OFFPRINT

Self-affirmation model for football goal distributions

E. Bittner, A. Nussbaumer, W. Janke and M. Weigel

EPL, 78 (2007) 58002

Please visit the new website
www.epljournal.org
Europhysics Letters (EPL) has a new online home at www.epljournal.org

Take a look for the latest journal news and information on:

- reading the latest articles, free!
- receiving free e-mail alerts
- submitting your work to EPL

www.epljournal.org
Self-affirmation model for football goal distributions

E. Bittner\(^1\), A. Nußbaumer\(^1\), W. Janke\(^1\) and M. Weigel\(^2\)

\(^1\) Institut für Theoretische Physik and Centre for Theoretical Sciences (NTZ), Universitäät Leipzig
Postfach 100 920, D-04009 Leipzig, Germany

\(^2\) Department of Mathematics and the Maxwell Institute for Mathematical Sciences, Herriot-Watt University
Riccarton, Edinburgh, EH14 4AS, UK

received 14 February 2007; accepted in final form 12 April 2007
published online 16 May 2007

PACS 89.20.-a – Interdisciplinary applications of physics
PACS 02.50.-r – Probability theory, stochastic processes, and statistics

Abstract – Analyzing football score data with statistical techniques, we investigate how the
highly co-operative nature of the game is reflected in averaged properties such as the distributions
of scored goals for the home and away teams. It turns out that in particular the tails of the
distributions are not well described by independent Bernoulli trials, but rather well modeled by
negative binomial or generalized extreme value distributions. To understand this behavior from
first principles, we suggest to modify the Bernoulli random process to include a simple component
of self-affirmation which seems to describe the data surprisingly well and allows to interpret the
observed deviation from Gaussian statistics. The phenomenological distributions used before can
be understood as special cases within this framework. We analyzed historical football score data
from many leagues in Europe as well as from international tournaments and found the proposed
models to be applicable rather universally. In particular, here we compare men’s and women’s
leagues and the separate German leagues during the cold war times and find some remarkable
differences.

Copyright © EPLA, 2007

Football (soccer) is one of the most popular sports world-wide, attracting millions of spectators each year. Its popularity and economical importance also captivated scientists from many fields, for instance in the attempt to improve the game tactics, etc. Much less effort has been devoted, it seems, to the understanding of football (and other ball sports) from the perspective of the stochastic behavior of co-operative “agents” (i.e., players) in abstract models. Such problems recently have come into the focus of physicists in the hope that the model-based point-of-view and methodological machinery of statistical mechanics might add a new perspective to the much more detailed investigations of more specific disciplines [1,2]. Some reports of such research are collected in ref. [3]. Score distributions of ball games have been occasionally considered by statisticians [4–7]. Very small data sets were initially found to be reasonably well described by the simplest Poissonian model resulting from constant and independent scoring probabilities [4]. Including more data, however, better phenomenological fits were achieved with models such as the negative binomial distribution (NBD), which can be constructed from a mixture of independent Poissonian processes [6], or even with models of generalized extreme value (GEV) statistics [7,8], which are particularly suited for heavy-tailed distributions. This yielded a rather inhomogeneous and purely phenomenological picture, without offering any microscopical justification. We argue that the crucial ingredient missed in previous studies are the correlations between subsequent scoring events.

In a broader context, this problem of extremes is of obvious importance, for instance, in actuarial mathematics and engineering, but the corresponding distributions with fat tails also occur in many physics fields, ranging from the statistical mechanics of regular and disordered systems [9–12] over turbulence [13] to earthquake data [14]. In these cases often average properties were considered instead of explicit extremes, and the empirical occurrence of heavy-tailed distributions led to speculations about hidden extremal processes, most of which could not be identified, though. It was only realized recently that GEV distributions can also arise naturally as the statistics of sums of correlated random variables [15], which could explain their ubiquity in nature.

For the specific example of scoring in football, correlations naturally occur through processes of feedback of...
scoring on both teams, and we shall see how the intro-
duction of simple rules for the adaptation of the success
probabilities in a modified Bernoulli process upon scoring
gain leads to systematic deviations from Gaussian statis-
tics. We find simple models with a single parameter of self-
affirmation to best describe the available data, including
cases with relatively poor fits of the NBD. The latter is
shown to result from one of these models in a particular
limit, explaining the relatively good fits observed before.

To investigate the importance of correlations, we con-
sider the distributions of goals scored by the home
away teams in football league or cup matches. To
determine the simplest possible approximation, both teams have
possibility to score in each unit with a probability
\( p = p(n) \)
depending on the number \( n \) of goals scored so far. Several
possibilities arise. For our model “A”, upon each goal the
scoring probability is modified as \( p(n) = p(n - 1) + \kappa \), with
some fixed constant \( \kappa \). Alternatively, one might consider
a multiplicative modification rule, \( p(n) = kp(n - 1) \), which
we refer to as model “B”. Finally, in our model “C” the
assumption of independence of the scoring of the two
teams is relaxed by coupling the adaptation rules, namely
by setting \( p_h(n) = p_h(n - 1)\kappa_h, p_a(n) = p_a(n - 1)\kappa_a \)
on a goal of the home (h) team, and vice versa for an away (a)
goal. If both teams have \( \kappa > 1 \), this results in an incentive
for the scoring team and a demotivation for the opponent,
but a value \( \kappa < 1 \) is conceivable as well. The resulting,
distinctly non-Gaussian distributions \( P_N(n) \) for the total
number of goals scored by one team can be computed
exactly for models “A” and “B” from a Pascal recurrence
relation [17],

\[
P_N(n) = [1 - p(n)]P_{N-1}(n) + p(n)P_{N-1}(n - 1),
\]

where \( p = p(n) \) is the probability to score a goal
in a particular unit for the scoring team, and \( p(n) \) is the
probability to score a goal in a particular unit for the
opposing team.

The NBD form has been found to describe football
score data rather well [6,7]. It appears rather ad hoc,
however, to assume that \( f(\lambda) \) follows a gamma form, and
fitting different seasons of our data with the Poissonian
model (1), the resulting distribution of \( \lambda \) does not resemble
a gamma distribution. As a phenomenological alternative
to the NBD, Greenhough et al. [7] considered fits of the
GEV distributions

\[
P_{\xi,\mu,\sigma}(n) = \frac{1}{\sigma} \left( 1 + \frac{n - \mu}{\sigma} \right)^{-1-\frac{1}{\xi}} e^{-\frac{(1+\frac{n-\mu}{\sigma})^{-1-\frac{1}{\xi}}}{\xi}}
\]
to the data, obtaining clearly better fits than with the
NBD in some cases. Depending on the value of the
parameter \( \xi \), these distributions are known as Weibull
(\( \xi < 0 \)), Gumbel (\( \xi \rightarrow 0 \)) and Fréchet (\( \xi > 0 \)) distributions, respectively [8].

In the present context of scoring in football, goals
are likely not independent events but, instead, scoring
certainly has a profound feedback on the motivation
and possibility of subsequent scoring of both teams (via
direct motivation/demotivation of the players, but also,
e.g., by a strengthening of defensive play in case of a
lead). Such feedback can be taken into account starting from
a simple Bernoulli model: consider a match divided
into, e.g., \( N = 90 \) time steps with both teams having the
possibility to score in each unit with a probability
\( p = p(n) \) depending on the number \( n \) of goals scored so far. Several
possibilities arise. For our model “A”, upon each goal the
scoring probability is modified as \( p(n) = p(n - 1) + \kappa \), with
some fixed constant \( \kappa \). Alternatively, one might consider
a multiplicative modification rule, \( p(n) = kp(n - 1) \), which
we refer to as model “B”. Finally, in our model “C” the
assumption of independence of the scoring of the two
teams is relaxed by coupling the adaptation rules, namely
by setting \( p_h(n) = p_h(n - 1)\kappa_h, p_a(n) = p_a(n - 1)\kappa_a \)
on a goal of the home (h) team, and vice versa for an away (a)
goal. If both teams have \( \kappa > 1 \), this results in an incentive
for the scoring team and a demotivation for the opponent,
but a value \( \kappa < 1 \) is conceivable as well. The resulting,
distinctly non-Gaussian distributions \( P_N(n) \) for the total
number of goals scored by one team can be computed
exactly for models “A” and “B” from a Pascal recurrence
relation [17],

\[
P_N(n) = [1 - p(n)]P_{N-1}(n) + p(n)P_{N-1}(n - 1),
\]

where \( p = p(n) \) is the probability to score a goal
in a particular unit for the scoring team, and \( p(n) \) is the
probability to score a goal in a particular unit for the
opposing team.
Table 1: Fits and their $\chi^2$ per degree-of-freedom, $\tilde{\chi}^2 = \chi^2/d.o.f.$, of the phenomenological distributions (1), (2), and (3) as well as fits of our microscopic feedback models “A” and “B” to the data for the East German “Oberliga”, the (West) German men’s premier league “Bundesliga”, the German women’s premier league “Frauen-Bundesliga” and the qualification stages of all past “FIFA World Cups”.

<table>
<thead>
<tr>
<th></th>
<th>Oberliga</th>
<th>Bundesliga</th>
<th>Frauen-Bundesliga</th>
<th>FIFA World Cup</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Home</td>
<td>Away</td>
<td>Home</td>
<td>Away</td>
</tr>
<tr>
<td>Poisson</td>
<td>$\lambda$</td>
<td>1.85(2)</td>
<td>1.05(1)</td>
<td>1.91(1)</td>
</tr>
<tr>
<td>$\tilde{\chi}^2$</td>
<td>12.5</td>
<td>12.8</td>
<td>9.21</td>
<td>9.13</td>
</tr>
<tr>
<td>NBD</td>
<td>$p$</td>
<td>0.17(1)</td>
<td>0.14(1)</td>
<td>0.11(1)</td>
</tr>
<tr>
<td></td>
<td>$r$</td>
<td>9.06(88)</td>
<td>6.90(84)</td>
<td>16.2(1.9)</td>
</tr>
<tr>
<td></td>
<td>$p_0$</td>
<td>0.0191</td>
<td>0.0112</td>
<td>0.0202</td>
</tr>
<tr>
<td></td>
<td>$\kappa$</td>
<td>0.0021</td>
<td>0.0016</td>
<td>0.0012</td>
</tr>
<tr>
<td>$\tilde{\chi}^2$</td>
<td>0.99</td>
<td>4.09</td>
<td>1.08</td>
<td>2.22</td>
</tr>
<tr>
<td>GEV</td>
<td>$\xi$</td>
<td>$-0.05(1)$</td>
<td>0.02(1)</td>
<td>$-0.10(1)$</td>
</tr>
<tr>
<td></td>
<td>$\mu$</td>
<td>1.12(2)</td>
<td>0.49(2)</td>
<td>1.17(2)</td>
</tr>
<tr>
<td></td>
<td>$\sigma$</td>
<td>1.30(2)</td>
<td>0.90(2)</td>
<td>1.33(1)</td>
</tr>
<tr>
<td>$\tilde{\chi}^2$</td>
<td>1.93</td>
<td>5.04</td>
<td>3.43</td>
<td>7.95</td>
</tr>
<tr>
<td>Model “A”</td>
<td>$p_0$</td>
<td>0.0188(2)</td>
<td>0.0112(1)</td>
<td>0.0199(2)</td>
</tr>
<tr>
<td></td>
<td>$\kappa$</td>
<td>0.0024(2)</td>
<td>0.0018(2)</td>
<td>0.0015(1)</td>
</tr>
<tr>
<td>$\tilde{\chi}^2$</td>
<td>1.07</td>
<td>4.23</td>
<td>1.01</td>
<td>2.21</td>
</tr>
<tr>
<td>Model “B”</td>
<td>$p_0$</td>
<td>0.0189(2)</td>
<td>0.0112(1)</td>
<td>0.0200(2)</td>
</tr>
<tr>
<td></td>
<td>$\kappa$</td>
<td>1.1115(83)</td>
<td>1.153(15)</td>
<td>1.0679(60)</td>
</tr>
<tr>
<td>$\tilde{\chi}^2$</td>
<td>0.75</td>
<td>3.35</td>
<td>1.25</td>
<td>1.96</td>
</tr>
</tbody>
</table>

$\tilde{\chi}^2 = \chi^2/d.o.f.$, naturally taking into account the different numbers of free parameters in the fits considered.

We first considered fits of the PDFs of the phenomenological descriptions (1), (2), and (3). Not to our surprise, and in accordance with previous findings [6,7], the simple Poissonian ansatz (1) is not found to be an adequate description for any of the data sets. Deviations occur mainly in the tails with large numbers of goals which in general are found to be rarer than can be accommodated by a Poissonian model. On the contrary, the NBD form (2) models all of the above data well as is illustrated in fig. 1. Considering the fits of the GEV distributions (3), we find that extreme value statistics are in general also a reasonably good description of the data. The shape parameter $\xi$ is always found to be small in modulus and negative in the majority of the cases, indicating a deviation of the Weibull type (which is in agreement with the findings of ref. [7] for different leagues). Fixing $\xi = 0$ yields overall clearly larger $\chi^2$ values. Comparing “Oberliga” and “Bundesliga”, we consistently find larger values of the parameter $\xi$ for the former, indicative of the comparatively fatter tails of these data, see table 1 and fig. 1. Comparing to the results for the NBD, we do not find any cases where the GEV distributions would provide the best fit to the data, so clearly the leagues considered here are not of the type for which Greenhough et al. [7] found better matches with the GEV statistics than for the NBD. Similar conclusions hold true for the comparison of “Bundesliga” and “Frauen-Bundesliga”, with the latter taking on the role of the “Oberliga”.

Representing the continuum limit of our model “A”, the good performance of the NBD fits observed so far implies that the feedback models proposed here can indeed capture the main characteristics of the game. To test this conjecture directly we performed fits of the exact distributions resulting from the recurrence relation (4), employing the simplex method to minimize the total $\chi^2$ deviation for the home and away scores. Comparing the results of model “A” to the fits of the limiting NBD, we observe in table 1 almost identical fit qualities for the final scores. However, for sums and differences of scores we find a considerably better description by using our model “A”, indicating deviations from the continuum limit there [17]. The overall best modeling of the league data is achieved with fits of model “B” which feature on average an even higher quality than those of model “A”, cf. table 1. We also performed fits to the more elaborate model “C”, but found the results rather similar to those of the simpler model “B” and hence do not discuss them here.

Comparing the leagues, we see in table 1 that the parameters $\kappa$ for the “Oberliga” are significantly larger than for the “Bundesliga”, whereas the parameters $p_0$ are slightly smaller for the “Oberliga”. That is to say, scoring a goal in a match of the East German “Oberliga” was
We have shown that football score data can be understood from a certain class of modified binomial models with a built-in effect of self-affirmation of the teams upon scoring a goal. The NBD fitting many of the data sets can in fact be understood as a limiting distribution of our model “A” with an additive update rule of the scoring probability. It does not provide very good fits in cases with heavier tails, such as the qualification round of the “FIFA World Cup” series. The overall best variant is our model “B”, where a multiplicative feedback model “B”, on the other hand, also handles this case extremely well and, for the away team, considerably better than the GEV distribution (3). The difference to the league data can be attributed to the possibly very large differences in skill between the opposing teams occurring since all countries are allowed to participate in the qualification round. The parameters in table 1 reveal a remarkable similarity with the parameters of the “Frauen-Bundesliga”, where a similar explanation appears quite plausible since the very good players are concentrated in just two or three teams.

Finally, we also considered the score data of the qualification stage of the “FIFA World Cup” series from 1930 to 2002 (≈3400 matches)\(^3\)\(^4\). Compared to the domestic league data discussed above, the results of the World Cup show distinctly heavier tails, cf. fig. 1. Consequently we obtain good fits for the heavy-tailed distributions, and, in particular, in this case the GEV distribution provides a better fit than the NBD, similar to what was found by Greenhough et al. [7], cf. table 1. The fits of model “A” are again rather similar to the NBD. The multiplicative feedback model “B”, on the other hand, also handles this case extremely well and, for the away team, considerably better than the GEV distribution (3). The difference to the league data can be attributed to the possibly very large differences in skill between the opposing teams occurring since all countries are allowed to participate in the qualification round. The parameters in table 1 reveal a remarkable similarity with the parameters of the “Frauen-Bundesliga”, where a similar explanation appears quite plausible since the very good players are concentrated in just two or three teams.

![Fig. 1: Histograms of final scores of home and away teams and corresponding fits.](image)

(a) East German “Oberliga” (b) (West) German “Bundesliga”. (c) The qualification stage of the “FIFA World Cup” series.

**Table 1: Model Fits for the Different Data Sets**

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Model “A”</th>
<th>Model “B”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oberliga</td>
<td>NBD</td>
<td>GEV</td>
</tr>
<tr>
<td>Bundesliga</td>
<td>NBD</td>
<td>GEV</td>
</tr>
<tr>
<td>FIFA World Cup</td>
<td>NBD</td>
<td>GEV</td>
</tr>
</tbody>
</table>

We disregarded all games played in tournaments on neutral grounds. The final knockout stage follows different rules [17].

\(^3\)www.rdasilva.demon.co.uk/football.html.

\(^4\)We disregarded all games played in tournaments on neutral grounds.
progresses, thus supporting the presence of feedback as discussed here.

***

The authors are grateful to O. Penrose and S. Zachary for discussions. This work was partially supported by the DFG under grant Nos. JA483/22-1 and 23-1 and the EC RTN-Network “ENRAGE”: Random Geometry and Random Matrices: From Quantum Gravity to Econophysics under grant No. MRTN-CT-2004-005616. MW acknowledges support by the EC MC-EIF programme under contract No. MEIF-CT-2004-501422.

REFERENCES