contributions of data science in the field is reported in Rein and Memmert (2016), where the complexity of the data management phases is also acknowledged. In the conclusive remarks of such a paper, one reads that "future soccer research will have to embrace a stronger multi-disciplinary approach". The suggestion surely includes the idea of collecting and analysing data about players (e.g., in-game performances). In this respect, Filetti et al. (2017)'s authors monitor the technical–tactical and physical efficiency of Italian "Serie A" season 2013–2014's players through a semi-automatic video analysis system. In Memmert et al. (2017), one can find an overview of the developments in the analysis of players' positional data in 11 matches between Bayern Munich and FC Barcelona.
Interestingly, the recent systematic literature review presented in Goes et al. (2021) outlines challenges and successes of big data potentialities to support tactical performance. In this respect, some contributions propose a data science-based analysis of the constellation of the position of the ball and the players during a match for guessing the probability of scoring a goal (see, e.g. Link et al., 2016; Ric et al., 2017). An analogous perspective is presented in Gonçalves et al. (2017), where one can find the complex networks-based analysis of the interactions among the players during a match. In a very different context, Frick et al. (2010) analyse the probability of the so-called "premature termination of a contract" of teams' head coach by employing a logit approach. In the context of the methodological tools used for dealing with data complexity, we mention Hassan et al. (2020), where the authors use neural networks to predict matches' outcomes.

This paper enters the debate on data science for the scientific exploration of football by providing a detailed analysis of the Italian Football Championships, which is undoubtedly one of the most relevant football contexts in the EU under different perspectives. For example, regarding the economic relevance, we quote Kennedy and Kennedy (2012), where the authors wrote, "The first decade of the twenty-first century has witnessed market growth from €8 billion per annum in revenue to almost €16 billion per annum, powered by the so-called 'big five' leagues in England, Germany, Spain, Italy and France". Furthermore, by looking at the attendance data in Worldfootball.net (2021b), one notice that Italy had 9,590,166 spectators in 2018/2019, scoring a place in the top ten's leagues in the world.

In line with Goossens et al. (2012)—where the authors analyse and compare different regulations for the Belgian league championships – we here compare the official historical rankings of the Italian championships between 1930 and 2020 (by removing the possible presence of penalties, as we will see below) with the cases of two-points score and three-points score for the winning team in a match. In this respect, we enter the debate raised by Csató (2020), where the author deals with the rules of the UEFA Champion League by advancing doubts on their fairness. We also mention Cea et al. (2020), where the authors discuss the procedure employed by FIFA for ranking the national team. In this context, they present proper modifications of such a procedure for overcoming some inconsistencies of the ranking rules.

In entering the debate by means of a rank-size analysis, we are close to some relevant contributions in the literature. For example, Ausloos (2014a)'s author assesses the Union of European Football Associations (UEFA) affiliated teams performance between 2009 and 2014 with—among others—the Discrete Generalised Beta Distribution, a three parameters ranks-size law (see Mansilla et al., 2007). Specifically, Ausloos studies the yearly ranking of the teams as a function of their UEFA coefficient. Again on the UEFA data, we also mention Ausloos et al. (2014b) in which the authors discuss the possibility of finding dissipative structures, as in open systems acquiring (and losing) energy. In Ausloos et al. (2014a), the authors present the rank-size relationships for the International Federation of Association Football (FIFA) and UEFA rankings to assess ranking differences in terms of the FIFA's and UEFA's coefficients. More recently, Yoon and Sedaghat (2020)'s authors use the rank-size law to fit the ranked attendance data for the games in Major League Baseball (MLB); National Baseball Association (NBA), National Football League (NFL), and National Hockey League (NHL). Differently, in Malacarne and Mendes (2000), the authors state that "the goal distribution by goal-players is connected with an anomalous decay related to the Zipf-Mandelbrot law" in the main football league from Italy, England, Spain and Brazil and for the specific case of a couple of championships. The informative content of the rank-size laws is well illustrated in Rimmer and Johnston (1967), where a visual inspection of the rank-size relationship between cities' population and their ranks in Australia is used to demonstrate the Victorian Football League's influence beyond the boundaries of the state (see Fig. 2 in the quoted paper).

This paper investigates the teams and the championships in the Italian league by assessing the relationship between teams' final ranks and scores. To this aim, we analyse all the Italian Football "Serie A" championships between 1930-2020. We use the final rankings and the teams' scores to assess the rank-size regimes at the yearly championships level. Moreover, we implement an individual teams-based analysis by ranking their performance over the years and using the rank-size law's parameters to compare their performances across the tournaments. So, in a nutshell, we analyse the presence of comparable features by using rank-size law's parameters which have different interpretations when comparisons are made between championships or teams (from now on, in this paper, we refer to the analysis done per championship with the locution "by year", and to the one done per team by means of "by team"). In this respect, the rank-size analysis allows creating a unified system based on the disaggregated data, hence pointing to the global features of the relationship between the ranked data in the light of the related scores. We also propose a cluster analysis of the rank-size laws based on a k-means algorithm to provide additional insights and capture similarities and deviations among championships and teams. Finally, we follow the approach used in Ficcadenti and Cerqueti (2017) to propose a measure of competitiveness within championships and by the team.

In line with the mentioned literature, we tested different laws; first of all, the Zipf-Mandelbrot one (ZML hereafter) presented in Mandelbrot (1953, 1961), as a generalization of the Zipf's law Zipf (1949, 1935); then—and more satisfactorily, under a statistical perspective—we tested the Discrete Generalised Beta Distribution (DGBD hereafter). The range of applications where these laws play a crucial role is wide. Among the others, we mention Ficcadenti and Cerqueti (2017) where the ZML has been used to estimate the economic cost of earthquakes, Ficcadenti et al. (2019, 2020) for their investigation of rank-size relationships in corpora, Cerqueti and Ausloos (2015) where the Italian cities tax income distribution analysis is run through a rank-size approach, Dimitrova and Ausloos (2015), where the Bulgarian Urban system is studied across years using rank-size laws and Rotundo (2014); Ausloos (2013, 2014b) for their applications in the field of scientometrics. Referring

to sport, it is worth mentioning Ausloos (2020) where the rank-size law of the official gains for bicycle teams in Tour de France is studied.
In our case, the Discrete Generalised Beta Distribution presents outstanding best fit performances. The analysis points to connections of the estimated curves' parameters with the years, the rules in place for assigning points and teams' presence in the *"Serie A"*. Besides, the clusterization of these parameters via the famous k-means algorithm (see Jain et al., 1999, for a detailed description of this method in the framework of the clustering procedures) highlights regularities and deviations in the characteristics of championships and teams—on the basis of the interpretation of the calibrated parameters. Such findings point the attention also to stylized facts related to soccer and its surrounding socio-economic environments, such as the relationship between competitiveness and economic capacity of the teams involved in the "Serie A" championships. In this light, we introduce and also discuss a competitiveness indicator of the teams at the individual championship level by considering the relative relevance of the sizes at high and low ranks.

The rest of the paper is organized as follows. Section 2 contains a description of the considered datasets. Section 3 contains a description of the employed methodology, with the statement of the rank-size laws; moreover, the implementation of the cluster analysis is also presented, along with the aggregation leading to the measurement of the competitiveness at low and high ranks. Section 4 presents the findings of the study critically, with a related discussion. Section 5 offers some conclusive remarks. We have relegated long tables on the explored datasets in the "Appendix".

2 The datasets

This study puts together information regarding Italian football results collected from Worldfootball.net (2021a). Specifically, we download the final scores of the top Italian football leagues, whose winner is awarded the *"Scudetto"*. Namely, we have downloaded the final points obtained by each team competing in the so-called *"Serie A"*.

The dataset covers 88 championships, namely all those played between 1930 and 2020 (we indicate the year in which the challenge ends). Table 4 (in the appendix) contains a statistical summary of the final points scored per each championship. It is interesting to notice that the number of admitted teams to the "Serie A" has changed over the years. The season-ending in 1930 was the first to adopt the single group formula, with home and away matches. This regulation has not undergone any change except for the number of teams, mainly held with 16, 18, 20 or 21 teams. The number of teams presents in the league changed :

- From 1930 till 1934, 18 teams
- From 1935 till 1943, 16 teams
- 1947, 20 teams
- 1948, 21 teams
- From 1949 till 1952, 20 teams
- From 1953 till 1967, 18 teams
- From 1968 till 1988, 16 teams
- From 1989 till 2004, 18 teams
- From 2005 till now 20 teams

In Table 5 (in the appendix) one finds a statistical summary of the scores obtained by each team at the end of each championship in which games have been played.[1]

A few interventions on the raw data are needed to homogenise information across the years. In fact, on the website, the final results are reported in tables that contain the ranking of the teams at the end of the tournament, the teams' names, the score (number of points) obtained by each team, the number of played, won, drawn and lost matches, scored and conceded goals along with their difference. Sometimes, score reductions are imposed by law; they penalise teams for specific illegal behaviours of managers and players. We create an additional column to incorporate penalisations and deduct the penalisation points from the reported scores. Some cases have been treated carefully in that penalisations might be responsible for relevant biases on the final ranking of the championships. For example, in 2006, *Juventus* sits at the last position, even if it has the highest score in the championship; such a severe penalisation is a consequence of a legal decision after the scandal *Calciopoli*, Commissione d'Appello Federale—Federazione Italiana Giuoco Calcio (2006). In this circumstance, we set the final score to zero but incorporate such information in the penalties' column, where the official score is saved and considered as the magnitude of the sanction. So, for the case of *Juventus* in 2006, we discount a penalty of -91, which was the whole scored level without penalisation.

A methodological note is now needed. For each championship, each team's final score is given by the sum of the number of the drawn matches and h times the number of the won matches. The value of h was 2 before 1994; after this year, due to a change of regulation, h was set to 3—as it is currently.

The analysis described in the next section is applied to three instances here identified with Pt, Pt_2 and Pt_3:

- Pt is the case where the points scored by the teams in a year (Y) incorporate penalties and/or relegation deriving from legal disputes outcomes (as per the description above).
- Pt_2 regards the analysis made on the ranking that the teams would have reached if the points assigned for winning a match were still 2, as it was before 1994. Namely, the points are assigned using the old rule for all the tournaments after the one finished in 1994.
- Pt_3 regards the analysis on the ranking that the teams would have reached if the points assigned for winning a match had been 3, as it currently is. So, all the points obtained before 1994 are assigned using the current rule.

For an overview of the three considered instances at a championship and team level, see Tables 4 and 5.

3 Methodology

We run rank-size investigations on the Italian football datasets Pt, Pt_2 and Pt_3 described in the previous section. First, we provide the analysis at the championship level, namely an analysis of the tournaments disputed between 1930 and 2020. Per championship, the final

[1] This is a unique case in Italian history with an odd number of teams. The situation was born from a peculiar case generated by World War II consequences and the political dispute with Yugoslavia regarding the territories of Trieste and Istria. In the championship that ended in 1947, Triestina has scored the least position but has played in odd conditions. For example, the home stadium was not utilizable due to the political situation. In the tournament ending 1948, FIGC admitted Triestina in "Serie A" for sporting merit despite its previous performance. During that season, Triestina has scored 49 points, as much as Juventus and Milan, close to the "Grande Torino" of those years, who was winning the third "Scudetto" in a row and forth at that time.

points obtained by the teams are ranked in descending order and, the highest outcome is associated with the rank one, while the lowest one has the largest rank, corresponding to the number of teams competing in that specific championship. The case of the tournament ending in 1937 (instance Pt) is reported in Fig. 1a as an illustrative example. At this stage one should focus on the black dots only; in that year Bologna has won the "Scudetto" and Alessandria was the last in the final ranking. On the x and y-axis, the final ranks and the total points scored are respectively reported. Second, we analyse the datasets at the team level by taking the teams that have competed in the *"Serie A"* for a large enough number of the considered championships. To this aim, each team's score obtained in each tournament is ranked in descending order. The highest result obtained among the championships disputed by the considered team has a rank equal to one. In contrast, the team's lowest score is associated with the highest rank, corresponding to the number of tournaments disputed between 1930 and 2020. Specific teams have participated in a few championships over the years; therefore, they do not have enough observations to make a robust rank-size fit. Thus, as announced above, we set a threshold. Namely, the teams that have competed less than 16 times in the *"Serie A"* are dropped from the dataset, so the rank-size analysis is not performed on them. Such a threshold allows a meaningful implementation of the rank-size law's best fits. The case of the SPAL[2] (instance Pt) is presented in Fig. 1b as another illustrative example. SPAL has had its best performance in 2019 scoring 42 points and its worse in 2020 with 20 points. In the following, we refer to r as the team's rank whose score is the size z.

In line with the aforementioned literature, a first best-fit tentative is done with the ZML rank-size curve (Mandelbrot, 1953, 1961) here reported for reference:

$$z = \frac{\phi}{(\theta + r)^{\xi}}, \tag{1}$$

where ϕ, θ and ξ are the non-negative parameters to be calibrated. However, the best-fit exercise through Eq. (1) does not provide statistically sounding results, that are then not shown[3]

For all the considered datasets, scores and related ranks are much better represented by the DGBD (see, e.g., Naumis and Cocho, 2008; Martínez-Mekler et al., 2009), which exhibits outstanding capacity of fitting the championships outcomes, with high values of the goodness-of-fit parameters. The formulation of such a law is

$$z = \frac{\alpha(R + 1 - r)^{\beta}}{r^{\gamma}}, \tag{2}$$

where α, β and γ are the non-negative parameters to be calibrated, and $R = \max(r)$ over the considered data sample.

The application of Eq. (2) and the mentioned data pre-processing lead to a dataset made of 28 teams[4] (see Table 5 in the appendix).

The DGBD-based analysis at championship and team levels generate two sets of triplets representing the estimated α, β and γ, i.e. $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$, via the Trust Region Reflective algorithm (Branch et al., 1999) applied with boundaries conditions on the parameters, that are

[2] This team is based in Ferrara, Emilia-Romagna.

[3] The elaborations of the rank-size analysis by using the ZML in Eq. (1) are available upon request.

[4] The set of teams that competed in more than 16 tournaments between 1930 and 2020 is made of 28 clubs, i.e. 'Ascoli', 'Atalanta', 'Bari', 'Bologna', 'Brescia', 'Cagliari', 'Catania', 'Chievo', 'Fiorentina', 'Genoa', 'Inter', 'Juventus', 'Lanerossi Vicenza', 'Lazio', 'Lecce', 'Livorno', 'Milan', 'Napoli', 'Padova', 'Palermo', 'Parma', 'Roma', 'SPAL', 'Sampdoria', 'Torino', 'Triestina', 'Udinese', 'Verona'.

forced to be positive. Furthermore, a "brute-force" procedure is deployed to avoid local minima in estimating. A broad grid of starting values feeds the trust Region Reflective algorithm so that the $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ minimizing the Root Mean Square Error (RMSE) are taken as starting points for the final estimation run (see Ficcadenti and Cerqueti, 2017, where a similar process is used).

The two sets of $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ are stored in two tables, one for the estimations when year by year results are considered and another for the team by team analysis. Each of these reports the three instances Pt, Pt_2 and Pt_3.

As a further investigation and for all the considered instances, we cluster championships and teams through the obtained triplets. In particular, we employ a k-means clustering approach (see, e.g. Baker et al., 2020), which is undoubtedly one of the most popular clustering methods (for clustering, the calibrated DGBD parameters are suitably standardized). Through it, we want to identify two regimes in scored points distributions along with each championship and each team's history. Therefore, we are interested in having $k = 2$ to capture best and worst-performing teams in championships (e.g., first and last four teams for each championship), and teams' glorious and undistinguished moments characterised by the best and worse years in terms of scored points. Signs of two regimes are found in the distributions of the estimated parameters, as we point out in Sect. 4. To ensure that $k = 2$ is a suitable choice, we make more formal consideration calculating the common Silhouette index (D'Urso and Maharaj, 2012; Kaufman and Rousseeuw, 2009) on each clustering by varying k, and we report the values in Table 1. The results allow to conclude in favour of the choice $k = 2$ apart for the case Pt_3, when the analysis team by team is run. Only in that instance, $k = 3$ is favoured, but the difference is not material and the Silhouette value is still in an acceptable range for $k = 2$. Therefore, the employed clustering algorithm selects the two clusters' centroids

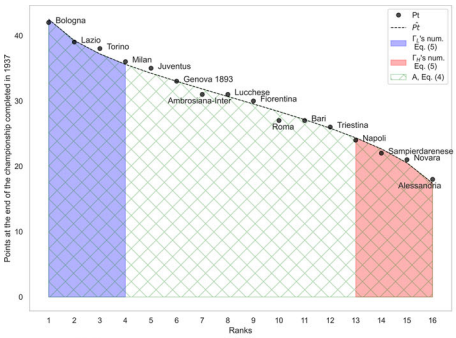
$$\mu_1 = (\alpha^{(1)}, \beta^{(1)}, \gamma^{(1)}), \qquad \mu_2 = (\alpha^{(2)}, \beta^{(2)}, \gamma^{(2)})$$

that minimise the within-clusters sum-of-squares criterion:

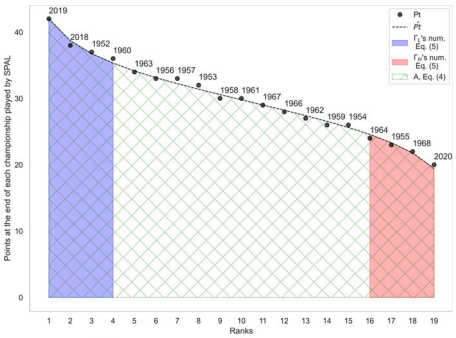
$$\sum_{i=1}^{n} \min_{\mu_j \in \mathbb{R}^3 : j=\{1,2\}} (||x_i - \mu_j||^2) \tag{3}$$

where $x_i = (\hat{\alpha}_i, \hat{\beta}_i, \hat{\gamma}_i,)$ is the triple of the estimated parameters α, β, γ in Eq. (2) for the i^{th} element of the considered sample, n is the cardinality of the sample and $||x - \mu||$ is the Euclidean distance between the three-dimensional vectors x and μ. More specifically, we use *"k-means++"* from the Python's Scikit-learn (Pedregosa et al., 2011) for its capacity in minimizing the chances of getting into local minima through the optimization procedure, see Arthur and Vassilvitskii (2006). We here report a summary of the algorithm that can be summarised in three steps. The first one is to choose the initial μ_1 and μ_2 (this is the point where the starting centroid selection process has been changed with the *"k-means++"* variant, see Ostrovsky et al., 2013). After that, the algorithm consists of a loop between the next two steps. The second step assigns each team to a cluster using the criteria of the nearest centroid, while in the third one, new centroids are calculated averaging $\hat{\alpha}_i$, $\hat{\beta}_i$, $\hat{\gamma}_i$ belonging to the same clusters. Then, the difference between the old and the new centroids are computed, and the algorithm repeats these last two steps until the new centroids and the old ones do not vary significantly. A summary of the process is reported in the Algorithm 1.

Additionally, we use the rank-size relationship reported in Eq. (2) to transform the points into an indication of competitiveness within championships and per team. For all the analysed instances, we denote with A the area underlying the curve in Eq. (2) and bounded from below



(a) Analysis *by year* referred to the championship finished in 1937.



(a) Analysis *by team* referred to the championships in which SPAL has played.

Fig. 1 In the figure are reported the rank-size's best fits run with Eq. (2). Furthermore, we report the elements necessary for the calculation of the competitive indexes Γ_H, Γ_L reported in eqs. (4) and (5)

Table 1 The Silhouette index is calculated for each k-means cluster analysis run. The parameter k goes from 2 to 5, we report it for the analysis "by year" and "by team" in each instance Pt, Pt_2 and Pt_3

N. clusters (k)	2	3	4	5
By year				
Pt	0.4353	0.4171	0.4090	0.3571
Pt_2	0.4781	0.4263	0.3489	0.3549
Pt_3	0.4941	0.4284	0.4073	0.3899
By team				
Pt	0.3946	0.3708	0.3606	0.3619
Pt_2	0.4494	0.3764	0.4188	0.4371
Pt_3	0.3966	0.4283	0.4135	0.4219

The parameter k goes from 2 to 5, we report it for the analysis "by year" and "by team" in each instance Pt, Pt_2 and Pt_3

by the abscissae, i.e.:

$$A = \int_{\min(r)}^{R} \hat{\alpha} \frac{(R+1-r)^{\hat{\beta}}}{r^{\hat{\gamma}}} dr \tag{4}$$

Moreover, we denote by Γ_L and Γ_H the proportion of the area of A given by the four lowest ranks and highest ranks, respectively, i.e.

$$\Gamma_L = \frac{\int_1^4 \hat{\alpha} \frac{(R+1-r)^{\hat{\beta}}}{r^{\hat{\gamma}}} dr}{A}, \qquad \Gamma_H = \frac{\int_{R-4}^{R} \hat{\alpha} \frac{(R+1-r)^{\hat{\beta}}}{r^{\hat{\gamma}}} dr}{A}. \tag{5}$$

The values of Γ_L and Γ_H are calculated for Pt, Pt_2 and Pt_3 in both the cases "by year" and "by team", providing a view of the competitiveness over the years and providing a relative measure of the capacity per each team. An exemplifying view of these indicators is reported in Fig. 1. For both the type of analysis, the areas are highlighted when the instance Pt is considered. Figure 1 shows the cases of the championship ended in 1937, and Fig. 1b gives a visual idea of the situation for the SPAL. One can inspect the colours of the areas to better understand the idea behind eqs. (4) and (5).

Table 2 Summary of the best fit results through Eq. (2) run per year

Type	Elements	Max	Min	μ	m	σ	Skew	Kurt
Pt	$\hat{\alpha}$	90.2132	7.3581	32.7264	30.1607	12.3471	1.4198	4.7461
	$\hat{\beta}$	0.6190	0.0014	0.2170	0.1966	0.1112	0.7303	1.0375
	$\hat{\gamma}$	0.2863	0.0000	0.1465	0.1465	0.0593	− 0.0827	0.0763
	R^2	0.9920	0.7941	0.9539	0.9657	0.0348	− 2.3340	6.6511
	RSME	5.7150	0.6654	2.2457	1.9747	1.0094	1.2269	1.5330
Pt_2	$\hat{\alpha}$	59.9453	12.8027	30.8773	29.0569	9.1704	0.6012	0.2496
	$\hat{\beta}$	0.4199	0.0149	0.1936	0.1791	0.0900	0.3431	− 0.3257
	$\hat{\gamma}$	0.2554	0.0000	0.1463	0.1465	0.0530	− 0.2045	− 0.1523
	R^2	0.9920	0.8453	0.9603	0.9671	0.0246	− 2.0959	6.5337
	RSME	3.1359	0.6654	1.7689	1.7559	0.5162	0.4421	0.2362
Pt_3	$\hat{\alpha}$	90.2132	15.5268	41.4253	38.3956	13.9737	0.7826	0.8434
	$\hat{\beta}$	0.4680	0.0014	0.2174	0.2007	0.1004	0.3038	− 0.3193
	$\hat{\gamma}$	0.3156	0.0000	0.1661	0.1625	0.0598	0.0492	0.0560
	R^2	0.9939	0.8822	0.9630	0.9682	0.0207	− 1.2471	2.0097
	RSME	4.6937	1.0259	2.6103	2.5893	0.7989	0.3412	− 0.2240

Regarding the notation, μ is the mean, m is the median, σ is the standard deviation, Skew. and Kurt. are skewness and kurtosis, respectively, R^2 is the coefficient of determination and RMSE stands for Root Mean Square Error

4 Results and discussion

The rank-size analysis results from the Eq. (2) best fits lead to the identification of the calibrated parameters $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$, along with some goodness-of-fit quantities—R^2 and RMSE in our context. The calibrated parameters, the goodness-of-fit measures, and the considered sample's cardinality form two sets of distributions—one associated with the championships and the other with the teams. The main descriptive statistics of such distributions are reported in Tables 2 and 3, and Fig. 2. The locutions "by year" and "by team" reported in that figure are respectively associated with the analysis run at year (championship) and team level.

The parameter α may be viewed as a proxy of the size at rank one so that a large (small) value of α is associated with a large (small) value of the maximum score of the considered sample.

The parameter β describes how the curve decreases as the rank grows. If β is small, then the curve tends to capture the so-called *"queen and harem effect"* turning the concavity of the curve at the highest ranks (Ausloos, 2013, 2014b).

The DGBD's γ captures the deviation of the size of two consecutive ranks. Specifically, the difference between the sizes at rank r and $r + 1$ decreases as the value of γ grows. One can better visualize what is described here observing Fig. 3 and jointly comparing the curves with the parameters (reported in caption) that generated them, for this exercise, one can disregard the clusters. Besides, the relationship between parameters can be visually inspected in Fig. 4.

At a championship level, one can notice that the distribution of $\hat{\alpha}$ is skewed on higher values in the cases Pt and Pt_3. This outcome highlights similarities between actual final rankings (Pt) and those resulting from the assignment of 3 points to won matches, with an asymmetry of the distribution of the maximum scores in the championships to its left tail. The positive

Table 3 Summary of the best fit results through Eq. (2) run per team

Type	Elements	Max	Min	μ	m	σ	Skew	Kurt
Pt	N.Obs.	87.0000	16.0000	45.1071	35.0000	26.6143	0.3718	− 1.6098
	$\hat{\alpha}$	50.2006	14.5640	30.6996	27.8887	9.5876	0.6247	− 0.2980
	$\hat{\beta}$	0.3955	0.0688	0.2111	0.1989	0.0818	0.3385	− 0.4869
	$\hat{\gamma}$	0.2390	0.0183	0.1384	0.1473	0.0623	− 0.4290	− 0.6290
	R^2	0.9938	0.8706	0.9605	0.9690	0.0275	− 1.5966	3.1017
	RSME	4.6539	0.4933	2.1580	1.9006	1.1472	0.6559	− 0.4795
Pt_2	N.Obs.	87.0000	16.0000	45.1071	35.0000	26.6143	0.3718	− 1.6098
	$\hat{\alpha}$	46.4931	12.6832	27.7349	27.7562	8.2021	0.3036	− 0.0192
	$\hat{\beta}$	0.3306	0.0688	0.1693	0.1532	0.0649	0.9426	0.7916
	$\hat{\gamma}$	0.1709	0.0000	0.0907	0.0948	0.0481	− 0.4366	− 0.4176
	R^2	0.9902	0.9241	0.9700	0.9729	0.0163	− 1.4316	1.9759
	RSME	1.9400	0.6592	1.1695	1.1550	0.2873	0.5814	0.4573
Pt_3	N.Obs.	87.0000	16.0000	45.1071	35.0000	26.6143	0.3718	− 1.6098
	$\hat{\alpha}$	65.7221	15.8673	35.8591	34.7041	11.6096	0.5711	0.4818
	$\hat{\beta}$	0.3494	0.0843	0.1920	0.1794	0.0679	0.9064	0.6229
	$\hat{\gamma}$	0.1890	0.0000	0.0994	0.1015	0.0534	− 0.3949	− 0.5252
	R^2	0.9897	0.8976	0.9735	0.9768	0.0182	− 3.0210	11.3890
	RSME	2.8439	1.0526	1.6314	1.5624	0.4518	1.0948	0.9875

Regarding the notation, μ is the mean, m is the median, σ is the standard deviation, Skew. and Kurt. are skewness and kurtosis, respectively, R^2 is the coefficient of determination and RMSE stands for Root Mean Square Error

skewness is associated with substantial values in the final rankings of the championships. In this respect, please refer to $\hat{\alpha}$ maxima, which is around 90 for Pt and Pt_3 and around 60 in Pt_2. Of course, such large values are registered for teams winning many competitions and in the cases where the points assigned to the won matches is 3, namely championships disputed from 1994-95; for this reason, we do not observe the same result for Pt_2. Moreover, the skewness for Pt is much larger than that of Pt_3. This finding agrees with the evidence that the dataset Pt_3 contains generally larger values so that its highest realizations are closer to the mean. This is confirmed by the resulting means of Pt and Pt_3—around 32 and 41, respectively—and by the medians—about 30 and 38, respectively. The interpretation of $\hat{\alpha}$ is quite similar in the rank-size analysis at a team level, and the same arguments proposed above apply. We only notice that the discrepancies among Pt, Pt_2 and Pt_3 are less evident for the team case than for the championship one. This outcome explains the less evident 3 points regulation effects on the scores of a given team rather than on the rankings of the championships. In other words, some teams have spent most of the years in "Serie A" before 1995, for example, Padova, therefore the differences between Pt, Pt_2 and Pt_3 are mitigated. The meaningfulness of the estimated parameter $\hat{\gamma}$ is evident if one compares the values obtained for the analysis "by team" in the instance Pt with those for Pt_2 and Pt_3 (see tables 3, 5 and figs. 2, 4). In particular, the mean value of $\hat{\gamma}$ is higher for Pt than for Pt_2 and Pt_3—the means in the three cases are around 0.14, 0.09 and 0.1, respectively. This result describes a situation with homogeneous sizes with quite steep rank-size curves. The case "by year" behaviour remains stable across the instances Pt, Pt_2 and Pt_3 with a slightly higher

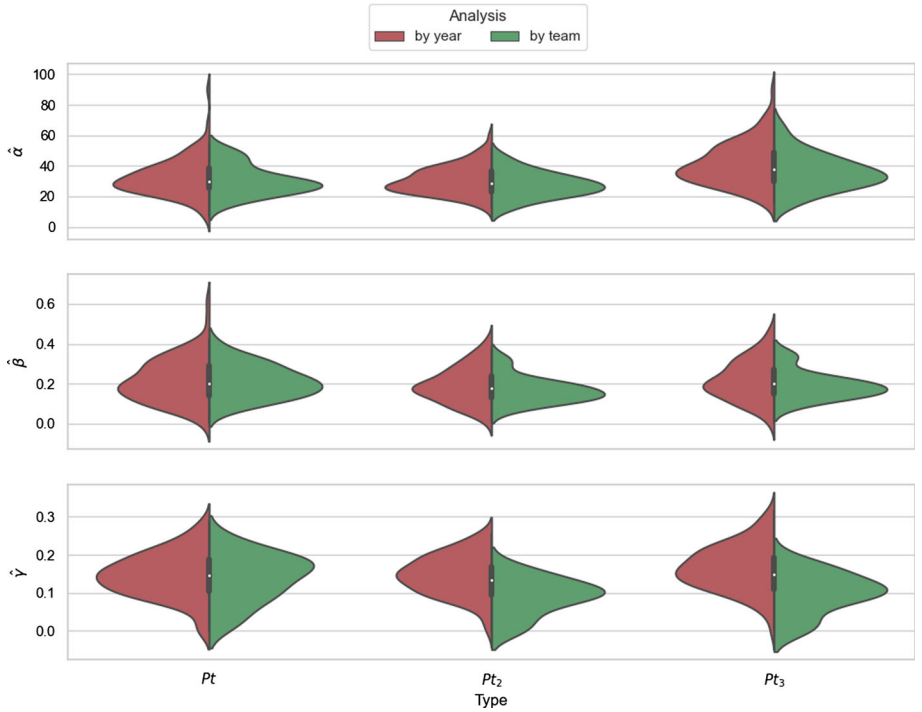
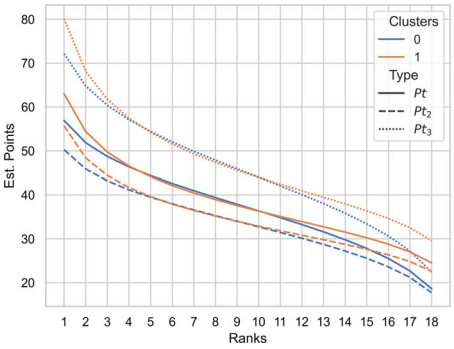


Fig. 2 The figure reports the DGBD estimated parameters' probability density (smoothed by a kernel density estimator) when one performs the best fits "by year" and "by team"

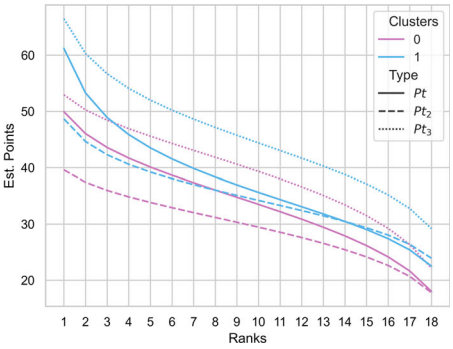
mean than those of the team cases, manifesting the presence in the championship analysis of a few more points on the tails of the rank-size law. Furthermore, most of the considered teams have played under both the rules of 2 and 3 points per won game. Therefore, the championships played during the latter periods led to higher scores than the former ones, forcing $\hat{\gamma}$'s to be more prominent in the Pt and smaller in Pt_2 and Pt_3—where the points decay is smoother along the ranks.

The values of $\hat{\beta}$ present similar behaviours for the cases Pt and Pt_3 when the analysis "by year" is performed. In contrast, when it is performed "by team", a stronger similarity can be observed between the cases Pt_2 and Pt_3 confirming the teams' performances regardless the rules in force. In both of cases, all the distributions of the $\hat{\beta}$ for Pt, Pt_2 and Pt_3 have positive skewness and present signs of bi-modality. So, if the points associated with low ranks are shallow and not that distanced, $\hat{\beta}$ tends to be smaller. For the cases of the analysis "by team", $\hat{\beta}$ captures the capacity of the teams in the "Serie A"—specifically those having played at least in 16 championships—to perform relatively well (with respect to them-self) for several years and rather bad in the majority of the competitions. Indeed, Fig. 5 shows that most of the analysed teams have had their "glory moments", being in the first five official positions of the competitions at least once and most of the teams spent their carriers in the central-low part of the rankings.

The analysis of the DGBD best fits' parameters proves the presence of two regimes across the different instances. Its strongest evidence consists of bi-modality in parameters' distributions which also strengthens the decision to use the k-means calibrated with two clusters, $k = 2$, further supported by the Silhouette index reported in Table 1. In figs. 6 and 7



(a) Cluster analysis *by year*. Notice that, the y-axis is the size estimation, while we set the variation range of the ranks between 1 and 18 – see x-axis – for a clear illustration of the final outputs



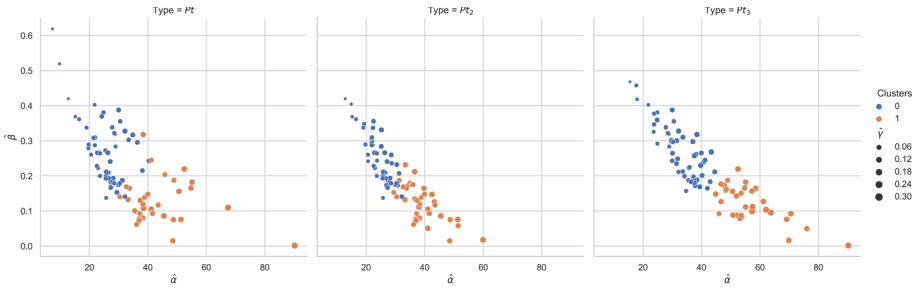
(b) Cluster analysis *by team*. Notice that, the y-axis is the size estimation, while we set the variation range of the ranks between 1 and 18 – see x-axis – for a clear illustration of the final outputs

Type	Clusters	$\bar{\hat{\alpha}}$	$\bar{\hat{\beta}}$	$\bar{\hat{\gamma}}$
Pt	0	25.6997	0.2750	0.1118
	1	43.8864	0.1250	0.2017
Pt_2	0	24.4065	0.2499	0.1111
	1	39.3915	0.1196	0.1927
Pt_3	0	32.8237	0.2724	0.1329
	1	56.4781	0.1210	0.2243

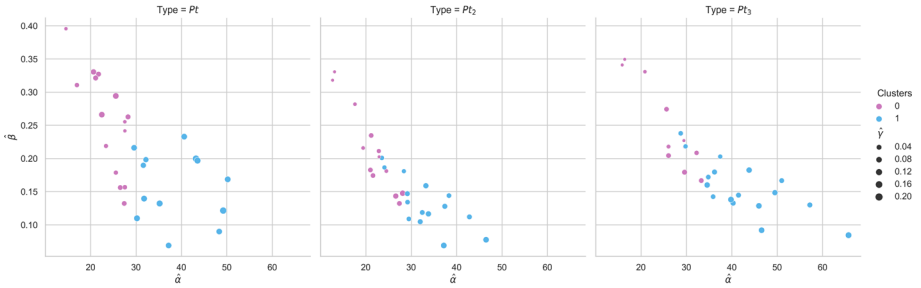
Type	Clusters	$\bar{\hat{\alpha}}$	$\bar{\hat{\beta}}$	$\bar{\hat{\gamma}}$
Pt	0	23.7836	0.2565	0.0960
	1	38.6797	0.1587	0.1873
Pt_2	0	21.4645	0.2119	0.0649
	1	33.1692	0.1324	0.1131
Pt_3	0	25.9234	0.2469	0.0545
	1	42.2880	0.1565	0.1283

Fig. 3 The clusters regimes are represented in terms of Eq. (2). Namely, the $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ belonging to the resulting clusters {0,1} have been averaged within the subdivisions to report two representative curves per each instance Pt, Pt_2 and Pt_3. The averaged parameters are indicated with $\bar{\hat{\alpha}}$, $\bar{\hat{\beta}}$ and $\bar{\hat{\gamma}}$

the results of the cluster analysis "by year" are reported. The different distributions of the observations across clusters and along the years manifest interesting aspects. Apart from the first nine years (first bin in Fig. 7) and the last years, the clusters capture the two regimes mentioned above. The years belonged to different groups until, more recently, they split more equally between the clusters; this is particularly evident in Pt. Potential justification for that can be found in Özaydin and Donduran (2019) where the authors describe how the teams' competitiveness grows together with the economic power. Similar arguments are detailed in Michie and Oughton (2004), where the authors state that "in Italy, there has been a marked deterioration in competitive balance since 1992 so that at the end of the period (2004) Italy had the highest degree of imbalance of the top 5 leagues" and connect this fact with the increase in revenues by the firms owning the soccer teams. Finally, in Nicoliello and Zampatti (2016) the impact of the Financial Fair Play regulation starting from 2014 is studied for the case of Italy. The need for a Financial Fair Play regulation constitute relevant proof of the competitive implications of economics or financial imbalances, see Masters (2014). In the instance Pt, the split years between clusters can be additionally affected by the changes in rules. For example, $\hat{\alpha}$ shows two picks (see Fig. 2), providing indications of the distinct behaviour of the winning teams, which got higher points for the more recent championships. Another way for understanding the results from the cluster analysis done "by year" comes from a visual inspection of Fig. 3a. It reports the curves plotted by plugging in Eq. (2) the estimated parameters' averages per cluster and instance. One notices that the differences are mostly present at high and low ranks, where the clusters are characterized. On the other



(a) Cluster analysis *by year*



(b) Cluster analysis *by team*

Fig. 4 The parameters resulting from the best fit of Eq. (2) are reported, and the colours distinguish them considering the clusters

hand, Fig. 4a, instance Pt, shows the signs of the change in rules with a less concentrated distribution of the dots with respect to the cases Pt_2 and Pt_3.

Concluding, in Figs. 8 and 9 we report the results from the analysis of the areas subtended by Eq. (2), namely the figures generated from Eq. (5). It is interesting to notice the presence of different regimes. There has been a drastic drop in the first years before 1950 for both Γ_H and Γ_L and then a recovery predominately obtained by the teams sitting at the first four positions of the championships until 1975-1980, see Fig. 8. The areas below the curve related to the last four positions in each championship did not recover enough to go back to the level occurred before the Second World War. In recent years, we can appreciate a stabilization for Γ_L, but, regarding Γ_H the situation has deteriorated, further strengthening the idea of having weaker and weaker teams at the bottom of the rankings.

The results from the analysis "by teams" are driven by similar factors and lead to conclusions regarding teams performances. In the analysis run on the official rankings Pt, the teams that had less success and have played more recently belong to the same cluster (see figs. 5, 10, 11). For example, one can notice that Brescia, Cagliari, Catania, Chievo and Lecce have a recent history in "Serie A" (Fig. 11) not that successful in terms of positions in the final ranks (Fig. 5) but still good for the points obtained (Fig. 10). On the other hand, teams like Fiorentina, Inter, Juventus, Napoli, Roma and Torino have a more successful history. Namely, they have got higher ranks (Fig. 5) and more points (Fig. 10). They have played more often in the "Serie A" (Fig. 11). The analysis of Pt_2 and Pt_3 are less affected by the time factor magnified by the official penalties applied and, more importantly, by the changes in rules. This is confirmed by the fact that changes in clusters are not that frequent when comparing the colours of the boxplots reported in Fig. 10, Pt_2 and Pt_3. Furthermore, these teams with

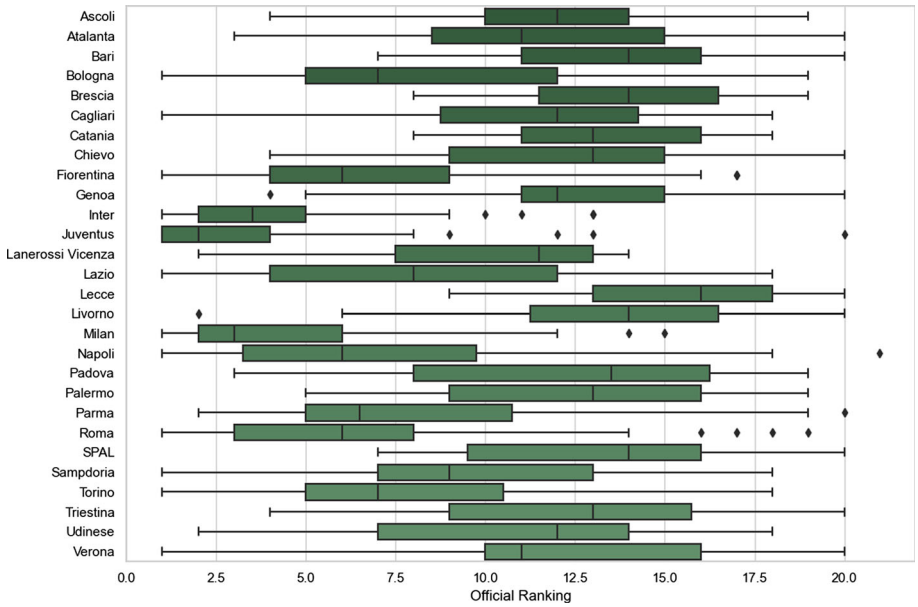


Fig. 5 This figure reports information about variability of the official ranking of each team analysed, namely those having played at least 16 "Serie A" championships. The coloured areas represent the observations distributed around the media (black vertical line), namely those comprised between the 25th ($Q1$) and 75th percentiles ($Q3$), the interquartile range ($IQR = Q3 - Q1$). The edges of the continuous lines respectively represent the data between $Q1 - 1.5 * IQR$ and $Q3 + 1.5 * IQR$. The rhombus represent the outliers positions

a more successful history in "Serie A" tend to have lower levels of Γ_L and Γ_H, as reported in Fig. 9. The presence of teams moving from one cluster to another when looking at Pt, Pt_2 and Pt_3 can be explained through three factors: (i) the change in rule for assigning points to won matches (from 2 points to 3 points for each victory), so that teams having played and won most of the matches recently (so getting 3 points per won match), are likely to stay in the same cluster for the instances Pt and Pt_3, see for example the case of Livorno. (ii) The proportion of draws had during the different championships, so in different periods. In other words, the teams having scored draws often have the rank-size curve more characterised by the few times they performed well. In this context, the points assigned to the won games—2 or 3—play a relevant role in identifying the clusters. This can be visualised by observing the different behaviours of the curves and clusters at low ranks in Fig. 3b, e.g., for the instances Pt and Pt_3. In this case, the distance between the two clusters at low ranks is more significant for Pt_3 than Pt; therefore, there is a remarkable difference in the probabilities of falling in one cluster or another. (iii) In Pt, one has the penalties included in the final results, hence leading to more extreme values (figs 10 and 12).

5 Conclusions

The present paper deals with a rank-size analysis of the Italian football final ranking obtained by teams in the so-called "Serie A" in the championships disputed between 1930 and 2020. The parameters' calibration procedure run with the DGBD, Eq. (2), presents an outstanding

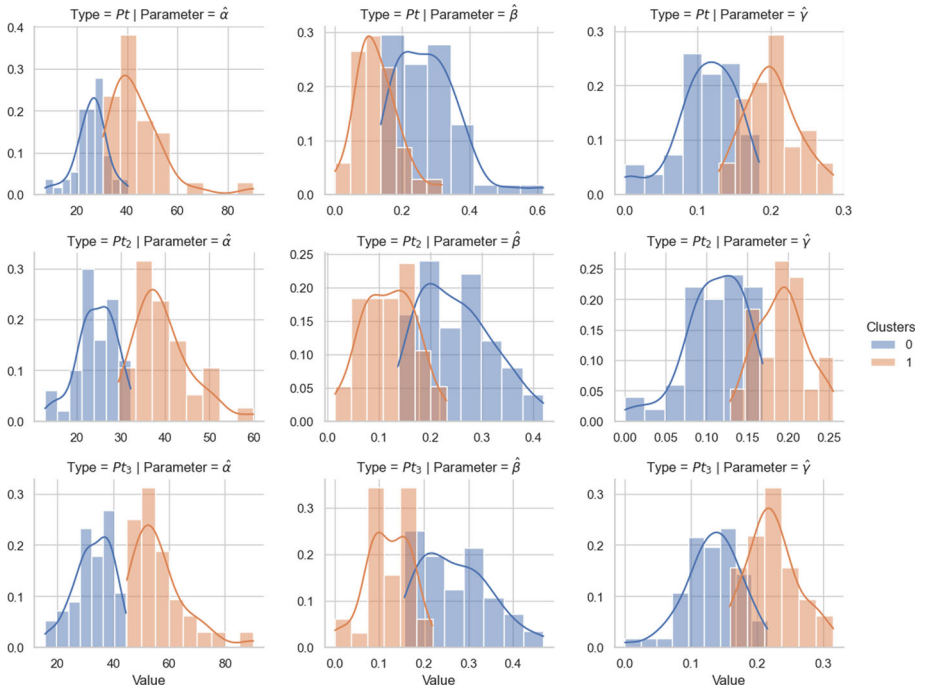


Fig. 6 Empirical distribution of the DGBD's estimated parameters divided by clusters when the analysis "by year" is run

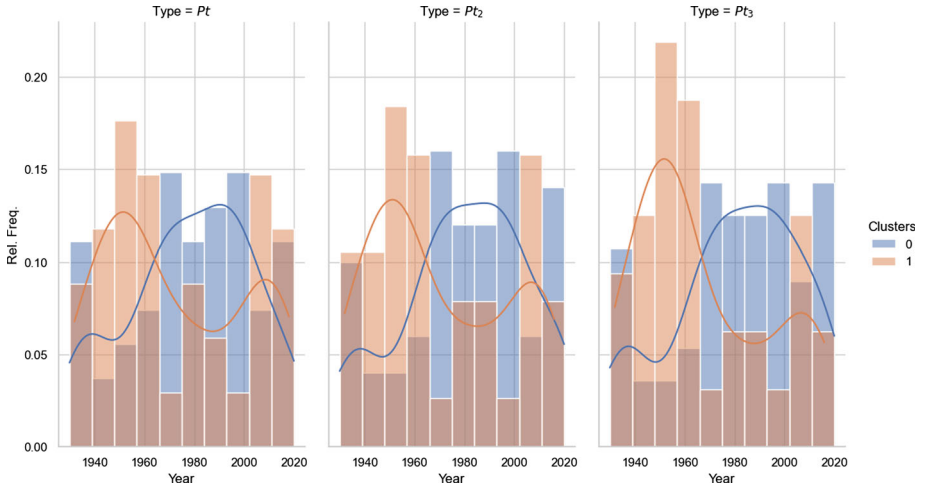


Fig. 7 Empirical distributions of the years belonging to the two clusters. Each bin covers nine years, starting from 1930 and ending in 2020

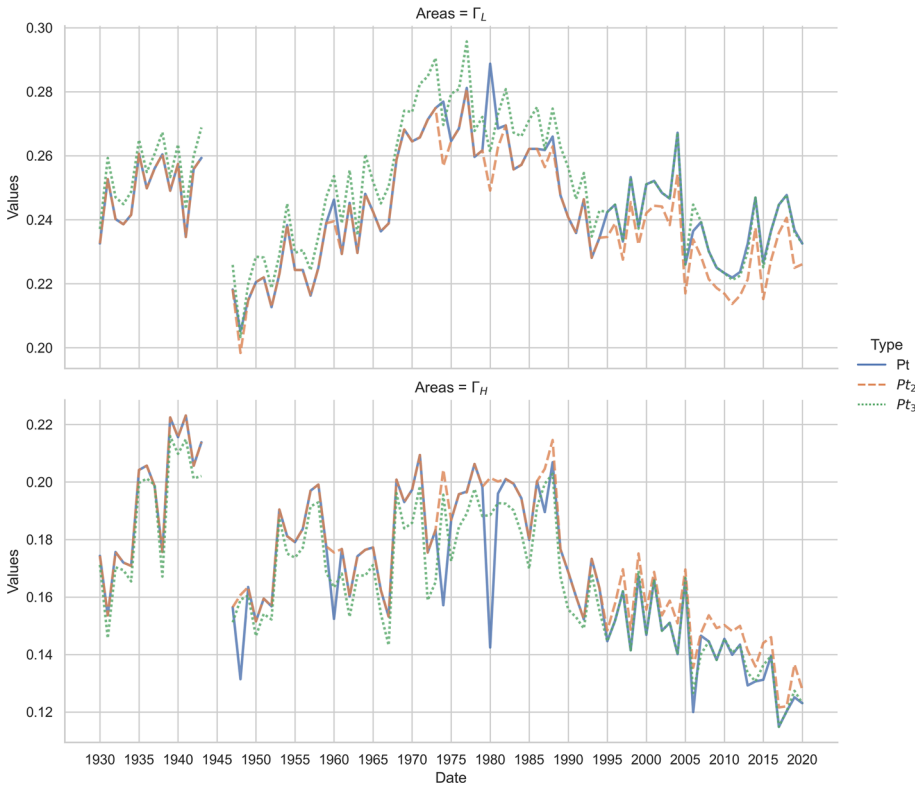


Fig. 8 Time series of the areas subtended by the curve in Eq. (2) and captured with Eq. (5) using the parameters estimated in the analysis "by year". The tournaments were not disputed between 1944 and 1946

performance with a R^2 ranging between 0.80 and 0.99. The DGBD's parameters α, β, and γ effectively capture relevant features related to the teams' performance. The highest points scored by the competitors in the championships, in the case "by year", and in the teams' history for the case "by team", are captured by $\hat{\alpha}$. The performance of the teams in the play-off area is captured by $\hat{\beta}$ and the decay as well as the concentration at highest rank is captured by $\hat{\gamma}$. The triplets estimated in the instances Pt, Pt_2 and Pt_3 successfully represent the conditions of teams and championships had over the years. The rank-size regimes and their meaning are evident once used to feed the k-means algorithm with $k = 2$. The results prove the relevance of historical phases of Italian football, suggesting a solid characterisation of them by the economic condition of the teams and the rules in place. More specifically, we consider the results connected with the literature regarding the teams' economic power and competitiveness (see Michie and Oughton, 2004; Nicoliello and Zampatti, 2016; Özaydin and Donduran, 2019), and the change in rules that occurred in 1994-95. In details, the connection with the deterioration in competitive balance since the '90s can be associated with the change of the European Champion Clubs' Cup tournament formula, occurred in 1992. The cup was initially a straight knockout tournament open only to the champions of Europe's domestic leagues. The competition got its current name "UEFA Champions League" during the season ended in 1993, after that other changes occurred. For example, it was added a round-robin group stage, and later it was allowed multiple entrants from other countries than the whole

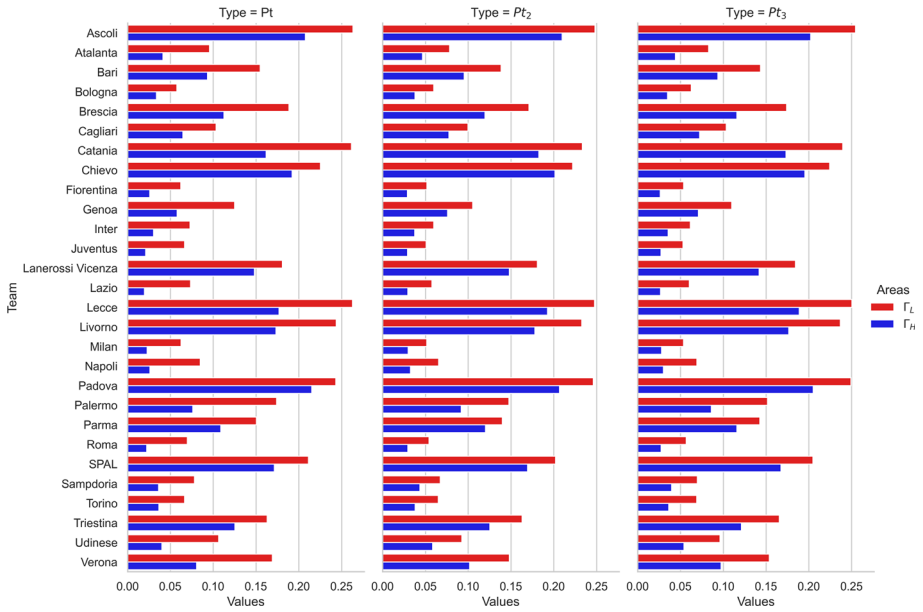


Fig. 9 This figure reports the results from Eq. (5) when the analysis "by team" is performed. The levels of the bar plots represent the areas subtended by the Eq. (2) when the first four (red) and the last four (blue) ranks are considered

EU group itself. Given that, the TV rights value and the business associated to soccer (e.g., appetite for sponsors) increased, mostly driven by the increasing number of matches and the spread of the competition across more countries. Such a phenomenon transformed the soccer dynamic in Italy as well as in other leagues. The teams at the top of the "Serie A" ranking got increased visibility and changes of increasing their earnings through the new business opportunities. With specific reference to the TV rights, the biggest change occurred during the season ending in 1995, namely when the contracts between UEFA and European Broadcasting Union ended. One can notice from Fig. 8, that around that period, the two Γs change trend's direction. Namely, the teams at low ranks have been increasing / holding their competitiveness (see Γ_L), while the others had the opposite (decreasing level of Γ_H).
With the analysis "by team" one captures the phenomenon for which many teams have had "glory years" entering in the "Serie A" and being able to have a "decent" performance for a few times, spending the rest of the permanence in floating in the middle of the ranking. The analysis of the areas performed via Eq. (5) further confirms that in recent years, the teams at the bottom of the rankings have performed worse and worse, underfeeding a general trend followed by the teams present at the top.

Funding Open access funding provided by Alma Mater Studiorum - Universitá di Bologna within the CRUI-CARE Agreement

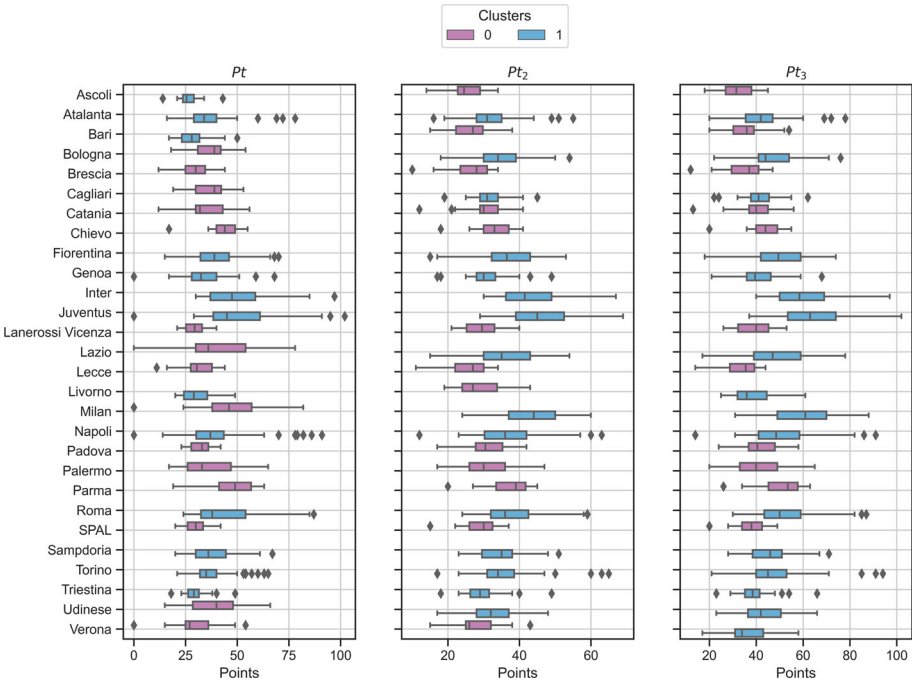


Fig. 10 This figure reports information about the scores' variability for each analysed team, namely those having played at least 16 "Serie A" championships. The coloured areas represent the observations distributed around the mean (black vertical line), namely those comprised between the 25th (Q1) and 75th percentiles ($Q3$), the interquartile range ($IQR = Q3 - Q1$). The edges of the black continuous lines respectively represent the data between $Q1 - 1.5 * IQR$ and $Q3 + 1.5 * IQR$. The rhombus are the outliers positions

Fig. 11 This figure reports information about the years in which the considered teams have competed. It gives an indication of the points obtained via the dots' size. Furthermore, the colour indicates the cluster to which it belongs

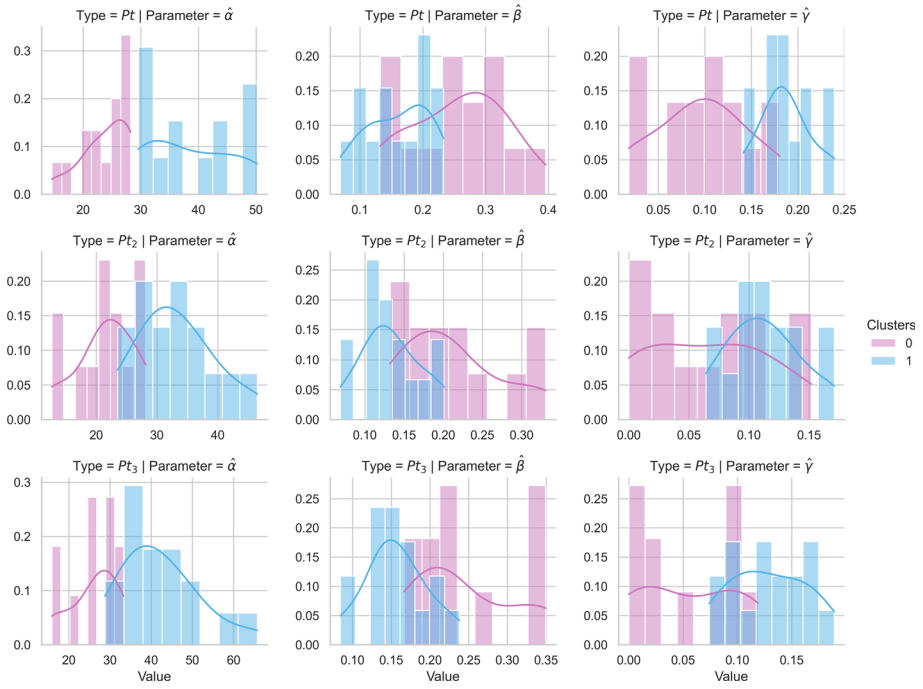


Fig. 12 Empirical distribution of the DGBD's estimated parameters divided by clusters when the analysis "by team" is run

Appendix

See Tables 4 and 5.

Table 4 The column "Date" reports the years in which the championship is finished, "N. Games" contains the number of teams matches played in that season and "N. Teams" reports the number of teams participating. Pt contains the statistical summary of the points scored at the end of the championship, with the penalties deducted from the score reached (official results). In this respect, the value 0 is associated with the relegation of at least a team over the considered year. Pt_2 and Pt_3 contain the summary of the ranked final points scored as if there was in place the rule that assigned two and three points per won match respectively.

Date	N. Games	N. Teams	Pt				Pt_2				Pt_3			
			Max	Min	μ	σ	Max	Min	μ	σ	Max	Min	μ	σ
2020	38	20	83	20	52	18	58	15	38	12	83	20	52	18
2019	38	20	90	17	51	18	62	18	38	11	90	20	51	18
2018	38	20	95	21	52	20	65	15	38	13	95	21	52	20
2017	38	20	91	18	53	20	62	15	38	13	91	18	53	20
2016	38	20	91	28	52	17	62	23	38	11	91	28	52	17
2015	38	20	87	19	50	16	61	20	38	10	87	24	51	15
2014	38	20	102	25	52	19	69	19	38	12	102	25	52	19
2013	38	20	87	22	51	17	60	16	38	10	87	22	52	16
2012	38	20	84	22	51	14	61	18	38	9	84	22	51	14
2011	38	20	82	24	52	14	58	19	38	9	82	24	52	14
2010	38	20	82	29	51	14	58	22	38	9	82	29	51	14
2009	38	20	84	30	52	15	59	25	38	10	84	30	52	15
2008	38	20	85	30	51	15	60	24	38	10	85	30	51	15
2007	38	20	97	26	49	16	67	21	38	11	97	26	51	17
2006	38	20	76	0	42	16	64	18	38	12	91	21	51	19
2005	38	20	86	35	50	13	60	27	38	9	86	35	50	13
2004	34	18	82	13	46	17	57	11	34	11	82	13	46	17
2003	34	18	72	21	45	14	51	17	34	9	72	21	45	14
2002	34	18	71	18	46	14	51	15	34	10	71	18	46	14
2001	34	18	75	20	46	14	53	15	34	9	75	20	46	14
2000	34	18	72	21	45	14	51	17	34	9	72	21	45	14
1999	34	18	70	20	46	12	50	18	34	8	70	22	46	12
1998	34	18	74	14	46	15	53	12	34	10	74	14	46	15
1997	34	18	65	19	45	11	48	17	34	7	65	19	45	11
1996	34	18	73	24	46	14	52	17	34	9	73	24	46	14
1995	34	18	73	12	46	15	50	10	34	10	73	12	46	15
1994	34	18	50	11	34	9	50	11	34	9	69	14	45	13
1993	34	18	50	17	34	8	50	17	34	8	68	23	45	11
1992	34	18	56	14	34	10	56	14	34	10	78	18	44	14
1991	34	18	51	18	34	9	51	18	34	9	71	22	44	13
1990	34	18	51	21	34	9	51	21	34	9	72	25	44	14

Table 4 continued

Date	N. Games	N. Teams	Pt				Pt_2				Pt_3			
			Max	Min	μ	σ	Max	Min	μ	σ	Max	Min	μ	σ
1989	34	18	58	22	34	9	58	22	34	9	84	28	44	14
1988	30	16	45	20	29	7	45	23	30	7	62	28	39	11
1987	30	16	42	15	29	7	42	21	30	6	57	28	39	10
1986	30	16	45	16	30	7	45	16	30	7	63	21	39	11
1985	30	16	43	15	30	8	43	15	30	8	58	17	38	12
1984	30	16	43	12	30	7	43	12	30	7	60	13	39	11
1983	30	16	43	13	30	7	43	13	30	7	59	15	38	10
1982	30	16	46	17	30	7	46	17	30	7	65	20	39	12
1981	30	16	44	16	29	7	44	16	30	7	61	22	39	10
1980	30	16	41	0	26	11	41	16	30	6	55	20	38	10
1979	30	16	44	15	30	7	44	15	30	7	61	17	38	10
1978	30	16	44	17	30	7	44	17	30	7	59	21	39	10
1977	30	16	51	14	29	9	51	14	30	9	74	17	39	14
1976	30	16	45	19	30	7	45	19	30	7	63	24	39	12
1975	30	16	43	17	30	8	43	17	30	8	61	20	39	12
1974	30	16	43	0	27	10	43	17	30	6	61	21	39	10
1973	30	16	45	16	30	9	45	16	30	9	63	19	39	14
1972	30	16	43	13	30	9	43	13	30	9	60	14	39	14
1971	30	16	46	21	30	7	46	21	30	7	65	26	38	11
1970	30	16	45	19	30	7	45	19	30	7	62	24	40	11
1969	30	16	45	19	30	8	45	19	30	8	61	23	39	11
1968	30	16	46	17	30	7	46	17	30	7	64	20	40	10
1967	34	18	49	17	34	9	49	17	34	9	67	20	44	14
1966	34	18	50	15	34	9	50	15	34	9	70	17	45	13
1965	34	18	54	21	34	8	54	21	34	8	76	28	45	13
1964	34	18	54	22	34	9	54	22	34	9	77	28	45	15
1963	34	18	49	20	34	7	49	20	34	7	68	25	45	11
1962	34	18	53	17	34	9	53	17	34	9	77	23	46	15
1961	34	18	49	18	34	7	49	18	34	7	71	23	46	12
1960	34	18	55	0	33	11	55	18	34	8	80	22	45	13
1959	34	18	52	23	34	8	52	23	34	8	72	29	45	13
1958	34	18	51	26	34	6	51	26	34	6	74	34	45	10
1957	34	18	48	22	34	5	48	22	34	5	69	29	45	8
1956	34	18	53	15	34	7	53	15	34	7	73	18	45	11
1955	34	18	48	21	34	7	48	21	34	7	67	27	45	10

Table 4 continued

Date	N. Games	N. Teams	Pt				Pt_2				Pt_3			
			Max	Min	μ	σ	Max	Min	μ	σ	Max	Min	μ	σ
1954	34	18	51	25	34	8	51	25	34	8	71	31	45	12
1953	34	18	47	22	34	6	47	22	34	6	66	29	46	9
1952	38	20	60	17	38	9	60	17	38	9	86	21	52	13
1951	38	20	60	27	38	10	60	27	38	10	86	36	52	15
1950	38	20	62	16	38	10	62	16	38	10	90	21	52	15
1949	38	20	60	26	38	8	60	26	38	8	85	35	52	13
1948	40	21	65	0	38	11	65	26	40	8	94	36	55	12
1947	38	20	63	18	38	9	63	18	38	9	91	23	52	14
1943	30	16	44	21	30	6	44	21	30	6	64	28	41	10
1942	30	16	42	19	30	6	42	19	30	6	58	25	40	9
1941	30	16	39	17	30	4	39	17	30	4	55	22	40	7
1940	30	16	44	22	30	6	44	22	30	6	64	29	41	9
1939	30	16	42	24	30	5	42	24	30	5	58	31	40	8
1938	30	16	41	15	30	8	41	15	30	8	57	18	40	12
1937	30	16	42	18	30	6	42	18	30	6	57	26	40	9
1936	30	16	40	16	30	6	40	16	30	6	55	21	40	9
1935	30	16	44	15	30	7	44	15	30	7	62	20	41	10
1934	34	18	53	17	34	9	53	17	34	9	76	21	46	13
1933	34	18	54	21	34	8	54	21	34	8	79	29	47	12
1932	34	18	54	22	34	8	54	22	34	8	78	29	46	13
1931	34	18	55	19	34	10	55	19	34	10	80	25	47	16
1930	34	18	50	16	34	8	50	16	34	8	72	20	47	12

Table 5 The column "Teams" reports the names of the teams that have played during the years, "N. Games" contains the number of matches played by those teams and "N. Championships" reports the number of tournaments played. Pt contains the statistical summary of the points scored at the end of the championship, with the penalties deducted from the score reached (official results). The value 0 is associated with the relegation of the considered team. Pt_2 and Pt_3 contain the summary of the ranked final points scored as if there was in place the rule that assigned two and three points per won match respectively.

Teams	N. Games	N. Championships	Pt				Pt_2				Pt_3			
			Max	Min	μ	σ	Max	Min	μ	σ	Max	Min	μ	σ
Alessandria	440	13	38	18	29.31	5.15	38	18	29.31	5.15	53	26	39.77	7.53
Ambrosiana	68	2	50	38	44.00	8.49	50	38	44.00	8.49	72	53	62.50	13.44
Ambrosiana-Inter	372	12	49	26	38.25	6.52	49	26	38.25	6.52	69	33	53.08	10.30
Ancona	68	2	19	13	16.00	4.24	19	11	15.00	5.66	25	13	19.00	8.49
Ascoli	508	16	43	14	26.50	6.39	34	14	25.62	5.19	45	18	32.50	7.28
Atalanta	2036	59	78	16	36.03	12.07	55	16	31.93	7.17	78	20	42.32	10.86
Avellino	300	10	30	23	26.40	1.90	30	23	26.90	2.13	40	28	34.80	3.52
Bari	1010	30	50	17	29.13	7.97	38	15	26.73	5.78	54	20	35.33	8.31
Benevento	38	1	21	21	21.00	–	15	15	15.00	–	21	21	21.00	–
Bologna	2460	73	54	18	37.36	7.53	54	18	34.52	6.55	76	22	46.49	10.30
Brescia	774	23	44	12	29.78	8.41	34	10	26.65	6.21	47	12	35.09	8.73
Cagliari	1368	40	53	19	36.55	8.04	45	19	31.35	5.19	62	22	41.42	8.04
Carpi	38	1	38	38	38.00	–	29	29	29.00	–	38	38	38.00	–
Casale	136	4	28	17	22.50	4.65	28	17	22.50	4.65	40	21	30.75	7.93
Catania	602	17	56	12	34.76	11.03	41	12	29.71	6.88	56	13	39.24	10.27
Catanzaro	210	7	29	13	23.43	5.68	29	13	23.43	5.68	37	15	28.86	7.60
Cesena	426	13	43	14	26.00	6.89	32	14	24.54	5.55	43	17	30.69	7.71
Chievo	634	17	55	17	43.53	8.86	41	18	32.76	5.54	55	20	43.71	8.31
Como	426	13	41	17	27.38	7.07	41	17	27.08	7.31	58	20	35.46	11.91
Cremonese	234	7	41	15	24.86	9.30	32	15	22.57	6.48	41	19	28.71	9.03

Table 5 continued

Teams	N. Games	N. Championships	Pt				Pt_2				Pt_3			
			Max	Min	μ	σ	Max	Min	μ	σ	Max	Min	μ	σ
Crotone	76	2	35	34	34.50	0.71	26	25	25.50	0.71	35	34	34.50	0.71
Empoli	462	13	54	20	35.46	10.32	40	18	28.00	5.85	54	22	37.08	8.33
Fiorentina	2762	82	70	15	40.70	11.63	53	15	36.93	7.65	74	18	50.22	11.78
Foggia	256	8	35	18	28.50	5.93	35	24	28.25	4.33	47	30	37.00	6.23
Foggia & Incedit	102	3	31	24	28.00	3.61	31	24	28.00	3.61	41	31	36.33	5.03
Frosinone	76	2	31	25	28.00	4.24	23	20	21.50	2.12	31	25	28.00	4.24
Genoa	1410	40	68	0	34.75	11.49	49	17	31.02	5.96	68	21	41.02	9.17
Genova 1893	410	13	48	24	34.54	6.88	48	24	34.54	6.88	69	32	47.77	10.97
Inter	2522	74	97	30	50.19	15.53	67	30	43.46	8.45	97	40	60.39	12.96
Juventus	2924	87	102	0	51.64	18.68	69	29	46.08	9.34	102	37	64.57	14.70
Lanerossi Vicenza	708	22	40	21	29.64	5.41	40	21	29.64	5.41	53	26	39.23	8.08
Lazio	2624	77	78	0	40.94	15.81	54	15	35.90	8.26	78	17	48.84	12.81
Lecce	564	16	44	11	31.12	9.08	34	11	25.44	6.02	44	14	33.12	8.15
Lecco	102	3	29	17	23.00	6.00	29	17	23.00	6.00	39	20	29.33	9.50
Legnano	106	3	25	17	20.33	4.16	25	17	20.33	4.16	31	21	25.67	5.03
Liguria	150	5	31	21	25.40	3.78	31	21	25.40	3.78	43	28	34.20	5.89
Livorno	626	18	49	20	31.44	8.54	43	19	28.56	6.65	61	25	38.06	9.30
Lucchese	282	8	38	21	30.50	5.66	38	21	30.50	5.66	52	26	40.50	8.47
Mantova	230	7	34	17	26.29	6.52	34	17	26.29	6.52	44	20	33.14	8.45
Messina	182	5	48	22	31.00	10.05	36	21	26.40	6.02	48	26	34.20	8.70

Table 5 continued

Teams	N. Games	N. Championships	Pt				Pt_2				Pt_3			
			Max	Min	μ	σ	Max	Min	μ	σ	Max	Min	μ	σ
Milan	2752	81	82	0	48.31	15.00	60	24	43.27	8.53	88	31	59.86	13.23
Milano	150	5	34	27	29.20	2.77	34	27	29.20	2.77	46	37	39.60	3.65
Modena	444	13	51	19	30.77	9.07	51	19	29.62	8.97	72	25	39.62	13.37
Napoli	2470	74	91	0	40.55	16.54	63	12	36.91	9.31	91	14	50.16	14.39
Novara	446	13	40	21	29.23	4.59	40	21	28.69	4.64	56	29	38.92	7.04
Padova	560	16	42	23	32.06	5.71	42	17	30.88	6.23	58	24	42.25	8.87
Palermo	1030	29	65	17	36.14	13.24	47	17	30.97	7.66	65	20	41.55	11.36
Parma	932	26	63	19	48.42	10.34	45	20	37.38	5.90	63	26	50.88	8.93
Perugia	418	13	46	18	35.31	7.76	41	23	30.15	4.26	52	28	39.85	5.89
Pescara	234	7	27	16	20.14	4.22	27	15	18.86	4.67	32	18	24.00	5.69
Piacenza	272	8	42	21	34.38	6.97	31	17	27.25	5.04	42	21	35.38	6.82
Pisa	218	7	27	20	23.00	2.16	27	20	23.00	2.16	35	25	29.00	3.27
Pistoiese	30	1	16	16	16.00	–	16	16	16.00	–	22	22	22.00	–
Pro Patria	430	12	40	15	28.17	7.64	40	15	28.17	7.64	57	18	38.00	11.25
Pro Vercelli	200	6	34	15	28.50	7.15	34	15	28.50	7.15	46	20	39.33	10.07
Reggiana	102	3	31	18	22.67	7.23	31	14	20.67	9.07	41	18	26.00	13.00
Reggina	326	9	44	31	38.33	3.91	39	25	30.33	4.18	51	31	39.56	5.77
Roma	2924	87	87	24	44.01	16.60	59	24	38.38	8.54	87	30	52.54	13.14
SPAL	658	19	42	20	30.00	5.90	37	15	28.74	5.34	49	20	37.95	7.38
Salernitana	74	2	38	34	36.00	2.83	34	28	31.00	4.24	47	38	42.50	6.36

Table 5 continued

Teams	N. Games	N. Championships	Pt				Pt_2				Pt_3			
			Max	Min	μ	σ	Max	Min	μ	σ	Max	Min	μ	σ
Sampdoria	2164	63	67	20	38.02	10.51	51	23	33.79	6.38	71	28	45.30	9.90
Sampierdarenese	90	3	27	22	25.00	2.65	27	22	25.00	2.65	36	28	33.00	4.36
Sassuolo	266	7	61	34	46.71	8.34	45	25	34.71	6.07	61	34	46.71	8.34
Siena	338	9	44	30	38.78	5.72	35	24	30.33	3.84	44	31	39.56	4.88
Talmone Torino	34	1	23	23	23.00	–	23	23	23.00	–	29	29	29.00	–
Ternana	60	2	19	16	17.50	2.12	19	16	17.50	2.12	23	19	21.00	2.83
Torino	2500	75	65	21	37.56	9.17	65	17	35.33	8.03	94	21	47.57	12.32
Treviso	38	1	21	21	21.00	–	18	18	18.00	–	21	21	21.00	–
Triestina	874	26	49	18	29.77	6.11	49	18	29.77	6.11	66	23	39.50	8.74
Udinese	1638	47	66	15	39.87	13.52	48	17	32.77	7.08	66	23	44.11	10.86
Varese	218	7	32	13	22.14	7.47	32	13	22.14	7.47	44	14	27.29	11.13
Venezia	400	12	42	16	26.25	7.89	38	15	24.58	6.89	53	18	32.42	10.13
Verona	942	29	54	0	29.90	10.73	43	15	27.72	6.28	58	17	36.21	9.18
Vicenza	278	8	49	25	36.38	8.68	39	25	30.00	5.68	55	33	40.62	8.43

Algorithm 1 K-means algorithm

Choose the number of clusters, here $k = 2$; ▷ The choice is driven by the Silhouette scores resulting from tests at different ks, see Table 1, and by the need of capturing two regimes in the rank-size representations of the phenomenon. Further details are reported in Sect. 3.
Place the clusters' centroids $\mu_1 = (\alpha^{(1)}, \beta^{(1)}, \gamma^{(1)})$, $\mu_2 = (\alpha^{(2)}, \beta^{(2)}, \gamma^{(2)})$ according to the *"k-means++"* variant, see Ostrovsky et al. (2013);
repeat
  for $i = [1, \ldots, n]$ **do** ▷ note that in our case n changes for the instances "by years" and "by teams".
    Find the x_i's nearest centroid μ_1 or μ_2 using the minimum Euclidean distance: $\min_{\mu_j \in \mathbb{R}^3 : j=[1,2]}(||x_i - \mu_j||^2)$; ▷ note that x_i is the i^{th} triplet of parameters
    Assign i^{th} data point to the cluster having the closest centroid;
  end for
  Update μ_1 and μ_2 with the average of the values belonging to the respective clusters;
until convergence of centroids reach steady points or until a fixed number of iterations is reached.

References

Arthur, D., & Vassilvitskii, S. (2006). *k-means++: The advantages of careful seeding*. Technical Report 2006–13 Stanford InfoLab.

Ausloos, M. (2013). A scientometrics law about co-authors and their ranking: The co-author core. *Scientometrics, 95*, 895–909.

Ausloos, M. (2014a). Intrinsic classes in the Union of European Football Associations soccer team ranking. *Central European Journal of Physics, 12,* 773–779.

Ausloos, M. (2014b). Zipf–Mandelbrot–Pareto model for co-authorship popularity. *Scientometrics, 101*, 1565–1586.

Ausloos, M. (2020). Rank-size law, financial inequality indices and gain concentrations by cyclist teams. The case of a multiple stage bicycle race, like Tour de France. *Physica A: Statistical Mechanics and Its Applications, 540*, 123161.

Ausloos, M., Cloots, R., Gadomski, A., & Vitanov, N. K. (2014). Ranking structures and rank-rank correlations of countries: The FIFA and UEFA cases. *International Journal of Modern Physics C, 25*, 1450060.

Ausloos, M., Gadomski, A., & Vitanov, N. K. (2014). Primacy and ranking of UEFA soccer teams from biasing organization rules. *Physica Scripta, 89*, 108002.

Baker, B. J., Du, J., Sato, M., & Funk, D. C. (2020). Rethinking segmentation within the psychological continuum model using Bayesian analysis. *Sport Management Review, 23*, 764–775.

Branch, M. A., Coleman, T. F., & Li, Y. (1999). A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM Journal on Scientific Computing, 21*, 1–23.

Cea, S., Durán, G., Guajardo, M., Sauré, D., Siebert, J., & Zamorano, G. (2020). An analytics approach to the FIFA ranking procedure and the World Cup final draw. *Annals of Operations Research, 286*, 119–146.

Cerqueti, R., & Ausloos, M. (2015). Evidence of economic regularities and disparities of Italian regions from aggregated tax income size data. *Physica A: Statistical Mechanics and its Applications, 421*, 187–207.

Commissione d'Appello Federale—ederazione Italiana Giuoco Calcio (2006). Testo della decisione relativa al Comm. Uff. N. 1/c - Riunione del 29 Giugno/3 - 4 - 5 - 6 - 7 Luglio 2006. http://download.ju29ro.com/sentenze/Calciopoli_-_Sentenza_Caf_(14_luglio_2006).pdf in Italian.

Csató, L. (2020). The UEFA Champions League seeding is not strategy-proof since the 2015/16 season. *Annals of Operations Research, 292*, 161–169.

Dimitrova, Z., & Ausloos, M. (2015). Primacy analysis of the system of Bulgarian cities. *Open Physics, 13*, 218–225.

D'Urso, P., & Maharaj, E. A. (2012). Wavelets-based clustering of multivariate time series. *Fuzzy Sets and Systems, 193*, 33–61. Theme: Data Analysis.

Ficcadenti, V., & Cerqueti, R. (2017). Earthquakes economic costs through rank-size laws. *Journal of Statistical Mechanics: Theory and Experiment, 2017*, 083401.

Ficcadenti, V., Cerqueti, R., & Ausloos, M. (2019). A joint text mining-rank size investigation of the rhetoric structures of the US Presidents' speeches. *Expert Systems with Applications, 123*, 127–142.

Ficcadenti, V., Cerqueti, R., Ausloos, M., & Dhesi, G. (2020). Words ranking and Hirsch index for identifying the core of the hapaxes in political texts. *Journal of Informetrics, 14*, 101054.

Filetti, C., Ruscello, B., D'Ottavio, S., & Fanelli, V. (2017). A study of relationships among technical, tactical, physical parameters and final outcomes in elite soccer matches as analyzed by a semiautomatic video tracking system. *Perceptual and Motor Skills, 124*, 601–620.

Frick, B., Barros, C. P., & Prinz, J. (2010). Analysing head coach dismissals in the German "Bundesliga" with a mixed logit approach. *European Journal of Operational Research, 200*, 151–159.

Galariotis, E., Germain, C., & Zopounidis, C. (2018). A combined methodology for the concurrent evaluation of the business, financial and sports performance of football clubs: the case of France. *Annals of Operations Research, 266*, 589–612.

Goes, F. R., Meerhoff, L. A., Bueno, M., Rodrigues, D., Moura, F. A., Brink, M., Elferink-Gemser, M., Knobbe, A., Cunha, S., Torres, R., & Lemmink, K. (2021). Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *European Journal of Sport Science, 21*, 481–496.

Gonçalves, B., Coutinho, D., Santos, S., Lago-Penas, C., Jiménez, S., & Sampaio, J. (2017). Exploring team passing networks and player movement dynamics in youth association football. *PLoS One, 12*, e0171156.

Goossens, D. R., Beliën, J., & Spieksma, F. C. (2012). Comparing league formats with respect to match importance in Belgian football. *Annals of Operations Research, 194*, 223–240.

Hassan, A., Akl, A.-R., Hassan, I., & Sunderland, C. (2020). Predicting wins, losses and attributes' sensitivities in the soccer world cup 2018 using neural network analysis. *Sensors, 20*, 3213.

Hughes, M., Caudrelier, T., James, N., Donnelly, I., Kirkbride, A., & Duschesne, C. (2012). Moneyball and soccer: An analysis of the key performance indicators of elite male soccer players by position. *Journal of Human Sport and Exercise*, *7*.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR), 31*, 264–323.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis* (Vol. 344). Wiley.

Kennedy, P., & Kennedy, D. (2012). Football supporters and the commercialisation of football: Comparative responses across Europe. *Soccer & Society, 13*, 327–340.

Link, D., Lang, S., & Seidenschwarz, P. (2016). Real time quantification of dangerousity in football using spatiotemporal tracking data. *PLoS One, 11*, e0168768.

Malacarne, L. C., & Mendes, R. D. S. (2000). Regularities in football goal distributions. *Physica A: Statistical Mechanics and its Applications, 286*, 391–395.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. *Communication Theory, 84*, 486–502.

Mandelbrot, B. (1961). On the theory of word frequencies and on related Markovian models of discourse. *Structure of Language and Its Mathematical Aspects, 12*, 190–219.

Mansilla, R., Köppen, E., Cocho, G., & Miramontes, P. (2007). On the behavior of journal impact factor rank-order distribution. *Journal of Informetrics, 1*, 155–160.

Martínez-Mekler, G., Martínez, R. A., del Río, M. B., Mansilla, R., Miramontes, P., & Cocho, G. (2009). Universality of rank-ordering distributions in the arts and sciences. *PLoS One, 4*, 1–7.

Masters, J. (2014). Financial Fair Play—Fair or farce? Retrieved May, 2021, from http://edition.cnn.com/2014/05/09/sport/football/football-financial-fair-play/index.html. [Online].

Memmert, D., Lemmink, K. A., & Sampaio, J. (2017). Current approaches to tactical performance analyses in soccer using position data. *Sports Medicine, 47*, 1–10.

Michie, J., & Oughton, C. (2004). *Competitive balance in football: Trends and effects*. The sportsnexus London.

Naumis, G., & Cocho, G. (2008). Tail universalities in rank distributions as an algebraic problem: The beta-like function. *Physica A: Statistical Mechanics and Its Applications, 387*, 84–96.

Neale, W. C. (1964). The peculiar economics of professional sports. *The Quarterly Journal of Economics, 78*, 1–14.

Nicoliello, M., & Zampatti, D. (2016). *Football clubs' profitability after the Financial Fair Play regulation: Evidence from Italy*. Sport, Business and Management: An International Journal.

Ostrovsky, R., Rabani, Y., Schulman, L. J., & Swamy, C. (2013). The effectiveness of Lloyd-type methods for the k-means problem. *Journal of the ACM, 59*, 1–22.

Özaydin, S., & Donduran, M. (2019). An empirical study of revenue generation and competitive balance relationship in European football. *Eurasian Journal of Business and Economics, 12*, 17–44.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. *SpringerPlus, 5*, 1–13.

Ric, A., Torrents, C., Gonçalves, B., Torres-Ronda, L., Sampaio, J., & Hristovski, R. (2017). Dynamics of tactical behaviour in association football when manipulating players' space of interaction. *PLoS One, 12*, e0180773.

Rimmer, P., & Johnston, R. (1967). Areas of community interest in Victoria as indicated by competitive sport. *Australian Geographer, 10*, 311–313.

Rotundo, G. (2014). Black–Scholes–Schrödinger–Zipf–Mandelbrot model framework for improving a study of the coauthor core score. *Physica A: Statistical Mechanics and Its Applications, 404*, 296–301.

Worldfootball.net (2021a). Serie A—Archive. Retrieved May, 2021, from https://www.worldfootball.net/history/ita-serie-a/ [Online].

Worldfootball.net (2021b). Serie A 2018/2019—Attendance data. Retrieved May, 2021, from https://www.worldfootball.net/attendance/ita-serie-a-2018-2019/1/ [Online].

Yoon, K. P., & Sedaghat, M. (2020). Rank power analysis for comparative strength of professional sports franchises. *Journal of Applied Business Research, 36*, 181–196.

Zipf, G. K. (1935). *The psycho-biology of language*. Houghton-Mifflin.

Zipf, G. K. (1949). *Human behaviour and the principle of least-effort*. Addison-Wesley Press.