Computer Physics Communications 184 (2013) 661–670

Contents lists available at SciVerse ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

A simple non-Markovian computational model of the statistics of soccer leagues: Emergence and scaling effects

Roberto da Silva [a,*], Mendeli H. Vainstein [a], Luis C. Lamb [b], Sandra D. Prado [a]

[a] Instituto de Física, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves 9500, Caixa Postal 15051 91501-970, Porto Alegre RS, Brazil

[b] Instituto de Informática, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves 9500, Caixa Postal 15064 91501-970, Porto Alegre, RS, Brazil

ARTICLE INFO

Article history:
Received 7 June 2012
Received in revised form
13 September 2012
Accepted 30 October 2012
Available online 5 November 2012

Keywords:
Statistics in sports
Evolutionary automata for soccer tournament modeling
Scaling effects

ABSTRACT

We propose a novel probabilistic model that outputs the final standings of a soccer league, based on a simple dynamics that mimics a soccer tournament. In our model, a team is created with a defined potential (ability) which is updated during the tournament according to the results of previous games. The updated potential modifies a team future winning/losing probabilities. We show that this evolutionary game is able to reproduce the statistical properties of final standings of actual editions of the Brazilian tournament (*Brasileirão*) if the starting potential is the same for all teams. Other leagues such as the Italian (Calcio) and the Spanish (La Liga) tournaments have notoriously non-Gaussian traces and cannot be straightforwardly reproduced by this evolutionary non-Markovian model with simple initial conditions. However, we show that by setting the initial abilities based on data from previous tournaments, our model is able to capture the stylized statistical features of double round robin system (DRRS) tournaments in general. A complete understanding of these phenomena deserves much more attention, but we suggest a simple explanation based on data collected in Brazil: here several teams have been crowned champion in previous editions corroborating that the champion typically emerges from random fluctuations that partly preserve the Gaussian traces during the tournament. On the other hand, in the Italian and Spanish cases, only a few teams in recent history have won their league tournaments. These leagues are based on more robust and hierarchical structures established even before the beginning of the tournament. For the sake of completeness, we also elaborate a totally Gaussian model (which equalizes the winning, drawing, and losing probabilities) and we show that the scores of the Brazilian tournament "Brasileirão" cannot be reproduced. This shows that the evolutionary aspects are not superfluous and play an important role which must be considered in other alternative models. Finally, we analyze the distortions of our model in situations where a large number of teams is considered, showing the existence of a transition from a single to a double peaked histogram of the final classification scores. An interesting scaling is presented for different sized tournaments.

1. Introduction

Soccer is an extremely popular and profitable, multi-billion dollar business around the world. Recently, several aspects regarding the sport and associated businesses have been the subject of investigation by the scientific community, including physicists who have devoted some work and time to describe statistics related to soccer (see for example [1–6]). In the literature about soccer models, one can find applications of complex networks [7] and fits with generalized functions [8]; however, they ofttimes have only one focus: goal distribution (see e.g. [1,2,4]).

Outside the soccer literature, it is important to mention other interesting studies which do not necessarily focus on the scores of the games, such as models that investigate properties of patterns emerging from failure/success processes in sports. In the case of basketball, it has been suggested [9] that the "hot hand" phenomenon (the belief that during a particular period a player's performance is significantly better than expected on the basis of a player's overall record), a definitively non-random pattern can be modeled by a sequence of random independent trials.

Returning to soccer, some authors [10] have devoted attention to the influence of the perceptual-motor bias associated with reading direction in foul judgment by referees. Heuer and collaborators have explored very interesting aspects in soccer and other sports: in [5], they explore the possibility of goal scoring being characterized as independent Poissonian processes with pre-determined expectation values, and obtain a good match between their model and experimental data. In [6], the authors

* Corresponding author.
E-mail addresses: rdasilva@if.ufrgs.br (R. da Silva), vainstein@if.ufrgs.br (M.H. Vainstein), lamb@inf.ufrgs.br (L.C. Lamb), prado@if.ufrgs.br (S.D. Prado).

0010-4655/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.cpc.2012.10.030

study the effects of dismissal of coaches on teams. Smiatek and Heuer show that, differently from soccer, the home advantage is nearly negligible compared to the total sum of goals (g) in handball tournaments [11]. More precisely, the authors calculated the difference $\Delta g = g_H - g_A$ between the number of goals in home and away matches, respectively. In handball, the authors obtain $\Delta g/g = 0.033$ while for soccer this number is much higher $\Delta g/g = 0.17$.

However, it is interesting to notice that there is a void in the literature: few studies have been carried out under the game theoretical approach of considering the outcome of a tournament from a simple dynamics among the competing teams. In other words, in looking at the statistics that emerge from this complex system called soccer, one can ask if the properties of the distribution of final tournament classification points can be seen as an emerging property of a soccer tournament dynamics established by simple rules among the different competing teams, or how these classification point distributions emerge from a soccer tournament by considering all "combats" among the teams. Here, we propose a model that combines previous studies concerning goal distribution [4] and a game theoretical approach to football tournaments that produces realistic final tournament scores and standings.

In this paper, we explore the statistics of standing points at the end of tournaments disputed according to the "Double Round Robin System" (DRRS)[1] in which the team with the most tournament points at the end of the season is crowned the champion, since many soccer tournament tables around the world are based on this well-known system. In general, 20 teams take part in the first tier tournament, such as "Serie A" in Italy, the English "Premier League", the Spanish "La Liga" and the Brazilian "Brasileirão" (from 2003 onwards) soccer tournaments. During the course of a season, each team plays every other team twice: the "home" and "away" games. Moreover the points awarded in each match follow the 3-1-0 points system: teams receive three points for a win and one point for a draw; no points are awarded for a loss. The Serie A Italian soccer tournament, or simply the "Calcio", has been played since 1898, but only from 1929 was it disputed in its current format and system. Their main champions have been Juventus, winner of the league 27 times, and Milan and Internazionale which won the league 18 times each. The Spanish "La Liga" also started in 1929, and over its history, the tournament has been widely dominated by only two teams: Real Madrid and Barcelona.

In Brazil, the national tournament, popularly known as "Brasileirão", was first organized in a modern format in 1971. In 2010 the Brazilian Soccer Confederation (CBF) recognized as national champions the winners of smaller national tournaments such as the "Taça Brasil" (played from 1959 to 1968) and another tournament known as "Roberto Gomes Pedrosa" (played from 1967 to 1970). However, only in 2003 the Brazilian League started being disputed via the DRRS. In all past editions of the tournament the league table was based on the method of preliminaries, typically used in tennis tournaments, which will not be considered in this paper. In the 10 editions played under the DRRS, the Brazilian tournament has already been won by 6 different football clubs: Cruzeiro, Santos, São Paulo, Corinthians, Flamengo, and Fluminense.

The statistics, as well as the fluctuations, associated to the standings and scores of teams in tournaments with 20 teams playing under the DRRS can be very interesting. Moreover, if we are able to reproduce such statistics via a simple automaton considering the teams as "agents" which evolve according to definite "rules" based on their previous performances and conditions, one could use this information when preparing or building up a team before a competition. Thus, models (e.g. automata) of games in a tournament, whose results are defined by the evolving characteristics of the teams, could provide important knowledge. Therefore, by exploring the conditions under which the standing and scores of tournaments can be mimicked by a model, we propose a simple, but very illustrative, evolutionary non-Markovian process.

It is known that many events can alter the performance of teams during a season besides their initial strengths, such as the hiring of a new player, renewed motivation due to a change in coach, key player injuries, and trading of players, among others. For the sake of simplicity, we consider that the teams in the model initially have the same chance of winning the games and that the combination of events that can lead to an improvement of a team will be modeled solely by increasing the probability of a team winning future games after a victory. Similarly, a loss should negatively affect their future winning probabilities.

Our main goal is to verify if the Brazilian soccer tournament has final standing scores with the same statistical properties that emerge from our simple model, and to check whether the properties of the Brazilian tournament differ from other leagues and, if so, the reasons for that behavior. In the first part of the paper we calibrate our model by using constant draw probabilities introduced ad hoc, based on data from real tournaments. In the second part, we have used draw probabilities that emerge from the model dynamics itself, being dependent on the teams "abilities". Both situations are able to reproduce real tournament data. The advantage of the second approach is the independence of extra parameters, i.e., the first one uses pre-calculated draw rates from previous statistics. Additionally, we show that it is possible to reproduce the European tournaments results and obtain good fits by slightly altering the model and seeding it initially with data from a previous championship. We also analyze distortions of our model under the hypotheses of inflated tournaments. Finally, we show a transition from single to double peaked histograms of final standing scores, which occurs when we analyze a small league and large tournaments. However, it is possible to obtain a scaling for different tournaments with different sizes.

2. A first model: ad hoc draw probabilities

In our model, each team starts with a potential $\varphi_i(0) = \varphi_0$, where $i = 1, \ldots, n$ indexes the teams. Each team plays once with the other $n - 1$ teams in each half of the tournament; a team A plays with B in the first half of the tournament and B plays with A in the second, i.e. the same game occurs twice in the tournament and there is no distinction between home and away matches (the "home field advantage" could be inserted in the potential of the teams). In a game between team i and team j, the probability that i beats j is given by

$$\Pr(i \succ j) = \frac{\varphi_i}{(\varphi_i + \varphi_j)}. \tag{1}$$

The number of games in the tournament is $N = n(n-1)$ and in each half of the tournament, $n-1$ rounds of $n/2$ games are played. In each round, a matching is performed over the teams by a simple algorithm, that considers all circular permutations to generate the games. We give an illustration for $n = 6$ teams, starting with the configuration:

1 2 3
4 5 6.

This configuration implies that in the first round, team 1 plays team 4, 2 plays 5 and team 3 plays team 6. To generate the second

[1] http://en.wikipedia.org/wiki/Round-robin_tournament.