2 Professional Football and Bessel-Functions: A Statistical Analysis

by M. Tolan, M. Paulus, R. Fendt, J. Stolze
Fakultät Physik/DELTA, Technische Universität Dortmund,
Otto-Hahn-Straße 4, D-44227 Dortmund, Germany

Football is a simple game. The outcome of matches seems to be quite erratic and unpredictable when teams of similar strength and ability are playing. The aim of this work is to show that football results averaged over a large number of games can be explained surprisingly well in terms of a model leading to modified Bessel functions and based on the simple assumption that all teams within one league are of equal strength and that their scoring rates can be approximated by the same Poisson distributions. An analysis of 27000 games from the German 1st and 2nd Bundesliga, 5300 games from the English Premier League, and 2000 results from the Italian Serie A shows an almost perfect agreement with the model. Furthermore it is shown that the correlation between goals in football is very weak.

1 Introduction

Football is the world’s most popular sport with professional leagues in many countries. The football world championships in Germany 2006 and South Africa 2010 were the sports events which obtained the most widespread attention any such event ever had in history. Thus, an in-depth analysis of football results is of fundamental interest for all fans. Such an analysis will be provided by our work, where we have tested the most simple model for the football game by looking at a large number of matches from the English Premier League, the Italian Serie A, and the 1st and 2nd German Bundesliga.

As early as 1980 Jack Dowie [1] pointed out that there is a remarkable correlation between the number of draws and the average scoring rate of the teams of the English Premier League. Figure 1 shows such a plot for the 1st German football league (i.e. the 1st Bundesliga), where this (anti-)correlation becomes clearly visible. The data were collected from a public database available on the Web [2]. While the scoring rate drops from 3.3 goals per game in the seasons before the year 1985 to approximately 2.9 goals per game after the year 1990, one can see that the percentage of draws increases from 24% to 28% during these periods. This trend to more draws in the game due to lower scoring rates has been stopped in the year 1995 by introducing the three-point rule, i.e. by making it more attractive for teams to go for a win yielding three points instead of two as in the years before. Since Figure 1 suggests that there may be a deeper relationship between the frequency of draws and the average number of goals which are scored by the football teams, one may look for a model which is
Figure 1: (Anti-)Correlation between the average scoring rate $A$ per football match of a season (solid line) and the fraction of draws (broken line) in the first German professional football league. Note that in the year 1995 the “three points rule” was introduced (vertical broken line) leading to a decrease of the absolute numbers of draws.

able to explain averaged trends in football results.

2 Theory and discussion

A simple model for the football game assumes Poisson distribution for the scoring of a team. This may be justified since football, at least on a professional level, is a game where relatively few goals are scored. Thus, each team is treated as a radioactive source “emitting” goals independently according to Poisson statistics. Such a crude model has already been proposed in the literature by several authors [3–6]. Let $a$ be the average scoring rate of a team and $k$ be the number of goals scored in a single match then

$$p_a(k) = \frac{a^k}{k!} e^{-a}$$

defines the probability that this team scores exactly $k$ goals, if it scores $a$ goals on average. For the teams in the German 1$^{\text{st}}$ and 2$^{\text{nd}}$ Bundesliga the scoring rate averaged over all games played is $a \approx 1.5$ reflecting the fact that roughly 3 goals are shot at each game. In order to test this crude model, one can calculate the probability for an outcome of a football game with a goal difference of $z \geq 0$, where $z = 0$ means a draw. A simple calculation yields for this probability [6]

$$P_z(A) = (2 - \delta_{z,0}) \cdot e^{-A} \cdot I_z(A)$$

(2)
where $A = 2a$ is the number of goals scored on average in a single game, $I_z(A)$ is the modified Bessel function of order $z$, and $\delta_{z,0} = 1$ for $z = 0$ and $\delta_{z,0} = 0$ for $z \neq 0$. The factor of 2 for $z \neq 0$ takes into account that a goal difference of $z > 0$ can be achieved by the home or away team, respectively, while a draw, $z = 0$, is only counted once. Eq. (2) is the reason for the name “Bessel-function model of football” [7], which we will use throughout the rest of this paper. This model will now be compared with real data.

For the English Premier League $A = 2.612$ holds, if all 5300 end results of the 13 seasons are considered. If the number of goals are counted which have been scored up to a certain minute of the game, then average scoring rates between $A = 0$ and $A = 2.612$ may be realized. All results with an outcome of $z$ goals difference were counted and the relative percentages with respect to the total number of 5300 games were calculated. This has been done for $z = 0, 1, 2, 3$, using again the public football database mentioned before [2]. Figure 2 shows the results for the Premier League data.

The solid lines are calculated for $z = 0, 1, 2, 3, 4$ using Eq. (2). The symbols are the percentages of the results with $z$ goals difference as obtained from the football database. In all figures the statistical errors are smaller than the symbol size. It can be seen that the simple Bessel-function model of football is able to explain the data almost perfectly. This is surprising since the Bessel-function formula is based on the drastic assumption of equal scoring rates and Poisson statistics for all teams.
In order to define a quantitative measure, a “randomness factor” \( R \) of a football league is defined by

\[
R = \frac{1}{(z_{\text{max}} + 1)N}.
\]

where \( P_z(A_j) \) is given by Eq. (2), \( P_{\text{data}}(z, A_j) \) are the probabilities obtained from the database, \( A_j \) is the scoring rate after a certain amount of time has passed in the games, and \( z \geq 0 \) is the goal difference as used before. In our case \( z \) always runs from \( z = 0 \) to \( z = z_{\text{max}} = 4 \), and \( j \) from \( j = 1 \) to \( j = N = 14 \) since the goal differences were analyzed for 14 different times between 8 minutes and 90 minutes. It is obvious that \( R = 0 \) means a perfect agreement of the data with the Bessel-function model and \( R = 1 \) is the largest possible \( R \)-value. For the Premier League we found the small value of \( R_{\text{England}} = 8.6 \cdot 10^{-4} \) indicating the perfect agreement between the data and the model.

It is remarkable that “better teams”, which do clearly exist in football leagues, e.g. Manchester United which dominated many of the last 13 seasons in England, do not lead to strong systematic deviations from the Bessel-function model. In order to investigate the effect of “better teams” further, 2000 games of the last 6 seasons of the Italian Serie A have been analyzed in a similar way as described above. The results are shown in Figure 3.

Again, the Bessel-function model of football is able to explain the data very well. Only small systematic deviations for the \( z = 0 \) and the \( z = 1 \) results are visible. The \( R \)-factor for the Serie A is \( R_{\text{Italy}} = 12.7 \cdot 10^{-4} \). This is 32% larger than the \( R \)-factor for the Premier League, but still very small indicating also the excellent agreement between data and the model. However, since one finds that in the last 6 seasons basically five teams were dominating, namely Juventus Torino, AC Milan, Inter Milan, AS Roma, and Lazio Roma, which were always ranked among the top 6 teams, the assumption of equal strength is certainly not valid for the Italian Serie A. Thus, the formula given by Eq. (2) seems to be quite robust against such systematic trends in a football league.

This effect will be further analyzed by looking at the data of the German 1st and 2nd Bundesliga. Figure 4 shows the same plot as Figures 2 and 3, but now taking into account all 27000 games played so far from 1963 until November 2005 in the German first and second professional football leagues.

Here the larger average scoring rate of \( A = 3.058 \) is found, if only the end results are considered. All data were again obtained by analyzing results from the public football database [2]. The calculated \( R \)-factor is now \( R_{\text{Germany}} = 20.6 \cdot 10^{-4} \), i.e. more than a
factor of two larger than $R_{\text{England}} = 8.6 \cdot 10^{-4}$, but still quite small. Figure 4 shows that the Bessel-function model of football is able to explain the data perfectly until $A = 1.827$, which corresponds to the scoring rate until the 60th minute. Then for larger scoring rates, which means for increasing duration of the game, small systematic deviations from the model are visible. These deviations are most prominent for the number of results with one goal difference ($z = 1$). Here the Bessel-function formula Eq. (2) overestimates the total number of results, while all other goal differences including the number of draws are slightly underestimated. This systematic behavior may indicate that the model of independent Poisson distributions with the same scoring rate for each team is too simple for a full explanation, since football is not a purely erratic game. A reason for this is that teams change their tactics with increasing duration of the game. In Figure 4 the points for $A = 2.975$ (corresponds to 89 minutes of the game) and $A = 3.058$ (end results) show a remarkably large fluctuation in the case of $z = 0, 1, 2$. A similar but weaker fluctuation in the $z = 0$ and $z = 1$ results for the 89th and 90th minute is also visible in Figure 2 for the English Premier League. While the probability for a $z = 1$ goal difference drops at the end significantly the probability for a draw ($z = 0$) and a two goals difference is increasing accordingly. This reflects the fact, that for a team being only one goal behind, opening its defence in favor of increased attacking efforts to obtain a draw is a risk which is worth to take. However, this scenario enhances at the same time the probability of a goal by the opponent due to the weaker defence, thus increasing the number of $z = 2$ results. Such an effect is neither expected for $z = 0$ where both teams may be satisfied with a draw, nor for $z \geq 2$, where it is unlikely that a team may score two or more goals in the last minute. Of course, such effects are not
Figure 4: Comparison between the data and the Bessel-function model of football for the 1st and 2nd German Bundesliga for the 27,000 matches of all seasons played until November 2005 (see also Figures 2 and 3).

described by Eq. (2).

The Bessel-function model of football as given by Eq. (2) relies on the fact that all teams have the same scoring rate \( a \) leading to \( A = 2a \) goals on average in a single game. However, it is well-known that the home teams win more often than the away teams. Thus, a better approach is to choose the average scoring rates
\[
\begin{align*}
    a_{\text{home}} &= A(1 + h) / 2 \\
    a_{\text{away}} &= A(1 - h) / 2
\end{align*}
\]
for the home and away teams, respectively, with \( A = a_{\text{home}} + a_{\text{away}} \) being again the total number of goals in a single game and the “home advantage factor” \( h \geq 0 \). Assuming again Poisson statistics yields the result [6]

\[
P_z(A) = \left[ \frac{(1 + h)^{z/2}}{(1 - h)^{z/2}} + \left( \frac{1 + h}{1 - h} \right)^{-z/2} - \delta_{z,0} \right] \cdot e^{-A} \cdot I_z \left( A\sqrt{1 - h^2} \right) \quad (4)
\]

for the probability of an outcome of a football game with \( z \) goals difference. For \( h = 0 \) Eq. (4) is equal to Eq. (2). One could also speculate about a “correlation effect” by introducing a bivariate Poisson distribution as performed in reference [6]. In this model the scoring rate

\[
A = a_{\text{home}} + a_{\text{away}} + a_{\text{corr}} \quad (5)
\]

of a match between two teams consists of the two “pure” scoring rates \( a_{\text{home}} = A(1 + h)(1 - c) / 2 \) and \( a_{\text{away}} = A(1 - h)(1 - c) / 2 \) of the home and away team, respectively, and a correlated scoring part \( a_{\text{corr}} = c A \) with \( h \geq 0 \) being being again the home advantage factor, and \( c \geq 0 \) being the fraction of goals of a match which are correlated, i.e. this fraction of goals is not independent of the scoring rates of the two teams. Such a
correlation between goals exists clearly in basketball where the possession of the ball already implies the next point with a high probability. Since the ball possession goes to the other team after one team has scored, points in basketball are highly correlated which means that $c$ might be large. However, in football it is unlikely that a large correlation between the goals of the two teams exist and a small correlation factor is expected. It should be noted that the correlation parameter $c$ does not alter the result given by Eq. (4) [6]. It enters indirectly into the results via the modified scoring rate given by Eq. (5). A discussion of Eq. (4) shows that the introduction of the home advantage parameter $h > 0$ leads to a suppression of the number of $z = 1$ results compared with the case $h = 0$. The $z = 2$ curve is almost unaffected while the probability for $z \geq 3$ increases slightly. These effects are in qualitative agreement with the data shown for the German professional leagues. However, the number of draws ($z = 0$) drops significantly as compared with the $h = 0$ case, which is not observed in the data. This cannot be corrected by the correlation factor $c$, which only stretches the curves along the $A$-axis. A best fit of Eq. (4) to the data for the German Bundesliga leads to $h = 0.3$ and $c = 0.05$. A home advantage factor of $h = 0.3$ is reasonable if the data for the German professional leagues are checked [2]. The small value of $c$ indicates that correlations between goals do not play a significant role in football. The $R$-factor drops by introducing the parameters $h = 0.31$ and $c = 0.05$ from $R_{\text{Germany}} = 20.6 \cdot 10^{-4}$ to $R_{\text{Germany,refined}} = 16.2 \cdot 10^{-4}$. Thus, the refined model including home advantage and goal correlation does explain the data for the German professional football leagues slightly better than the “pure” Bessel-function model of football.

3 Summary

In summary, we have shown that the data for all investigated professional football leagues are very close to the theoretical curves given by the simple Bessel-function model of football. The presented refinements of this model due to the home advantage factor and possible goal correlations yield only minor corrections - the overall shapes of the curves are always dominated by the simple model. In view of this result, one may speculate if professional football players know details about the mathematics of (modified) Bessel functions. While it appears unlikely that they know the explicit formulas for the Bessel functions given in textbooks, the Figures 2, 3, and 4 clearly indicate that professional players at least obey the statistics given by them very well. The reason is that the Bessel-function model of football seems to be quite robust against fluctuations of the ability, i.e. the scoring rate, of the teams within such a professional league. Since fluctuations of team abilities become in general much larger when non-professional football leagues are considered, it is not expected that the Bessel-function model of football is able to explain such data with the same accuracy as in Figures 2, 3, and 4, although the fraction of players who know details about Bessel functions may even be higher among non-professional football players!
Unfortunately detailed data for non-professional football leagues which yield over many seasons all goals and the minutes in which they have been scored are not available.

Acknowledgements

The authors would like to thank Prof. Dr. Alfred Pflug (TU Dortmund) for stimulating discussions and Tim Tenhumberg for bringing the essential reference [2] to our attention.

References


[2] The public database with virtually all information about the German 1st and 2nd professional football leagues and information about leagues from other countries may be found at http://www.dasfussballstudio.de


