Anomalous diffusion in the evolution of soccer championship scores: Real data, mean-field analysis, and an agent-based model

Roberto da Silva,[*] Mendeli H. Vainstein,[†] and Sebastián Gonçalves[‡]
Instituto de Física, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, 9500, CEP 91501-970, Porto Alegre, Rio Grande do Sul, Brazil

Felipe S. F. Paula[§]
Instituto de Informática, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, 9500, CEP 91501-970, Porto Alegre, Rio Grande do Sul, Brazil

(Received 31 March 2013; revised manuscript received 17 July 2013; published 21 August 2013)

Statistics of soccer tournament scores based on the double round robin system of several countries are studied. Exploring the dynamics of team scoring during tournament seasons from recent years we find evidences of superdiffusion. A mean-field analysis results in a drift velocity equal to that of real data but in a different diffusion coefficient. Along with the analysis of real data we present the results of simulations of soccer tournaments obtained by an agent-based model which successfully describes the final scoring distribution [da Silva *et al.*, Comput. Phys. Commun. **184**, 661 (2013)]. Such model yields random walks of scores over time with the same anomalous diffusion as observed in real data.

DOI: 10.1103/PhysRevE.88.022136 PACS number(s): 02.50.Le, 01.80.+b, 05.40.Fb

I. INTRODUCTION

Soccer, commonly known as football (outside the USA), is a sport played between two teams of typically 11 players with a spherical ball. The game, presently played in more than 200 countries according to FIFA, is the world's most popular sport whose beauty is appreciated even by scientists, physicians, and economists [1]. Recently it has even been shown that the performance of a country's football team influences that of its stock market [2], indicating the vast amounts of money involved in the sport. Several aspects regarding soccer and its associated businesses have been the subject of interest of the scientific community. Indeed, some statistical descriptions related to soccer appeared in the physics literature, using concepts of complex networks [3] and generalized functions [4]. However, they generally focus on goal distribution (see, for example, Refs. [5–7]) but not on the evolving properties during a tournament season, which is precisely our intention. Without going into greater technical details we should mention that apart from the passion, enthusiasm, and money which revolve around this sport, the factors that may determine which one of the two teams is the favorite in a game are difficult to grasp. As in any other sport, even the undoubtedly favorite may lose the game. It has been shown that computer models such as developed in Ref. [8] outperform expert human tipsters in predicting the outcome of a game; in fact, the latter are easily beaten by using a simple strategy: betting on the home team in every game guarantees a winning chance of 47% while tipsters are right only 42% of the time. On the other hand, such computer models cannot guarantee riches [9]. Since a random component is always present, we wonder whether the fluctuations in the scoring process during soccer tournaments can be captured by a model based on a few simple assumptions.

Focusing on the simple aspects that govern the properties of the temporal evolution of team scores during a tournament, at the end of a game, a team can have one of three possible results: win (if it scored more than its opponent), draw (if the score is tied), or loss (if it scored less than its opponent), which counts as 3, 1, or 0 points, respectively, in a typical double round robin system (DRRS). Our goal in this contribution is to check whether a previous model presented by some of us [10] can statistically reproduce the drift and dispersion of the accumulated points in DRRS soccer tournaments, in which, during the course of a season, each team plays every other team twice, the "home" and "away" games, and no team is eliminated until the end of the season when the team with the most tournament points is crowned the champion.

As we are interested in the scoring process of teams during championship seasons to infer about its diffusive properties, we will denote by $x_i(t)$ the number of points accumulated by team $i = 1,\ldots,n$ at time t in a tournament with n teams. We start our analysis checking the behavior of the average number of accumulated points, i.e., $\langle x(t)\rangle = \frac{1}{n}\sum_{i=1}^{n} x_i(t)$ in round t, and the dispersion $\sigma(t) = \sqrt{\langle x(t)^2\rangle - \langle x(t)\rangle^2}$. The championship score in soccer has an intrinsic monotonic behavior, i.e., $\langle x(t)\rangle > \langle x(t-1)\rangle$, even in the extreme event of a generalized draw. In order to check these quantities in a wide number of cases, we use data for all the teams from four international and recognized soccer leagues: Brazilian, Spanish, French, and English. All these leagues have $n = 20$ teams playing according to the DRRS. As an example, in Fig. 1 we show $\langle x(t)\rangle$ and $\sigma(t)$ as functions of t (round), calculated with data from the 2007–2008 season of the Spanish tournament (La Liga); plots for other countries and seasons are similar (not shown). In the left panel, we display the score evolution of each team with continuous lines (color version: gray); symbols correspond to their average, and the continuous

[*] rdasilva@if.ufgrs.br
[†] vainstein@if.ufrgs.br
[‡] sgonc@if.ufrgs.br
[§] felipesfpaula@inf.ufrgs.br

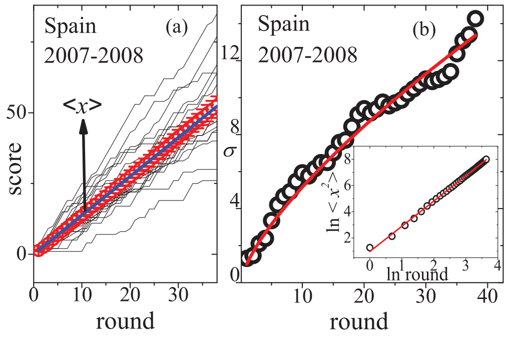

FIG. 1. (Color online) (a) Continuous thin lines correspond to the scores of the different teams in the Spanish tournament (season 2007–2008) throughout time (round). The points correspond to the average of the scores, $\langle x(t)\rangle$, with their respective error bars. The bold continuous line (color version blue) corresponds to the linear fit for $\langle x(t)\rangle$. (b) Time evolution of the standard deviation $\sigma(t)$ of the same scores. The continuous red line corresponds to a fit of a power function $\sigma(t) = Dt^{\beta}$. Inset: log-log plot for the time evolution of the second moment $\langle x^2(t)\rangle$.

bold line (color version: blue) is a linear fit of the average. On the right panel we show the corresponding plots for the dispersion.

What is the expected behavior for $\langle x(t)\rangle$? In every round, each team will score 0, 1, or 3 points; therefore the average above will always increase by a value between 1 (if all games in a round end in a draw) and 1.5 (no draws, so half of the teams score 3 and the other half, 0). The fact that Ct gives a good fit of $\langle x(t)\rangle$ for all championships and seasons, as can be seen in Table I, reveals that each championship could be classified in terms of the "drift velocity" C; the larger it is, the lower is the number of draws, and the more attractive is the championship. In Fig. 1(b) the continuous line corresponds to the fit $\sigma(t) = \sqrt{D t^{\beta}}$ where D is tentatively ascribed to the diffusion constant of the tournament. The log-log plot in the inset of the same figure shows the power law behavior of the second moment $\langle x^2(t)\rangle \propto t^{\xi}$. The analysis of the second moment and the related diffusive process allow one to infer about the properties of the complex systems.

In order to know whether a team scores in the course of a tournament can be represented by a diffusive process (and of what type), we analyze the four international soccer leagues performing linear fits in the log-log scale of $\sigma(t)$ and $\langle x^2(t)\rangle$. We also analyze the behavior of the third moment of scores expected to be as $\langle x^3(t)\rangle = E^* t^{\gamma}$, where E^* and γ are parameters to be compared among the different tournaments. The results for coefficients and exponents are presented in Table I, where we observe a reasonable agreement for the coefficient C for the different tournaments and seasons, with values ranging from 1.33 (France, 2010) to 1.4 (Spain, 2010). According to our arguments above we could say that the Spanish championship is more attractive than the other three.

The values obtained for D have a greater dispersion, ranging from 0.79 to 1.96. In addition, the values of the exponent $0.54 < \beta < 0.82$ suggest a case of superdiffusion, while we observe a universal exponent $1.83 < \xi < 1.94$ for the second moment. Such diffusive behavior suggests that a mean field model could be proposed, taking into account the simplest aspects of soccer dynamics.

Therefore, with the previous analysis at hand, the paper is organized as follows: in the next section we propose a mean-field analysis for the scoring process of teams in soccer leagues. In Sec. III we present some characteristics of the agent-based model aim to describe the dynamics of soccer tournaments based on DRRS. A comparison among the real data, mean-field approximation, and agent-based model is presented in Sec. IV. Finally, we summarize our main conclusions in Sec. V.

TABLE I. Propagation velocity C, diffusion coefficient D, coefficient E^*, and exponents obtained from fits $\langle x(t)\rangle = Ct$, $\sigma(t) = Dt^{\beta}$, $\langle x^2(t)\rangle \propto t^{xi}$, and $\langle x^3(t)\rangle = E^* t^{\gamma}$ to data from real tournaments.

	Coefficients C D E^*				Exponents ξ β γ			
	Brazil	Spain	France	England	Brazil	Spain	France	England
2007	1.37(2)	1.38(2)	1.34(2)	1.36(2)	1.86(2)	1.93(1)	1.89(1)	1.90(1)
	1.06(16)	1.21(8)	1.41(10)	0.88(9)	0.64(3)	0.68(1)	0.61(1)	0.82(2)
	7.2(7)	6.0(3)	6.7(3)	6.7(5)	2.72(3)	2.80(2)	2.73(2)	2.81(2)
2008	1.38(2)	1.38(2)	1.34(2)	1.36(2)	1.90(1)	1.92(1)	1.86(1)	1.89(1)
	1.14(8)	1.18(8)	1.44(10)	0.79(11)	0.63(2)	0.71(1)	0.66(1)	0.76(2)
	6.1(4)	6.3(4)	7.6(4)	6.5(5)	2.77(2)	2.80(2)	2.71(2)	2.78(3)
2009	1.37(2)	1.38(2)	1.38(2)	1.37(3)	1.94(1)	1.88(1)	1.87(1)	1.83(1)
	1.16(8)	1.28(7)	1.36(9)	1.34(14)	0.59(1)	0.74(1)	0.68(1)	0.71(2)
	5.3(2)	8.0(3)	7.7(4)	9.5(5)	2.80(2)	2.75(2)	2.73(2)	2.68(2)
2010	1.35(2)	1.40(2)	1.33(2)	1.35(2)	1.89(1)	1.92(1)	1.90(1)	1.91(1)
	1.10(8)	0.98(8)	0.90(11)	0.86(9)	0.62(1)	0.78(1)	0.66(2)	0.68(2)
	6.1(4)	6.4(3)	5.5(3)	5.6(3)	2.75(2)	2.83(2)	2.77(2)	2.79(2)
2011	1.36(2)	1.37(2)	1.35(2)	1.38(2)	1.85(1)	1.90(1)	1.90(1)	1.92(1)
	1.96(11)	1.27(7)	1.16(9)	1.25(7)	0.54(1)	0.71(1)	0.68(1)	0.74(1)
	8.2(4)	6.8(3)	6.1(4)	7.0(2)	2.68(2)	2.78(2)	2.77(2)	2.80(1)

II. MEAN-FIELD REGIME

Defining r as the mean draw probability, the win and loss mean probabilities are $p_w = p_l = (1 - r)/2$. Based on the DRRS scores 3-1-0, and defining $P_m(n)$ as the probability for a team to have n points (position) at round m, and supposing a Markovian process, we have

$$P_{m+1}(n) = r P_m(n-1) + p_w P_m(n-3) + p_l P_m(n). \quad (1)$$

In order to obtain a partial differential equation that describes the diffusive process in soccer, we can think that $P_m(n)$ denotes the probability of movement of a particle that can either stay still or move to the right in steps of one or three units. Using the relation between draw, win, and loss probabilities we obtain the following relation:

$$P_{m+1}(n) - P_m(n) = -r\,[P_m(n) - P_m(n-1)] - (1-r)/ \\ 2\,[P_m(n) - P_m(n-3)]\,. \quad (2)$$

The expression in the last term can be written as

$$P_m(n) - P_m(n-3) = \Delta^{(2)} P_m(n) - \Delta^{(2)} P_m(n-1) \\ + 3\Delta P_m(n-1), \quad (3)$$

where $\Delta P_m(n) = P_m(n) - P_m(n-1)$ is the first finite difference and $\Delta^{(2)} P_m(n) = \Delta P_m(n) - \Delta P_m(n-1) = P_m(n) + P_m(n-2) - 2P_m(n-1)$ is the second finite difference. Using the definition of the third finite difference, $\Delta^{(3)} P_m(n) = \Delta^{(2)} P_m(n) - \Delta^{(2)} P_m(n-1)$ and its relation to the first finite difference in the last term in the Eq. (2), and considering a similar finite difference for the time variable, it is possible to arrive at the related partial differential equation (PDE)

$$\frac{\partial P}{\partial t} = -C\frac{\partial P}{\partial x} + D\frac{\partial^2 P}{\partial x^2} - E\frac{\partial^3 P}{\partial x^3}, \quad (4)$$

where $P(x,t)$ denotes the probability of a particle (team) having score (position) x at time (round) t, and $C = r + 3(1 - r)/2$, $D = 3(1 - r)/2$ and $E = (1 - r)/2$.

At this point we want to highlight the assumptions of the mean-field approximation:

(1) Each team $i = 1,2,\ldots,n$ has an scoring rate (drift coefficient), i.e., c_i, which is not time dependent.

(2) The individual stochastic process of each team can be fitted as $x_i(t) = c_i t + \eta_i$, which can be thought of as a particular realization of a unique stochastic process.

(3) The constants η_i are considered identically equal to 0.

These three assumptions will be tested and discussed below. For that we perform statistical analysis for c_i and η_i. In order to check our hypothesis, we perform a linear fit for each team trajectory for the different championships. In Fig. 2 we show the distribution of parameters c_i and η_i obtained from linear fits. In plot (a) we exhibit the histograms of c_i values from two different linear fits: (1) the regular fit and (2) by fixing $\eta_i = 0$. As can be observed, both histograms are very similar, which is confirmed by a statistical comparison performed on them and presented in the first two columns of Table II. No significant differences among the samples drawn by the two different fits (regular and origin fixed) were found at level of 0.05, so the third assumption is well supported.

Normality was tested by means of Kolmogorov-Smirnov (KS) and Shapiro-Wilk (SW) tests. No evidence of normality was found at a 5% level coherent with the nonsymmetric distribution in Fig. 2(a). Consistently the values of skewness are similar and different from 0 for both samples. This is an indicator that soccer scores do not behave as normal diffusion.

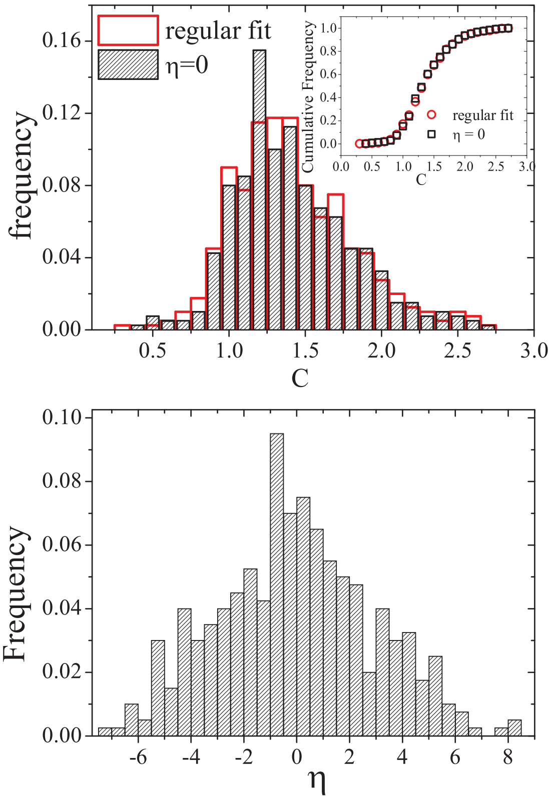

FIG. 2. (Color online) (a) Distribution of slopes c_i for the individual linear fits made by fixing $\eta = 0$ and without fixing it (regular fit). (b) Distribution of linear coefficients η.

The mean-field approximation assumes no time dependence of coefficients c_i. Is such really the case? We know that each team can have c_i between 0 (no wins in a tournament) and 3 (wins all games). However, taking all teams together, the average C can have values between 1 and 3/2. The first one represents the very odd situation of a whole championship with all games tied. The second one corresponds to the other extreme situation where no draw occurred. For a particular round the average score is calculated by

TABLE II. The first four columns describe the results from statistical tests for the slopes c_i performing different linear fits: (a) $x_i(t) = \eta_i + c_i t$, (b) $x_i(t) = c_i t$, (c) fit of first half of seasons, and (d) fit of second part of the seasons. The last column corresponds to the results for the linear coefficient η in the fit corresponding to the first column.

Statistics	Regular fit	$\eta_i = 0$	First half	Second half	η
Average	1.366	1.364	1.362	1.364	−0.036
Standard deviation	0.389	0.388	0.446	0.384	2.96
N. test 5% (SW)	No	No	No	No	Yes
N. test 5% (W)	No	No	No	No	Yes
Skewness	0.61	0.63	0.52	0.64	0.11

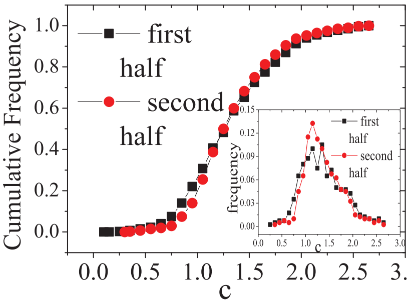

FIG. 3. (Color online) Cumulative histograms of c_i values for all tournaments (Brazil, Spain, France, and English) considering linear fits obtained for the two different parts of tournaments: Part I: 19 initial rounds; Part II: 19 final rounds. Inset plot shows the regular histogram of data.

$\langle x \rangle = p_{\rm draw} + 3p_{\rm win} = (3-r)/2$, which is exactly the constant C. Its value is 4/3 when $r = 1/3$. For the analyzed tournaments we observed $r < 1/3$, corresponding to $1.36 < \overline{C} < 1.37$. Therefore, as a final *tour de force* for the time independence of c_i, we do the following: we split all the time series in two halves and apply a linear fitting for all trajectories for the first 19 rounds and the second 19 rounds. Then we make the corresponding histograms for each set and compare them.

We can see an excellent agreement (visual) between the cumulative histograms of the two analyzed periods as shown in Fig. 3. Such agreement is numerically checked by statistics of the third and fourth columns in Table II. We have similar values of average, skewness, and the normality tests, which allows us to conclude that the hypothesis of no time dependence of the coefficients c_i is very well founded. By performing a simple hypothesis test we also find evidence that there is no difference between the two periods analyzed since the averages 1.363(22) and 1.365(19) corresponding to the two different time intervals are identical at a level of 5%.

We mention in passing that the agent-based model [10] implicitly considers a time dependence of parameters by means of the score-dependent potential of each team.

In addition, in Fig. 2(b) we display a histogram of η_i obtained by the general linear fit. In this case we can observe a distribution centered around 0. Actually $\overline{\eta_i} = -0.04 \pm 0.14$ and is normally distributed since for both methods (SW and KS) the distribution of η_i is normal at 5% level (such results are shown in the last column in Table II). Therefore this is another indicator that the η_i are not relevant for a statistical description of c_i.

Here it is important to comment on generalizations of Fokker-Planck equations, of which Eq. (4) is a case. The most embracing generalization of Fokker-Planck equations is given by the Kramers-Moyal expansion

$$\frac{\partial P}{\partial t} = \sum_{n=1}^{\infty} \left(-\frac{\partial}{\partial x} \right)^n D^{(n)}(x,t) P(x,t) = \hat{L}_{\rm KM} P(x,t), \qquad (5)$$

where $\hat{L}_{\rm KM}$ is the Kramers-Moyal operator and the differential operators act on the product $D^{(n)}(x,t)\,P(x,t)$. It is required by the Pawula theorem [11] that for the probability $P(x,t)$ to be positive at all times, the Kramers-Moyal expansion should either retain an infinite number of terms or be truncated after the first or second term, resulting in a Fokker-Planck equation with Gaussian noise [12]. However, expansions containing more than the first two terms can be of use to approximate the distribution functions, even though the probability must then have negative values at least for sufficiently small times, since these negative values may be very small [for our results $O(10^{-12})$]. It should be also noted that this is not a problem in our case (as will be discussed later), since these small negative values occur where the probability is negligibly small and can be approximated by zero. In a similar case, reported by Risken (see Ref. [11], p. 79), the Kramers-Moyal expansion truncated at a proper $n \geqslant 3$ is in better agreement with the exact distribution for a Poisson process than the distribution function following from the Kramers-Moyal expansion truncated at $n = 2$.

We can obtain a solution of the previous PDE by using the Fourier transform method. Defining $\widehat{P}(k,t) = \int_{-\infty}^{\infty} P(x,t) e^{ikx}\, dx$, if we multiply Eq. (4) by $\exp(ikx)$ and integrate by parts supposing that $P(\pm\infty,t) = \frac{\partial P}{\partial x}(\pm\infty,t) = \frac{\partial^2 P}{\partial x^2}(\pm\infty,t) = \frac{\partial^3 P}{\partial x^3}(\pm\infty,t) = 0$, we obtain $\frac{d}{dt}\widehat{P}(k,t) = (iCk - Dk^2 - iEk^3)\widehat{P}(k,t)$, for which the solution is $\widehat{P}(k,t) = \widehat{P}(k,0)\exp[(ick - Dk^2 - iEk^3)t]$, where $\widehat{P}(k,0)$ is the transform of the initial distribution, i.e., $\int_{-\infty}^{\infty} P(y,0) e^{iky}\, dy$. Taking the inverse Fourier transformation, we get

$$P(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{(iCk - Dk^2 - iEk^3)t}\, dk \int_{-\infty}^{\infty} e^{ik(y-x)} P(y,0)\, dy. \qquad (6)$$

All teams start with zero points, so $P(y,0) = \delta(y)$; then, since $\cos(Ck - Ek^3)$ is an even function while $\sin(Ck - Ek^3)$ is an odd function of k, we obtain a closed form for the distribution $P(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \cos[-Ek^3 t + (Ct - x)k] e^{-Dk^2 t}\, dk$. Since we obtain the general solution in the mean-field regime

$$P(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \cos\big((r-1)k^3 t/2 + \{[r + 3(1-r)/2\,]t - x\}\,k\big) e^{-3(1-r)k^2 t/2}\, dk, \qquad (7)$$

but we are interested only in the solution for the integer values $x = 0,1,2,\ldots,6(n-1)$ and $t = 1,\ldots,2(n-1)$, considering soccer tournaments based on the DRRS, the scoring distribution is discrete and is therefore more properly described by

$$P_{\rm disc}(x,t) = \frac{P(x,t)}{\sum_{j=0}^{6(n-1)} P(j,t)}. \qquad (8)$$

To obtain $\langle x(t) \rangle, \langle x^2(t) \rangle$, and $\sigma(t)$ as functions of t to compare with the corresponding quantities from the real data, we need to numerically estimate the integral in Eq. (7) and the moments of the distribution $P_{\rm disc}(x,t)$, which we perform by means of Simpson's rule.

Having presented the mean-field approximation for the dynamics of soccer tournaments, the next section will be devoted to the agent-based model previously proposed in Ref. [10], used here to explore the diffusive properties of DRRS soccer tournaments.

III. AGENT-BASED MODEL

We turn now to a simple model presented by some of us [10], which has already reproduced some statistical results that emerge in soccer, with the aim of getting a more accurate statistical representation of the dynamics of soccer teams' participation in a tournament. The model is as follows: n teams are playing a DRRS tournament where, in the game between teams i and j at round t, the probability that i beats j is given by

$$\Pr(i \succ j,t) = \left[1 - r_{\text{draw}}^{(i,j)}(t)\right]\frac{\varphi_t^{(i)}}{\left(\varphi_t^{(i)} + \varphi_t^{(j)}\right)}, \qquad (9)$$

where $\varphi_t^{(i)}$, $i = 1,\dots,n$ is the ith team potential at round t. In the most elaborated version of the model (which will be used in this paper and is called prescription III in Ref. [10]), the initial condition $\varphi_0^{(i)}$ is equaled to the average number of goals per game of the team classified in the ith position in a previous real tournament. Here $r_{\text{draw}}^{(i,j)}(t)$ is the probability of draw between the two teams and is calculated based on the respective goal scoring probabilities, modeled by Poisson distributions whose parameters are their potentials $\varphi_t^{(i)}$ and $\varphi_t^{(j)}$. Accordingly, given their potentials, the draw probability is the probability that both teams score the same number of goals ($n_i = n_j = n$), which is calculated by

$$\begin{aligned} r_{\text{draw}} &= \Pr\left[(n_i = n_j)\middle|\left(\varphi_t^{(i)},\varphi_t^{(j)}\right)\right] = \sum_{n=0}^{\infty} \frac{\varphi_t^{(i)n}}{n!} e^{-\varphi_t^{(i)}} \frac{\varphi_t^{(j)n}}{n!} e^{-\varphi_t^{(j)}} \\ &= e^{-(\varphi_t^{(i)}+\varphi_t^{(j)})} I_0\left(2\sqrt{\varphi_t^{(i)}\varphi_t^{(j)}}\right), \qquad (10) \end{aligned}$$

where $I_\nu(z) = (\frac{1}{2}z)^\nu \sum_{k=0}^{\infty}(\frac{1}{4}z^2)^k/[k!\Gamma(\nu+k+1)]$ is the modified Bessel function of the first kind. In case of a victory, the winning team has its potential incremented by its own potential divided by the total number of rounds in a tournament $k = 2(n-1)$, i.e., $\varphi^{(i)} \rightarrow \varphi^{(i)} + \varphi^{(i)}/2(n-1)$, and the team gains three points; the defeated team has its potential decremented by the equivalent quantity $\varphi^{(j)} \rightarrow \varphi^{(j)} - \varphi^{(j)}/2(n-1)$ and does not receive any points. In case of a draw, both teams gain one point and the potentials remain unchanged. In this way, we can write the average score obtained by team A at round t, when playing against team B, given that at round $(t-1)$, their potentials were $\varphi_{t-1}^{(A)}$ and $\varphi_{t-1}^{(B)}$, respectively, as

$$\begin{aligned} \left\langle x_t^{(A)} \middle| \varphi_{t-1}^{(A)};\varphi_{t-1}^{(B)} \right\rangle = {}& 3\left[1 - \exp\left(-\varphi_{t-1}^{(A)} - \varphi_{t-1}^{(B)}\right)\right. \\ & \left. \cdot I_0\left(2\sqrt{\varphi_{t-1}^{(A)}\varphi_{t-1}^{(B)}}\right)\right]\left(\frac{\varphi_{t-1}^{(A)}}{\varphi_{t-1}^{(A)} + \varphi_{t-1}^{(B)}}\right) \\ & + \exp\left(-\varphi_{t-1}^{(A)} - \varphi_{t-1}^{(B)}\right) I_0\left(2\sqrt{\varphi_{t-1}^{(A)}\varphi_{t-1}^{(B)}}\right). \end{aligned} \qquad (11)$$

The above recurrence formula does not allow one to solve the time evolution of scores $x_t^{(i)}$ analytically, because the model captures the non-Markovian features overseen by the mean-field approach. In order to obtain the studied quantities, namely, $\langle x(t)\rangle$, $\langle x^2(t)\rangle$, $\langle x^3(t)\rangle$, and $\sigma(t)$, we must resort to numerical computation by means of Monte Carlo simulations. In a previous publication [10], we showed that the statistics of accumulated final scores matches closely that of real tournaments (Brazilian, Italian, and Spanish). Here we are interested in whether the temporal evolution of scores in the model also follows that of real tournaments. That will be discussed in the next section.

IV. RESULTS

In Fig. 4, we illustrate the moments and standard deviations of the Brazilian and English tournaments obtained by the different proposed methods. The tournaments of the other countries analyzed in this work present similar behavior. Therefore they were omitted here for the sake of brevity. We can observe that although the mean field presents a good matching, the agent-based model (algorithm) follows the data from real tournaments more closely. Table III displays results of all coefficients and exponents for the real tournaments and the two methods described above (mean field and model) averaged over five seasons. The parameters $(*)_{\text{real}}$ correspond to averages among all years of the data previously estimated in Table I. The parameters $(*)_{\text{alg}}$ correspond to the ones obtained from the model (algorithm), in which case we consider the statistics of five sequential runs of our algorithm, seeded with the season 2006–2007 average number of goals as the initial condition. The parameters $(*)_{\text{MF}}$ correspond to the ones calculated with the discrete distribution $P_{\text{disc}}(x,t)$ given in Eq. (8) (using data for the 2006–2007 season for the mean draw probability r). Finally C_d and D_d, are calculated directly from the expressions $C_d = (r + 3(1-r)/2)$, $D_d = 3(1-r)/2$ derived earlier.

The results are very conclusive: all methods result in similar values for the coefficient C, indicating that it represents a general and fundamental behavior of the DRRS studied. On the other hand, the value of D is consistently overestimated by the mean-field approximation, whereas our algorithm predicts a value closer to the real estimates. Moreover and more importantly, our non-Markovian model provides a better approximation to the diffusive exponents, which characterize

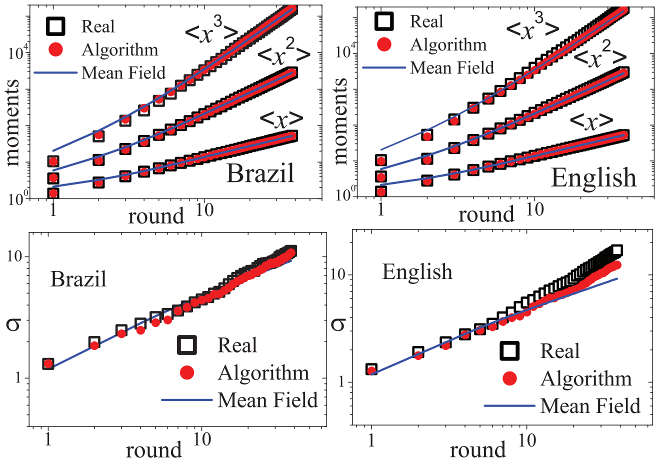

FIG. 4. (Color online) Moments and standard deviations of Brazilian and English tournaments. We can observe that although the mean field presents a good matching, the agent-based model (algorithm) follows the data from real tournaments more closely.

TABLE III. Coefficients and exponents obtained from real data, simulations (algorithm), mean-field (MF) approximation, and expressions $C_d = (r + 3(1-r)/2)$ and $D_d = 3(1-r)/2$ given in the text. The experimental ones for each country were obtained using its corresponding five seasons in the period 2007–2012. For the algorithm, five runs were simulated using as input the mean number of goals of the 20 teams in the 2006–2007 season. The mean-field approximation had as input r calculated for the same season.

Country	C_{real}	C_{alg}	C_{MF}	C_d	β_{real}	β_{alg}	β_{MF}	D_{real}	D_{alg}	D_{MF}	D_d	ξ_{real}	ξ_{alg}	ξ_{MF}
Brazil	1.37(4)	1.375(1)	1.364(2)	1.372	0.60(4)	0.593(9)	0.544(5)	1.28(17)	1.49(7)	1.70(5)	1.12	1.89(3)	1.962(7)	1.80(2)
Spain	1.38(4)	1.364(1)	1.363(2)	1.371	0.72(2)	0.680(1)	0.545(5)	1.18(5)	1.10(5)	1.70(5)	1.11	1.91(2)	1.957(4)	1.80(2)
France	1.35(4)	1.358(1)	1.340(2)	1.350	0.66(3)	0.583(6)	0.542(5)	1.25(10)	1.46(6)	1.60(5)	1.05	1.88(2)	1.919(6)	1.80(2)
England	1.36(5)	1.364(1)	1.363(2)	1.371	0.74(4)	0.636(8)	0.543(5)	1.02(11)	1.22(4)	1.70(5)	1.11	1.89(2)	1.933(6)	1.80(2)

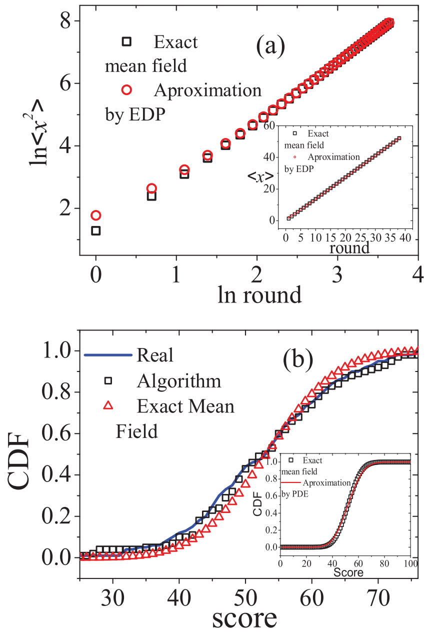

FIG. 5. (Color online) (a) Comparison of the time evolution of $\langle x^2 \rangle$ via an exact mean field [solution of finite difference Eq. (1)] and via solution of PDE [Eqs. (7) and (8)]. The inset plot shows the same results for the first moment $\langle x \rangle$. (b) This figure illustrates the cumulative distribution function (CDF) of scores by using real data, agent-based model (algorithm), and mean-field solution via finite difference. The inset plot shows that CDF via PDE corroborates the exact solution of the mean field via finite difference.

the kind of diffusive process. The mean-field approximation predicts a diffusive exponent $\beta_{\text{MF}}^{\text{PDE}} \approx 0.54$, very close to the one expected for normal diffusion, $\beta = 0.5$, which is obtained by directly solving the finite difference equation, Eq. (1), with a precision of $O(10^{-15})$. On the other hand, discarding the first 10 rounds, we obtain from the solution of the PDE [Eq. (4)], $\beta_{\text{MF}}^{\text{PDE}} = 0.5008$ with a precision of $O(10^{-4})$, which shows that the mean-field regime yields a diffusive process different from real data and the non-Markovian model.Anyway, the mean-field regime yields an excellent estimate of C, and reasonable estimates of D and ξ, capturing the essence of real championships. Figure 5(a) compares the time evolutions of $\langle x^2 \rangle$ and $\langle x \rangle$(plotted in the inset) as functions of t obtained from the solutions of the the exact finite difference equation and from the solution of the PDE, Eq. (4), by fixing $r = 0.2552$ (extracted from the Brazilian championship).

In Fig. 5(b) we illustrate a comparison among three cumulative distributions of scores in the 38th round: (1) from real data from the same five seasons of the Brazilian tournament previously used in this paper, (2) from the exact mean-field difference equation, and (3) from five generated seasons obtained from the non-Markovian algorithm. The inset plot illustrates the comparison between the two mean-field solutions (via PDE and via finite difference). The reader can observe that numerical solution of the related PDE corroborates the solution of the finite difference equation, showing that it is an acceptable approximation for our purpose. However, the real data suggest a case of superdiffusion, a feature which is captured by our non-Markovian model with a smaller percentage error. Moreover, such behavior seems to

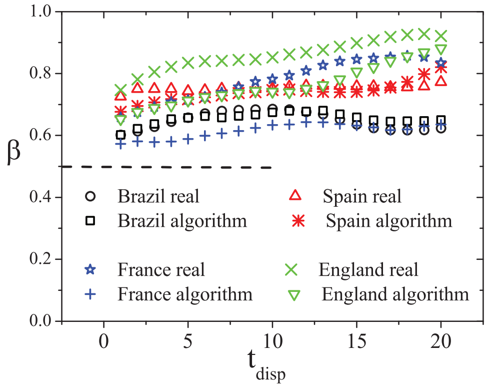

FIG. 6. (Color online) Behavior of exponent β as a function of the number of discarded rounds (t_{disp}) obtained from real data and by agent-based model (algorithm). In this case the fit is performed for the subsequent $2n-1-t_{\mathrm{disp}}$ rounds.

persist even when the exponent is calculated discarding the first rounds (large times). In Fig. 6 we can clearly observe that superdiffusivity is robust even when we use the $2n-1-t_{\mathrm{disp}}$ points to perform the linear fit in order to estimate β in $\ln \sigma(t)$ versus $\ln t$. All of our estimates calculated from real data or from non-Markovian modeling predictions indicate that $\beta > 0.5$ independently from t_{disp} first steps discarded, with a small tendency of growing as t_{disp} grows. Recently Ribeiro *et al.* [13] explored results from cricket, obtaining time series of up to 200 steps, in which they found exponents $\beta \approx 0.65$, a figure close to the one obtained in this work for soccer time series.

Here it is important to briefly return to a discussion about the Pawula theorem [11]. First, we observe the effects predicted by this theorem in the solution of the PDE in the mean-field regime. For example, we find $P(x=1,t=38) = -3.7 \times 10^{-12}$ in a point where a very small probability is expected. Rounding this value to zero introduces a negligible error. For the subsequent probability $[P(x=2,t=38) = 8.4 \times 10^{-12}]$ and all others ($x \geqslant 2$) we find positive values that closely match the distribution expected from the solution of the finite difference equation [Fig. 5(b)].

Finally, we explore the effect of coefficient E in Eq. (4). Is this coefficient necessary for the mean-field description? To answer this we present the distribution of scores obtained by the mean-field approximations with and without the third derivative term. Besides we compare both approaches with real data and agent-based model results. For the sake of simplicity, we choose one tournament, the French one and only two rounds: eighth and 38th). The other tournaments and rounds lead to the same conclusion.

Figure 7 shows the score distribution of the eighth round in log scale and the 38th in linear scale to better appreciate the differences among the distributions. The importance of E in the mean-field approximation is well evident; the mean-field with the $E=0$ curve clearly departs from the others three curves, which are close to each other, indicating that the third derivative term in Eq. (4) has an important role in the diffusive process.

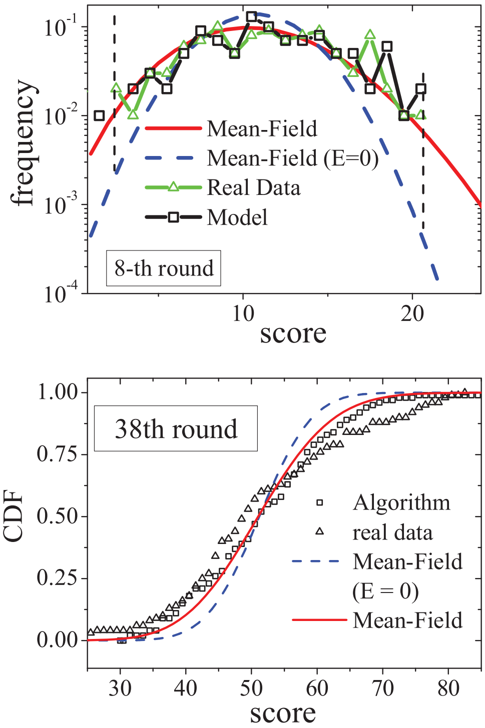

FIG. 7. (Color online) CDF in different rounds. A study of importance of E for the mean-field approximation.

V. CONCLUSIONS

Our agent-based model successfully mimics the anomalous superdiffusive behavior ($\sigma \sim t^{\beta}, \beta > 1/2$) of real soccer tournaments, while the mean-field analysis gives a partial description of them. Indeed, our model reproduces the statistical fluctuations in the time evolution of the scoring process of teams in soccer tournaments, which is here synthesized by two parameters related to the diffusion process: the drift speed C and the diffusion coefficient D.

ACKNOWLEDGMENT

The authors are partly supported by the Brazilian Research CNPq. We would like to thank the anonymous referees for carefully reading our manuscript.

[1] M. Hopkin, News of Nature, doi: 10.1038/news060612-17 (2006).

[2] A. Edmans, D. Garcia, and O. Norli, Sports Sentiment and Stock Returns: Sixteenth Annual Utah Winter Finance Conference, EFA 2005 Moscow Meetings. Available at SSRN: doi: 10.2139/ssrn.677103.

[3] R. N. Onody and P. A. de Castro, Phys. Rev. E **70**, 037103 (2004).

[4] L. Malacarne and R. Mendes, Physica A **286**, 391 (2000).

[5] A. Heuer, C. Muller, O. Rubner, N. Hagemann, and B. Strauss, PloS ONE **6**, e17664 (2011).

[6] G. Skinner and G. Freeman, J. Appl. Stat. **36**, 1087 (2009).

[7] A. Heuer, C. Muller, and O. Rubner, Europhys. Lett. **89**, 38007 (2010).

[8] D. Forrest and R. Simmons, Int. J. Forecast. **16**, 317 (2000).

[9] J. Giles, News of Nature, doi: 10.1038/news060605-10 (2006).

[10] R. da Silva, M. H. Vainstein, L. C. Lamb, and S. D. Prado, Comput. Phys. Commun. **184**, 661 (2013).

[11] H. Risken, *The Fokker-Planck Equation Methods of Solution and Applications* (Springer Verlag, Berlin, 1989).

[12] M. H. Vainstein and J. M. Rubí, Phys. Rev. E **75**, 031106 (2007).

[13] H. V. Ribeiro, S. Mukherjee, and Xiao Han T. Zeng, Phys. Rev. E **86**, 022102 (2012).