Using Skellam’s Distribution to Assess Soccer Team Performance

Peter Xenopoulos
Pomona College, Claremont CA 91711, USA, peter.xenopoulos@pomona.edu

Abstract. The revolution in sports analytics has yet to effectively reach soccer as it has to sports such as baseball and basketball. Using a theoretical framework based off of the Poisson and Skellam distributions, this paper contributes to soccer analytics by providing an easily calculate metric of predicted performance based off of the difference in distributions between a teams goals scored and allowed. Application of this method produces unbiased results of the amount of points a team scores in a season.

Keywords: Skellam’s distribution, Poisson, Soccer, Sports Prediction

1 Introduction

Methods to assess sport team performance are widely sought after by teams, fans and betting houses. While techniques such as Bill James’ Pythagorean Win-Loss for baseball and the Relative Power Index (RPI) in basketball each provide a glimpse of how teams compare in those sports, soccer is severely lacking in established metrics to measure relative team performance.

Increased popularity in soccer modeling, we believe is due to two reasons: the first being the relative size of the soccer sports betting markets and the second being the increased use of statistical analysis applied to soccer data. Because of the increased interest in analyzing soccer data, there are increasingly complex models. Some of these models try to predict the outcome of a single game and others try to determine the outcome of an entire season. In this paper, we argue, from a theoretical and empirical standpoint, that determining the outcome of an entire season can be accurately estimated, using just a team’s goals for and goals against, from what is known as a Skellam distribution.

The first part of this paper will address previous statistical literature in the area of evaluation of season data, particularly for soccer. In the second part, we will address the theoretical basis for our approach, which is based off of the Poisson and Skellam distributions. In the third part, we apply the theory to data and observe the results. Finally, we conclude in the fourth part with our findings. This paper expands on previous literature which explores the use of a bivariate Poisson distribution, known as a Skellam distribution.
2 Literature Review

Quantifying team performances was first pioneered in baseball with the Bill James Pythagorean formula, defined as:

\[
\text{WinPct} = \frac{RS^\gamma}{RS^\gamma + RA^\gamma}
\]  

(1)

Where \(RS\) denotes runs scored and \(RA\) denotes runs allowed. In the beginning of its use, \(\gamma\) was set to 2, hence the name "pythagorean" formula. This formula performed extremely well in predicting a team's win percentage over the course of Major League Baseball's 162 game season. It also amazed many analysts due to its simplicity, which inevitably led to its use as a pillar of mainstream baseball analysis. Through empirical research, it was determined that \(\gamma\) was around 1.83.  

In 2006, Miller proved that the formula indeed was true under a set of assumptions, one of which was that runs scored and allowed were distributed under a Weibull distribution. The formula is typically used at a variety of points in the season to represent the winning percentage a team "should" have and may allude to its sequencing of scoring (a team that is outperforming its Pythagorean win percentage may be having a lot of luck in close games, etc.)

This idea was further extended to soccer by Hamilton through the creation of the Extended Pythagorean Formula, which addressed Bill James’ Pythagorean Formula’s inability to deal with draws. The result, however, posited that, like runs in baseball, soccer goals were distributed from a three-parameter Weibull distribution. The use of the Weibull mainly seems to have been used for producing tractable results, as soccer goals are discrete, since they only occur in integer values (specifically, non-negative values).

There has been previous research indicating that soccer goals may be well described from a Poisson distribution. If we take a quick look at the data in Figure 1, the Poisson does indeed appear to fit the number of soccer goals in a game well. Taking the number of goals per game data from the English Premier League since 2002, we see a clear relationship with the Poisson distribution. The blue bars represent the frequency of said number of goals and the red square is the Poisson fit for a distribution with \(\mu = 2.651128\), the average number of goals per match during this time period. This distribution also maintains its Poisson shape when goals are broken down to distinguish between home and away goals.

---

1 Davenport and Woolner, Revisiting the Pythagorean Formula, 1999
2 Miller, A Derivation of the Pythagorean Win Loss Formula in Baseball, 2007
3 Hamilton, An Extension of the Pythagorean Expectation for Association Football, 2011
5 Karlis and Ntzoufras, Bayesian modelling of football outcomes: Using the Skellams distribution for the goal difference, 2009
6 Lee, Modelling Scores in the Premier League: Is Manchester United Really the Best?, 1997
3 Theory and Statistical Preliminaries

Before we go into a discussion of the probabilistic and statistical background of our approach, we must outline the Poisson and Skellam distributions.

3.1 Poisson Distribution

The Poisson distribution is a discrete probability distribution that is typically used to express the number of events in a fixed time interval, if these events occur with a known average rate and the occurrence of one event is independent of the other. For soccer, we believe it is reasonable to assume both that goals occur independently and that goals occur at an average rate. We discuss these assumptions later in the paper. The extended assumptions of a Poisson random variable include:

1. \( X \) is the number of times an event occurs in an interval and \( X \in \mathbb{N} \), where \( \mathbb{N} \) represents all of the natural numbers.
2. The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently.
3. The rate at which events occur is constant. The rate cannot be higher in some intervals and lower in other intervals.
4. Two events cannot occur at exactly the same instant.
5. The probability of an event in an interval is proportional to the length of the interval.

If these assumptions are true, then \( X \) is a Poisson random variable where \( X \sim \text{Poisson}(\lambda) \). The probability mass function is then defined as

\[
P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}
\]

Where \( X \) has a mean of \( \lambda \) and a variance of \( \lambda \). Because of the underlying Poisson assumptions, as well as the Poisson’s discrete nature, we believe the Poisson distribution to be the best probabilistic analogy for soccer goals.

3.2 Skellam’s Distribution

Suppose \( X \) is a Poisson random variable, with mean \( \lambda_X \) that describes the number of goals a team scores on average. Then, suppose \( Y \) is another Poisson random variable, with mean \( \lambda_Y \) that describes the number of goals a team allows, or gets scored on, on average. If \( Z = X - Y \), then \( Z \) represents the outcome of the game with respect to the home team. If \( Z > 0 \), then the home team has won. If \( Z = 0 \), the home team has drawn. If \( Z < 0 \), the home team has lost. \( Z \) has what we call a Skellam distribution, where \( Z \sim \text{Skellam}(\lambda_X, \lambda_Y) \).\(^7\) \( Z \) is said to have a probability mass function

\[
f_Z(z|\lambda_X, \lambda_Y) = P(Z = z|\lambda_X, \lambda_Y) = e^{-(\lambda_X + \lambda_Y)}(\frac{\lambda_X}{\lambda_Y})^{z/2}I_{|z|}(2\sqrt{\lambda_X \lambda_Y})
\]

\(^7\) Skellam. The frequency distribution of the difference between two Poisson variates belonging to different populations, 1946
for $z \in \mathbb{N}$, $\lambda_X, \lambda_Y > 0$ and where $I_r(x)$ is the modified Bessel function of order $r$.\footnote{Abramowitz and Stegun, 1974, pp. 375} \footnote{It is important to note that the definition of $f_Z$ also holds if $X$ and $Y$ are correlated. This is helpful in the case of soccer, because teams that typically score more also perform better on defense (the whole team is usually better)}

$$I_r(x) = \left(\frac{x}{2}\right)^r \sum_{k=0}^{\infty} \frac{\left(\frac{x^2}{4}\right)^k}{k! \Gamma(r+k+1)}$$  

(4)

We can then find the probability of a win ($P(Z > 0)$), the probability of a draw ($P(Z = 0)$) and the probability of a loss ($P(Z < 0)$) by taking sums to find the cumulative distribution function of $Z$.

We see a loss described as

$$P(Z < 0) = \sum_{i=-\infty}^{-1} f_Z(i)$$  

(5)

a draw described as

$$P(Z = 0) = f_Z(0)$$  

(6)

and a win described as

$$P(Z > 0) = \sum_{i=1}^{\infty} f_Z(i)$$  

(7)

Although infinite sums are not very tractable, when the Skellam distribution is applied on soccer data, the distributions are more or less well bounded by arbitrary cutoffs of -10 and 10. These cutoffs represent goal differences of 10, which by modern professional standards, especially in the English Premier League, where competition is close, is considered almost impossible. Therefore, the use of such a discrete distribution is actually quite easily calculable as infinite sums are avoided entirely.

### 3.3 Estimation of Parameters

In order to successfully apply Skellam’s distribution, we must determine $\lambda_X$ and $\lambda_Y$. The maximum likelihood estimator of a Poisson random variable is

$$\lambda_{MLE} = \frac{1}{n} \sum_{i=1}^{n} k_i$$  

(8)

where $k_i$ denotes the number of events observed in time period $i$. For $X$ and $Y$, the MLE estimate is the average number of goals scored per game and the average number of goals allowed per game. It is also important to note that the MLE estimator for $\lambda$ is unbiased.
4 Application

We will observe an application of the previously mentioned Skellam’s distribution to determine the number of points a team should have. Due to the fact that the MLE’s of $\lambda_X$ and $\lambda_Y$ are extremely easy to calculate, we start by calculating the average goals scored per game and goals allowed per game. This is standard with any soccer team standings table, which typically include team, games played, wins, draws, losses, goals for, goals against and points. Therefore, average goals scored per game would be calculated as $\frac{\text{Goals Scored}}{\text{Games Played}}$ and average goals allowed per game would be calculated as $\frac{\text{Goals Allowed}}{\text{Games Played}}$.

To find the loss, draw and win percentages for each team, we take their goals for per game and goals against per game and input them into equations (5), (6) and (7) respectively. However, instead of using infinity in our summations, we will use $-10$ for $-\infty$ and $10$ for $\infty$ as previously mentioned. This is done to make $f_{Z}$ calculable. We believe these substitutions should also not yield any difference in final result, because the cumulative distribution function on $(-\infty, -11)$ and $(11, \infty)$ is essentially 0 and 1, respectively.

Once we have found each team’s win, draw and loss percentages, we multiply them by the number of games a team has played, in order to find the number of wins, draws and losses. In the English Premier League, teams play 38 games by the end of the season. In Hamilton’s 2011 paper, using the extended Pythagorean Winning Percentage method, a mean point difference of 0.45 was observed.

In Table 1, we show the results of the using Skellam’s distribution to estimate points in the 2009-2010 English Premier League season. We see that the mean of the $\text{pointsDiff}$ column, which is the points difference in predicted and actual points, has a mean of $-0.37$. When compared to the average point difference of 0.45 observed in Hamilton’s 2011 paper, we see that there is no statistical difference between the two ($p = 0.71$), as well as no statistical difference from 0 ($p = 0.64$). This may suggest that using Skellam’s distribution may lead to an unbiased estimate of a soccer team’s points. However, it is important to note that Hamilton’s Extended Pythagorean Theorem also yields an unbiased estimate of a team’s season point total.

When applied on an expanded set of data, including each Premier League season from 2009-2010 to 2014-2015, we observed a mean point difference of $-0.1798$ with a standard deviation of 4.43 ($n = 100$). Using a two-tail test of means, with the null hypothesis that the mean point difference is 0, we again do not reject the null hypothesis on our expanded set of data. This furthers evidence that, by using Skellam’s distribution, one may attain an unbiased estimate of a team’s points – thereby its ranking compared to other teams.

If we look at the relationship between the goals per game difference, $\text{GDiff}$, defined as

$$\text{GDiff} = \frac{\text{Goals scored}}{\text{Games Played}} - \frac{\text{Goals allowed}}{\text{Games Played}}$$

In most soccer leagues, a team is awarded 1 point for a draw, 0 for a loss and 3 for a win.
we see a strong linear relationship in Figure 2, with no indication of het-
eroscedasticity. The estimated model is:

\[
\text{Points}_{\text{predicted}} = 51.9902 + 24.5274 \times (\text{Goals per Game, Scored} - \text{Goals per Game, Allowed})
\]

This relationship makes sense, as goals difference correlates highly with predicted points and points, with a \( \rho \) equal to 0.997 and 0.961 respectively. Predicted points correlates extremely highly with points, as also expected, with \( \rho \) equal to 0.964.

5 Conclusion

This paper presents an alternative to the expanded Pythagorean Win-Loss formula as presented in Hamilton’s 2011 paper. We believe the approach to be more intuitive, more simple in formulation and easily applicable across soccer leagues. In addition, the results are easily representable through a linear model. The major contribution of this work is that it presents a theoretical basis off of which the Skellam distribution may be used to calculate a league-agnostic win percentages, thereby bypassing the requirement of a Pythagorean league constant. Our method also requires very little data – normal standings data is all one needs to execute our method. Our method has been applied to years of historical data and has demonstrated its ability to give good estimates of team performance as a function of its offensive and defensive capabilities. Our result is useful in that it can be used to find teams that are under or over-performing relative to what they should be performing. Using the Skellam distribution, we present a method that takes team level data to quantify how well a team should be doing, much like the Pythagorean Win Formula in baseball.
Fig. 1. Poisson fit to goals per match
Fig. 2. Relationship between $GDiff$ and Predicted Points for 2009-2015 seasons.
**Table 1. Predicted 2009-2010 English Premier League Results**

<table>
<thead>
<tr>
<th>Team</th>
<th>W</th>
<th>D</th>
<th>L</th>
<th>Pts</th>
<th>Goals Per Game</th>
<th>Allowed Per Game</th>
<th>Predicted Pts</th>
<th>Pts Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chelsea</td>
<td>27</td>
<td>5</td>
<td>6</td>
<td>86</td>
<td>2.71</td>
<td>0.84</td>
<td>92.82</td>
<td>6.82</td>
</tr>
<tr>
<td>Manchester United</td>
<td>27</td>
<td>4</td>
<td>7</td>
<td>85</td>
<td>2.26</td>
<td>0.74</td>
<td>88.67</td>
<td>3.67</td>
</tr>
<tr>
<td>Arsenal</td>
<td>23</td>
<td>6</td>
<td>9</td>
<td>75</td>
<td>2.18</td>
<td>1.08</td>
<td>78.75</td>
<td>3.75</td>
</tr>
<tr>
<td>Tottenham Hotspur</td>
<td>21</td>
<td>7</td>
<td>10</td>
<td>70</td>
<td>1.76</td>
<td>1.08</td>
<td>69.78</td>
<td>-0.22</td>
</tr>
<tr>
<td>Manchester City</td>
<td>18</td>
<td>13</td>
<td>7</td>
<td>67</td>
<td>1.92</td>
<td>1.18</td>
<td>70.60</td>
<td>3.60</td>
</tr>
<tr>
<td>Aston Villa</td>
<td>17</td>
<td>13</td>
<td>8</td>
<td>64</td>
<td>1.37</td>
<td>1.03</td>
<td>61.25</td>
<td>-2.75</td>
</tr>
<tr>
<td>Liverpool</td>
<td>18</td>
<td>9</td>
<td>11</td>
<td>63</td>
<td>1.61</td>
<td>0.92</td>
<td>70.39</td>
<td>7.39</td>
</tr>
<tr>
<td>Everton</td>
<td>16</td>
<td>13</td>
<td>9</td>
<td>61</td>
<td>1.58</td>
<td>1.29</td>
<td>59.69</td>
<td>-1.34</td>
</tr>
<tr>
<td>Birmingham City</td>
<td>13</td>
<td>11</td>
<td>4</td>
<td>50</td>
<td>1.00</td>
<td>1.24</td>
<td>44.84</td>
<td>-5.16</td>
</tr>
<tr>
<td>Blackburn Rovers</td>
<td>13</td>
<td>11</td>
<td>4</td>
<td>50</td>
<td>1.08</td>
<td>1.45</td>
<td>42.09</td>
<td>-7.91</td>
</tr>
<tr>
<td>Stoke City</td>
<td>11</td>
<td>14</td>
<td>13</td>
<td>47</td>
<td>0.89</td>
<td>1.26</td>
<td>40.94</td>
<td>-6.06</td>
</tr>
<tr>
<td>Fulham</td>
<td>12</td>
<td>10</td>
<td>16</td>
<td>46</td>
<td>1.03</td>
<td>1.21</td>
<td>46.31</td>
<td>0.31</td>
</tr>
<tr>
<td>Sunderland</td>
<td>11</td>
<td>11</td>
<td>4</td>
<td>44</td>
<td>1.26</td>
<td>1.47</td>
<td>46.68</td>
<td>2.68</td>
</tr>
<tr>
<td>Bolton Wanderers</td>
<td>10</td>
<td>9</td>
<td>19</td>
<td>39</td>
<td>1.11</td>
<td>1.76</td>
<td>36.00</td>
<td>-3.00</td>
</tr>
<tr>
<td>Wolverhampton Wanderers</td>
<td>9</td>
<td>11</td>
<td>18</td>
<td>38</td>
<td>0.84</td>
<td>1.47</td>
<td>34.55</td>
<td>-3.45</td>
</tr>
<tr>
<td>Wigan Athletic</td>
<td>9</td>
<td>9</td>
<td>20</td>
<td>36</td>
<td>0.97</td>
<td>2.08</td>
<td>26.89</td>
<td>-9.11</td>
</tr>
<tr>
<td>West Ham United</td>
<td>8</td>
<td>11</td>
<td>19</td>
<td>35</td>
<td>1.14</td>
<td>1.74</td>
<td>40.05</td>
<td>5.05</td>
</tr>
<tr>
<td>Burnley</td>
<td>8</td>
<td>6</td>
<td>24</td>
<td>30</td>
<td>1.11</td>
<td>2.16</td>
<td>28.79</td>
<td>-1.21</td>
</tr>
<tr>
<td>Hull City</td>
<td>6</td>
<td>12</td>
<td>20</td>
<td>30</td>
<td>0.89</td>
<td>1.97</td>
<td>26.64</td>
<td>-3.36</td>
</tr>
<tr>
<td>Portsmouth</td>
<td>7</td>
<td>7</td>
<td>24</td>
<td>28</td>
<td>0.89</td>
<td>1.74</td>
<td>30.81</td>
<td>2.81</td>
</tr>
</tbody>
</table>