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 \( i \) 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...
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 \( \sigma(t) = D t^{\beta} \).

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 \( \gamma \) 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 \( \langle x(t) \rangle = C t, \sigma(t) = D t^{\beta}, \langle x^2(t) \rangle \propto t^\gamma, \) and \( \langle x^3(t) \rangle = E^* t^\gamma \) to data from real tournaments.

<table>
<thead>
<tr>
<th></th>
<th>Brazil</th>
<th>Spain</th>
<th>France</th>
<th>England</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>1.37(2)</td>
<td>1.38(2)</td>
<td>1.34(2)</td>
<td>1.36(2)</td>
</tr>
<tr>
<td>2008</td>
<td>1.38(2)</td>
<td>1.38(2)</td>
<td>1.34(2)</td>
<td>1.36(2)</td>
</tr>
<tr>
<td>2009</td>
<td>1.37(2)</td>
<td>1.38(2)</td>
<td>1.34(2)</td>
<td>1.36(2)</td>
</tr>
<tr>
<td>2010</td>
<td>1.35(2)</td>
<td>1.40(2)</td>
<td>1.33(2)</td>
<td>1.35(2)</td>
</tr>
<tr>
<td>2011</td>
<td>1.36(2)</td>
<td>1.37(2)</td>
<td>1.35(2)</td>
<td>1.38(2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Brazil</th>
<th>Spain</th>
<th>France</th>
<th>England</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C ) Coefficients ( E^* )</td>
<td>( C ) Exponents ( \beta )</td>
<td>( \gamma )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brazil</td>
<td>Spain</td>
<td>France</td>
<td>England</td>
<td>Brazil</td>
</tr>
<tr>
<td>2007</td>
<td>1.37(2)</td>
<td>1.38(2)</td>
<td>1.34(2)</td>
<td>1.36(2)</td>
</tr>
<tr>
<td>2008</td>
<td>1.38(2)</td>
<td>1.38(2)</td>
<td>1.34(2)</td>
<td>1.36(2)</td>
</tr>
<tr>
<td>2009</td>
<td>1.37(2)</td>
<td>1.38(2)</td>
<td>1.34(2)</td>
<td>1.36(2)</td>
</tr>
<tr>
<td>2010</td>
<td>1.35(2)</td>
<td>1.40(2)</td>
<td>1.33(2)</td>
<td>1.35(2)</td>
</tr>
<tr>
<td>2011</td>
<td>1.36(2)</td>
<td>1.37(2)</td>
<td>1.35(2)</td>
<td>1.38(2)</td>
</tr>
</tbody>
</table>
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

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

(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:

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

(2)

The expression in the last term can be written as

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

$$+ 3 \Delta P_m(n - 1),$$

(3)

where $\Delta P_m(n) = P_m(n) - P_m(n - 1)$ is the first finite difference and $\Delta^{(2)} P_m(n) = \Delta P_m(n) - \Delta P_m(n - 1) = P_m(n) + P_m(n - 2) - 2 P_m(n - 1)$ 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)

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

(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 $\eta_i$ 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 $\eta_i$. 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 $\eta_i$ 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.

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

\begin{center}
\begin{tabular}{|c|c|c|c|c|}
\hline
Statistics & Regular fit & $\eta_i = 0$ & First half & Second half \\
\hline
Average & 1.366 & 1.364 & 1.362 & 1.364 & -0.036 \\
Standard & 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 \\
\hline
\end{tabular}
\end{center}
\( \hat{\mu} = p_{\text{draw}} + 3p_{\text{win}} = (3 - r)/2 \), 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 \( 1.365 < C < 1.37 \). 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 obtained by the general linear fit. In this case we can observe the independence of \( c_i \) in the third and fourth columns in Table II. We have similar values when \( r < 1/3 \), corresponding to \( 1.365(19) \) corresponding to the two different time intervals we considered linear and 1.365(19) corresponding to the two different time intervals.

We can obtain a solution of the previous PDE by using the Fourier transform method. Defining \( \hat{P}(k,t) = \int_{-\infty}^{\infty} P(x,t)e^{ikx} \, dx \), if we multiply Eq. (4) by \( \exp(ikx) \) and integrate by parts supposing that \( P(\pm \infty,t) = \frac{4}{\pi} P(\pm \infty,t) = \frac{e^{ikx}}{i\pi} \), we obtain

\[
\frac{d}{dt} \hat{P}(k,t) = (i\hat{C}k - Dk^2 - iE_k^3)\hat{P}(k,t),
\]

for which the solution is \( \hat{P}(k,t) = \hat{P}(k,0)\exp((i\hat{C}k - Dk^2 - iE_k^3)t) \), where \( \hat{P}(k,0) \) 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

\[
P(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i(\hat{C}k - Dk^2 - iE_k^3)y} \, dk \int_{-\infty}^{\infty} e^{ik(y-x)} P(y,0) \, dy.
\]

(6)

All teams start with zero points, so \( P(y,0) = \delta(y) \); then, since \( \cos(\hat{C}k - E_k^3) \) is an even function while \( \sin(\hat{C}k - E_k^3) \) is an odd function of \( k \), we obtain a closed form for the distribution:

\[
P(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \cos((\hat{C}t - E_k^3)k) e^{-Dk^2/2} \, dk.
\]

(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

\[
P_{\text{disc}}(x,t) = \frac{P(x,t)}{\sum_{j=0}^{6(n-1)} P(j,t)}.
\]

(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 \( P_{\text{disc}}(x,t) \), 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

\[
\Pr(i > j, t) = \left(1 - r_{\text{draw}}^{(i,j)}(t)\right) \frac{\psi_i^{(i)}}{\psi_i^{(i)} + \psi_j^{(j)}},
\]

(9)

where \( \psi_i^{(i)} \) is the potential of the \( i \)th team 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 \( \psi_{0}^{(i)} \) is equated to the average number of goals per game of the team classified in the \( i \)th 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 \( \psi_i^{(i)} \) and \( \psi_j^{(j)} \). 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

\[
r_{\text{draw}} = \Pr[(n_i = n_j)|(\psi_i^{(i)}, \psi_j^{(j)})] = \sum_{n=0}^{\infty} \frac{\psi_i^{(i)} n! e^{-\psi_i^{(i)}}}{n!} \frac{\psi_j^{(j)} n! e^{-\psi_j^{(j)}}}{n!} = e^{-\psi_i^{(i)} - \psi_j^{(j)}} I_0(2\sqrt{\psi_i^{(i)} \psi_j^{(j)}},
\]

(10)

where \( I_0(z) = \frac{1}{\pi} \sum_{k=0}^{\infty} \frac{(\frac{1}{2}z)^{2k}}{k!(v + k + 1)} \) 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., \( \psi_i^{(i)} \rightarrow \psi_i^{(i)} + \psi_i^{(i)}/(2n - 1) \), and the team gains three points; the defeated team has its potential decremented by the equivalent quantity \( \psi_j^{(j)} \rightarrow \psi_j^{(j)} - \psi_j^{(j)}/(2n - 1) \) 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 \( \psi_{i-1}^{(A)} \) and \( \psi_{j-1}^{(B)} \), respectively, as

\[
\langle x_t^{(A)} | x_{t-1}^{(A)}; y_{t-1}^{(B)} \rangle = 3\left[1 - \exp\left(-\psi_{i-1}^{(A)} - \psi_{j-1}^{(B)}\right)\right] I_0(2\sqrt{\psi_{i-1}^{(A)} \psi_{j-1}^{(B)}}) \\
\cdot \left(\frac{\psi_{i-1}^{(A)}}{\psi_{i-1}^{(A)} + \psi_{j-1}^{(B)}}\right) + \exp\left(-\psi_{i-1}^{(A)} - \psi_{j-1}^{(B)}\right) I_0(2\sqrt{\psi_{i-1}^{(A)} \psi_{j-1}^{(B)}}).
\]

(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 \(*\)_{real} correspond to averages among all years of the data previously estimated in Table I. The parameters \(*\)_{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 \( C_d = (r + 3(1 - r))/2 \) and \( D_d = 3(1 - r)/2 \) 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 diffuse exponents, which characterize
TABLE III. Coefficients and exponents obtained from real data, simulations (algorithm), mean-field (MF) approximation, and expressions $C_d = \theta + (3 - \theta)/2$ and $D_d = (3 - \theta)/2$
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.

<table>
<thead>
<tr>
<th>Country</th>
<th>$C_{\text{real}}$</th>
<th>$D_{\text{real}}$</th>
<th>$\beta_{\text{MF}}$</th>
<th>$\xi_{\text{real}}$</th>
<th>$r_{\text{real}}$</th>
<th>$C_{\text{alg}}$</th>
<th>$D_{\text{alg}}$</th>
<th>$\beta_{\text{alg}}$</th>
<th>$\xi_{\text{alg}}$</th>
<th>$C_{\text{MF}}$</th>
<th>$D_{\text{MF}}$</th>
<th>$\beta_{\text{MF}}$</th>
<th>$\xi_{\text{MF}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>1.37(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
</tr>
<tr>
<td>Spain</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
<td>1.36(4)</td>
</tr>
<tr>
<td>France</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
</tr>
<tr>
<td>England</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
<td>1.35(4)</td>
</tr>
</tbody>
</table>

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 $\beta_{\text{MF}} \approx 0.54$, 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)], $\beta_{\text{PDE}} = 0.5008$ 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 $\xi$, 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 $\beta$ as a function of the number of discarded rounds ($t_{\text{disp}}$) obtained from real data and by agent-based model (algorithm). In this case the fit is performed for the subsequent $2n - 1 - t_{\text{disp}}$ 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 $2n - 1 - t_{\text{disp}}$ points to perform the linear fit in order to estimate $\beta$ 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 $t_{\text{disp}}$ first steps discarded, with a small tendency of growing as $t_{\text{disp}}$ 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 ($x \geq 2$) 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.

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.