SOCCERMETRICS RESEARCH PUBLICATIONS

An Extension of the Pythagorean Expectation for Association Football

Howard H. Hamilton

Summary: This publication presents a formulation of an extension to the Pythagorean expectation for association football and other sports in which a draw result is a nontrivial event. Instead of estimating win percentage as in baseball, the extended Pythagorean estimates points won per game. A least-squares algorithm is used to fit offensive and defensive goal distributions to a three-parameter Weibull distribution, of which the parameter of interest is the Pythagorean exponent. Further analysis reveals that the league Pythagorean exponent remains stable across multiple leagues in the same calendar year and within a single league over multiple seasons, which gives support to the notion of a "universal" Pythagorean exponent. Application of the extended Pythagorean to results of domestic soccer leagues in Europe, Asia, and the Americas shows excellent agreement between goal statistics and league records for a majority of teams, and it indicates the teams that strongly overperform or underperform with respect to their expected performance.

Publication Date: January 2011

SOCCERMETRICS
RESEARCH

1 Introduction

This paper presents an extension to the Pythagorean expectation for use in association football (soccer). The Pythagorean expectation was first developed by James (1980) to predict the win percentage of a baseball team from the observed number of runs scored *RS* and runs allowed *RA* during the season:

$$\frac{\hat{W}}{M} = \frac{RS_{obs}^{\gamma}}{RS_{obs}^{\gamma} + RA_{obs}^{\gamma}} \tag{1}$$

In the above equation, $\hat{W}$ is the estimated number of wins divided by M matches to give a win percentage, and γ is the Pythagorean exponent that minimizes the root-mean-square difference between the predicted and observed winning percentages. James initially set the value of γ to 2.0, which inspired the Pythagorean name, but subsequent sabermetricians such as Davenport and Woolner (1999) have derived through empirical analysis a Pythagorean exponent of 1.86. Miller (2007) derived the Pythagorean formula from basic statistical principles and the assumption that the runs scored and allowed were independent random variables drawn from a Weibull distribution. His formula was virtually identical to James', which confirmed that the Pythagorean expectation was a probabilistic estimation of team results based on run statistics.

The Pythagorean expectation is calculated at a few intervals during the season to assess whether a team is performing above or below expectations. A team that exhibits a large difference between expected and observed wins (referred to in this paper as a "Pythagorean residual") could see its win-loss record attributed to either luck or subtle yet significant factors in its play. It is thought that the Pythagorean expectation regresses toward the mean number of wins in the league, so that teams with extremely high win percentages significantly overperform their expectations and teams with extremely low win percentages significantly underperform them.

The Pythagorean has been applied to baseball, basketball (Oliver (2004)), American football (Schatz (2003)) and other sports leagues (Cochran and Blackstock (2009)) with varying degrees of success. However, its applications to soccer have not been as successful (Anonymous (2006)) and have generally resulted in an underprediction of points won over a season. One reason for this is that the Pythagorean formula does not allow for the possibility of a tied result, which happens in a nontrivial percentage of soccer matches during a given season. The inclusion of drawn results also requires a redefinition of a win, which precludes the use of James' Pythagorean expectation.

This publication uses Miller's derivation of the baseball Pythagorean to formulate an extension to the Pythagorean for soccer and other sports in which a draw

result is a nontrivial event. Instead of estimating win percentage, the extended Pythagorean estimates points won per game. This change is consistent with the common practice of domestic soccer leagues to award points for wins and draws (currently defined as three points per win and one point per draw). A least-squares algorithm is used to fit offensive and defensive goal distributions to a three-parameter Weibull distribution, of which one of these parameters is the Pythagorean exponent. The exponent is estimated for each team in a league competition and then averaged to obtain a league Pythagorean exponent. Further analysis reveals that the league Pythagorean exponent remains stable across multiple leagues in the same calendar year and within a single league over multiple seasons, which gives support to the notion of a "universal" Pythagorean exponent. Application of the extended Pythagorean to results of domestic soccer leagues in Europe, Asia, and the Americas shows excellent agreement between goal statistics and league records for a majority of teams, and indicates the teams that strongly overperform or underperform with respect to their expected performance.

The remainder of the paper is organized as follows. Section 2 presents the derivation of the extended Pythagorean, starting with mathematical definitions of a win and draw and proceeding to derivation of their respective probabilities. Section 3 describes the least-squares algorithm used to estimate the Pythagorean exponent and outlines suggested procedures for implementing the extended Pythagorean. Section 4 presents results of statistical analyses performed on the extended Pythagorean, and applies the formula to the analysis of soccer leagues in England, the Netherlands, and the USA.

2 Extended Pythagorean Derivation

2.1 Statistical Preliminaries

There are several statistical preliminaries to be presented before the derivation of the extended Pythagorean. These preliminaries are borrowed from the publication by Miller (2007).

The goals scored and allowed by a team are modeled as statistically independent random variables. This modeling assumption appears to be a fair one to make in soccer because of the possibility of a draw. Moreover, researchers such as Dixon and Robinson (1998) have presented results that give credence to the existence of statistical independence between offensive and defensive goals.

Furthermore, the goals scored and allowed by a team during a season are modeled as random variables drawn from a three-parameter Weibull distribution:

$$\begin{aligned} f(x;\alpha,\beta,\gamma) &= \frac{\gamma}{\beta}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} e^{-\left(\frac{x-\beta}{\alpha}\right)^{\gamma}}, x \geq \beta \\ &= 0, x < \beta \end{aligned}$$

The distribution parameter is defined as β, the scale parameter α, and the shape parameter the Pythagorean exponent γ. Of course, a soccer team can only score integer goals, but this assumption is useful in order to construct notions of a win or draw. The use of a continuous distribution permits the computation of probabilities with integration instead of discrete summations, which results in more tractable solutions.

The distribution parameter β establishes a lower bound of possible scores. In order to reconcile the issue of using a continuous statistical distribution with discrete data, the data are placed into N bins, defined

$$[-.5,.5] \cup [.5,1.5] \cup \cdots \cup [6.5,7.5] \cup [7.5,8.5] \cup \cdots \cup [N-.5, N+.5] \qquad (2)$$

This construction moves the means of the bins to their centers, which is where all of the data in the bins would be located. Because there are only integer goals in soccer, the statistical model is continuous and the translation parameter is therefore $\beta = -0.5$. The bins facilitate the integration of the distribution between the endpoints of the bin, which permits the computation of draw probability.

The scale parameters α_{GS} and α_{GA} are related to the means of the two Weibull distributions, the calculation of which was detailed by Miller (2007). The means are equal to the average goals scored and goals allowed – G_S and G_A, respectively, so the alpha terms are defined as the following:

$$\alpha_{GS} = \frac{G_S - \beta}{\Gamma(1+\gamma^{-1})} = \frac{\widehat{G}_S}{\Gamma(1+\gamma^{-1})} \qquad (3)$$

$$\alpha_{GA} = \frac{G_A - \beta}{\Gamma(1+\gamma^{-1})} = \frac{\widehat{G}_A}{\Gamma(1+\gamma^{-1})} \qquad (4)$$

The shape parameter γ defines the skewness in the goalscoring distributions. This parameter will be estimated by a nonlinear least-squares algorithm, which will be described further in the paper.

2.2 Definition of Wins and Draws

Figure 1 shows an XY-plane for which the axes represent the goals scored by team X and team Y in a match. The grey squares are centered at identical X- and Y-

coordinates – (0,0), (1,1), (2,2) – and represent regions where the goals scored by team X and Y are less than 0.5 goals apart. In the lower regions of the plane, team X has scored at least 0.5 goal more than team Y, and in the upper regions of the plane the situation is reversed.

The mathematical definition of a win by team X, having scored c goals, is expressed as the following:

$$\begin{aligned} c-\frac{1}{2} &< X < c+\frac{1}{2} \\ 0 &< Y < c-\frac{1}{2} \end{aligned}$$

A draw at c goals is defined:

$$\begin{aligned} c-\frac{1}{2} &< X < c+\frac{1}{2} \\ c-\frac{1}{2} &< Y < c+\frac{1}{2} \end{aligned}$$

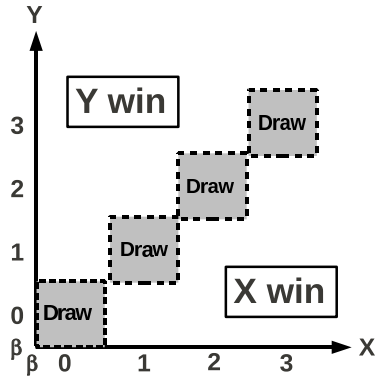

Figure 1: Illustration of result outcomes between teams X and Y in a soccer match. Grey regions represent draws at (c,c) goals, and blank regions represent wins.

2.3 Derivation of Win Probability

To derive the expression for win probability, one starts with the probability that team X will score c goals ($c \geqslant 0$) and team Y fewer:

$$\begin{aligned}
P(X=c, Y<c) &= \int_{c+\beta}^{c-\beta} \int_{\beta}^{c+\beta} \frac{\gamma}{\alpha_{GS}} \left(\frac{x-\beta}{\alpha_{GS}}\right)^{\gamma-1} e^{-\left(\frac{x-\beta}{\alpha_{GS}}\right)^{\gamma}} \frac{\gamma}{\alpha_{GA}} \left(\frac{y-\beta}{\alpha_{GA}}\right)^{\gamma-1} e^{-\left(\frac{y-\beta}{\alpha_{GA}}\right)^{\gamma}} dy\, dx \\
&= \int_{c+\beta}^{c-\beta} \frac{\gamma}{\alpha_{GS}} \left(\frac{x-\beta}{\alpha_{GS}}\right)^{\gamma-1} e^{-((x-\beta)/\alpha_{GS})^{\gamma}} \left[e^{-((y-\beta)/\alpha_{GA})^{\gamma}}\right]_{\beta}^{c+\beta} dx \\
&= \int_{c+\beta}^{c-\beta} \frac{\gamma}{\alpha_{GS}} \left(\frac{x-\beta}{\alpha_{GS}}\right)^{\gamma-1} e^{-((x-\beta)/\alpha_{GS})^{\gamma}} \left[1-e^{-(c/\alpha_{GA})^{\gamma}}\right] dx \\
&= \left[e^{-((x-\beta)/\alpha_{GS})^{\gamma}}\right]_{c+\beta}^{c-\beta} \left[1-e^{-(c/\alpha_{GA})^{\gamma}}\right] \\
&= \left[e^{-(c/\alpha_{GS})^{\gamma}} - e^{-((c-2\beta)/\alpha_{GS})^{\gamma}}\right] \left[1-e^{-(c/\alpha_{GA})^{\gamma}}\right]
\end{aligned}$$

A substitution for $\beta = -0.5$ yields the following expression:

$$P(X=c, Y<c) = \left[e^{-(c/\alpha_{GS})^{\gamma}} - e^{-((c+1)/\alpha_{GS})^{\gamma}}\right] \left[1-e^{-(c/\alpha_{GA})^{\gamma}}\right]$$

Finally sum over N, which defines the maximum number of goals, to obtain the total probability of a win by team X:

$$P(X>Y) = \sum_{c=0}^{N} \left[e^{-(c/\alpha_{GS})^{\gamma}} - e^{-((c+1)/\alpha_{GS})^{\gamma}}\right] \left[1-e^{-(c/\alpha_{GA})^{\gamma}}\right] \tag{5}$$

2.4 Derivation of Draw Probability

To derive the probability of a draw, one starts with the probability that teams X and Y will score the identical number of goals $c \geqslant 0$:

$$\begin{aligned}
P(X=Y=c) &= \int_{c+\beta}^{c-\beta}\int_{c+\beta}^{c-\beta} \frac{\gamma}{\alpha_{GS}}\left(\frac{x-\beta}{\alpha_{GS}}\right)^{\gamma-1} e^{-\left(\frac{x-\beta}{\alpha_{GS}}\right)^{\gamma}} \frac{\gamma}{\alpha_{GA}}\left(\frac{y-\beta}{\alpha_{GA}}\right)^{\gamma-1} e^{-\left(\frac{y-\beta}{\alpha_{GA}}\right)^{\gamma}} dy\,dx \\
&= \int_{c+\beta}^{c-\beta} \frac{\gamma}{\alpha_{GS}}\left(\frac{x-\beta}{\alpha_{GS}}\right)^{\gamma-1} e^{-((x-\beta)/\alpha_{GS})^{\gamma}} \left[e^{-((y-\beta)/\alpha_{GA})^{\gamma}}\right|_{c+\beta}^{c-\beta} dx \\
&= \int_{c+\beta}^{c-\beta} \frac{\gamma}{\alpha_{GS}}\left(\frac{x-\beta}{\alpha_{GS}}\right)^{\gamma-1} e^{-((x-\beta)/\alpha_{GS})^{\gamma}} \left[e^{-((c-2\beta)/\alpha_{GA})^{\gamma}} - e^{-(c/\alpha_{GA})^{\gamma}}\right] dx \\
&= \left[e^{-((x-\beta)/\alpha_{GS})^{\gamma}}\right|_{c+\beta}^{c-\beta} \left[e^{-((c-2\beta)/\alpha_{GA})^{\gamma}} - e^{-(c/\alpha_{GA})^{\gamma}}\right] \\
&= \left[e^{-((c-2\beta)/\alpha_{GS})^{\gamma}} - e^{-(c/\alpha_{GS})^{\gamma}}\right] \left[e^{-((c-2\beta)/\alpha_{GA})^{\gamma}} - e^{-(c/\alpha_{GA})^{\gamma}}\right]
\end{aligned}$$

A substitution for $\beta = -.5$ yields the following expression:

$$P(X=Y=c) = \left[e^{-((c+1)/\alpha_{GS})^{\gamma}} - e^{-(c/\alpha_{GS})^{\gamma}}\right] \left[e^{-((c+1)/\alpha_{GA})^{\gamma}} - e^{-(c/\alpha_{GA})^{\gamma}}\right]$$

The total probability for a draw between teams X and Y is obtained by summing over N:

$$P(X=Y) = \sum_{c=0}^{N} \left[e^{-((c+1)/\alpha_{GS})^{\gamma}} - e^{-(c/\alpha_{GS})^{\gamma}}\right] \left[e^{-((c+1)/\alpha_{GA})^{\gamma}} - e^{-(c/\alpha_{GA})^{\gamma}}\right] \quad (6)$$

2.5 Extended Pythagorean Formula

The extended Pythagorean is the sum of the win and draw probabilities defined in Sections 2.3 and 2.4 and is expressed as the number of points $\hat{P}$ won over M matches:

$$\begin{aligned}
\tfrac{\hat{P}}{M} &= 3P(X>Y) + P(X=Y) \\
&= 3 \cdot \sum_{c=0}^{N} \left[\exp\left\{-\left(\tfrac{c+1}{\alpha_{GS}}\right)^{\gamma}\right\} - \exp\left\{-\left(\tfrac{c}{\alpha_{GS}}\right)^{\gamma}\right\}\right] \left[1 - \exp\left\{-\left(\tfrac{c}{\alpha_{GA}}\right)^{\gamma}\right\}\right] \\
&+ \sum_{c=0}^{N} \left[\exp\left\{-\left(\tfrac{c+1}{\alpha_{GS}}\right)^{\gamma}\right\} - \exp\left\{-\left(\tfrac{c}{\alpha_{GS}}\right)^{\gamma}\right\}\right] \left[\exp\left\{-\left(\tfrac{c+1}{\alpha_{GA}}\right)^{\gamma}\right\} - \exp\left\{-\left(\tfrac{c}{\alpha_{GA}}\right)^{\gamma}\right\}\right]
\end{aligned}$$

It is more practical to use the goals scored/allowed statistics in the Pythagorean formula, so the expressions α_{GS} and α_{GA} are substituted into the expression:

$$\begin{aligned}\frac{\hat{P}}{M} &= 3 \cdot \sum_{c=0}^{N} \left[\exp\left\{-\left(\frac{\kappa(c+1)}{\hat{G}_S}\right)^{\gamma}\right\} - \exp\left\{-\left(\frac{\kappa c}{\hat{G}_S}\right)^{\gamma}\right\}\right]\left[1-\exp\left\{-\left(\frac{\kappa c}{\hat{G}_A}\right)^{\gamma}\right\}\right] \\ &+ \sum_{c=0}^{N} \left[\exp\left\{-\left(\frac{\kappa(c+1)}{\hat{G}_S}\right)^{\gamma}\right\} - \exp\left\{-\left(\frac{\kappa c}{\hat{G}_S}\right)^{\gamma}\right\}\right]\left[\exp\left\{-\left(\frac{\kappa(c+1)}{\hat{G}_A}\right)^{\gamma}\right\} - \exp\left\{-\left(\frac{\kappa c}{\hat{G}_A}\right)^{\gamma}\right\}\right]\end{aligned} \tag{7}$$

where κ is

$$\kappa = \Gamma\left(1+\frac{1}{\gamma}\right)$$

3 Implementation

3.1 Least-Squares Estimation of Weibull Parameters

A nonlinear least-squares algorithm is used to determine the parameters α and γ that give the best fit of the goalscoring histograms $(x_i,\, p_i)$ to the Weibull three-parameter distribution. (The translation parameter β has already been defined.) The critical parameter is γ, which must provide a satisfactory fit for the offensive and defensive goalscoring distributions simultaneously. The α terms are known precisely from the goal data, but the estimated values are useful in assessing the goodness of the data fit to the probability distribution.

The nonlinear least-squares algorithm will minimize the following cost function:

$$\min_{\{\alpha_{GF},\alpha_{GA},\gamma\}} ||\, p_i - f_{GF}(x_i;\alpha_{GF},-0.5,\gamma)\,||^2 + ||\, q_i - f_{GA}(x_i;\alpha_{GA},-0.5,\gamma)\,||^2 \tag{8}$$

in which f_{GF} and f_{GA} are the Weibull probability functions for goals scored and conceded, respectively, x_i the number of goals and $p_i(x_i)$ and $q_i(x_i)$ the proportion of goals scored and conceded, respectively. An iterative approach is used to calculate the best-fit parameters:

$$\Delta\xi = \left(\mathbf{J}^T\mathbf{J}\right)^{-1}\mathbf{J}^T\Delta y \tag{9}$$

$$\xi^{k+1} = \xi^k + \Delta\xi \tag{10}$$

where

$$\Delta y = \begin{bmatrix} p_i - f_{GF}\left(x_i;\alpha_{GF}^k,\gamma^k\right) \\ q_i - f_{GA}\left(x_i;\alpha_{GA}^k,\gamma^k\right) \end{bmatrix}$$

$$\xi = \begin{bmatrix} \alpha_{GF} & \alpha_{GA} & \gamma \end{bmatrix}$$

$$\mathbf{J} = \begin{bmatrix} \frac{\partial f_{GF}}{\partial \alpha_{GF}} & \frac{\partial f_{GF}}{\partial \alpha_{GA}} & \frac{\partial f_{GF}}{\partial \gamma} \\ \frac{\partial f_{GA}}{\partial \alpha_{GF}} & \frac{\partial f_{GA}}{\partial \alpha_{GA}} & \frac{\partial f_{GA}}{\partial \gamma} \end{bmatrix}$$

The partial derivatives of the Jacobian $\mathbf{J}$ are defined

$$\frac{\partial f}{\partial \alpha} = \left(\frac{\gamma}{\alpha}\right)^2 \left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} e^{-\left(\frac{x-\beta}{\alpha}\right)^{\gamma}} \left[\left(\frac{x-\beta}{\alpha}\right)^{\gamma} - 1\right] \tag{11}$$

and

$$\frac{\partial f}{\partial \gamma} = \frac{\gamma}{\alpha} \left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} e^{-\left(\frac{x-\beta}{\alpha}\right)^{\gamma}} \left[\frac{1}{\gamma} + \ln\left(\frac{x-\beta}{\alpha}\right) \left\{1 - \left(\frac{x-\beta}{\alpha}\right)^{\gamma}\right\}\right] \tag{12}$$

The least-squares algorithm is a gradient-based solver and converges to a solution within 20-30 iterations. There are some situations such as leagues with split season tournaments (which is currently common practice in Latin America) for the least-squares algorithm to become numerically sensitive. This problem can be remedied with the inclusion of a line search procedure, but the threshold for stopping the iteration may have to be raised.

3.2 Overall Procedures

There are two tasks to be carried out in the implementation of the extended Pythagorean: the computation of the league Pythagorean exponent, and the computation of the extended Pythagorean itself.

League Pythagorean Exponent

1. Collect match result data over a given season for all teams in the league. Arrange the data into goals scored and goals allowed columns for each team.
2. Compute a histogram of the goals scored and allowed per team.
3. Fit simultaneously the offensive/defensive goal histograms to the two-parameter Weibull distribution ($\beta = -0.5$). The result will be $\{\alpha_{GS}, \alpha_{GA}, \gamma\}$ for each team.
4. Calculate the mean γ over all of the league teams, resulting in the league Pythagorean exponent. The κ term should also be pre-calculated at this stage.

Computation of Extended Pythagorean

1. Collect total matches played, goals scored, and goals allowed for all teams in the league.
2. Divide goals by total matches played to obtain average number of goals scored or allowed.
3. Use the average goal values and league Pythagorean exponent to compute the extended Pythagorean. The resulting value is the estimated point average, which is multiplied by the number of matches and rounded up to obtain the estimated number of league points.

4 Results and Discussion

4.1 League Pythagorean Exponent

The League Pythagorean exponent is defined as the arithmetic mean of Pythagorean exponents of all of the teams in a league competition. It is essential to determine the behavior of the league Pythagorean exponent over multiple seasons, as well as the behavior of the exponent across multiple leagues in the same season. (Here, "season" is defined as the European soccer season which runs from August to May.) The behavior of the exponent across multiple leagues is of great interest because it could infer the existence of a "universal" Pythagorean exponent.

The behavior of the Pythagorean exponent for a single league over multiple seasons is illustrated by collecting league data for the English Premier League from the 1999-2000 season to the 2009-2010 season. The result data are used to develop goal histograms for all the teams in the league and then fit those histograms to the assumed Weibull distribution. Figure 2 shows a time history of the English Premier League's Pythagorean exponent. There is some oscillation present in the exponent and a wide deviation in the single-season exponents on the order of 10-30% of the mean, but the mean values remain within the 1.55-1.75 range.

To assess the behavior of the league Pythagorean exponent over multiple leagues, result data are collected from 41 top-level domestic leagues during the 2009-10 European season (2009 in the case of leagues played within a calendar year). A list of the leagues included in the study is presented in Table 1.

Figure 3 displays the league Pythagorean exponents grouped by region. There is some oscillation in the league Pythagorean exponent, but the values remain within a somewhat narrow region. The mean value of the exponent is 1.66 $\pm$ 0.26. There will always be uncertainty associated with the Pythagorean exponent, but it appears that an exponent between 1.60 and 1.85 will provide a good fit

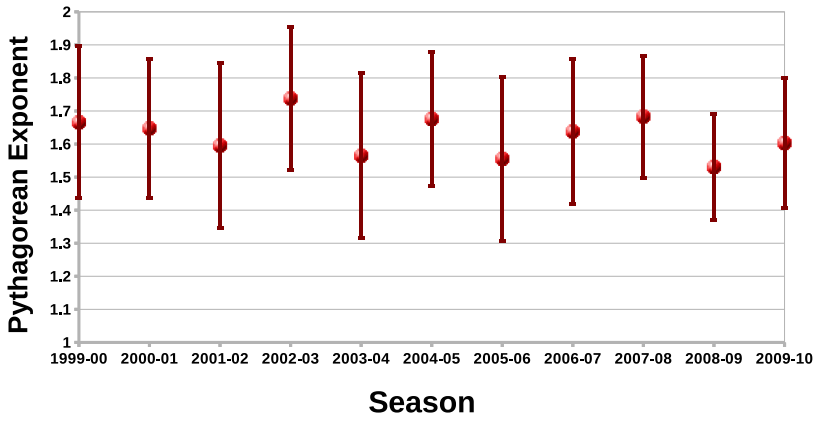

Figure 2: League Pythagorean exponent of English Premier League between the 1999-2000 and 2009-10 seasons.

Table 1: National soccer leagues included in Pythagorean study, 2009/2009-10 league season.

Region	National Leagues
Europe	Austria, Belgium, England, Finland, France, Germany, Greece, Israel, Italy, Netherlands, Norway, Poland, Portugal, Romania, Russia, Spain, Sweden, Switzerland, Turkey, Ukraine
North America	USA, Mexico, Honduras, Costa Rica, Guatemala, El Salvador, Panama
South America	Brazil, Argentina, Chile, Uruguay, Venezuela
Asia	Saudi Arabia, China, Iran, Qatar, Japan, South Korea
Africa	Egypt, South Africa, Tunisia

between observed and predicted point totals. In this study, the ”universal” Pythagorean exponent was set to 1.70.

4.2 League Expectation

With the league Pythagorean exponent determined, the extended Pythagorean is applied to various domestic soccer leagues around the world. The Pythagoreans for the league sides are generated according to the procedures described in Section 3.2, and a league Pythagorean exponent of 1.70 is used. The win and draw probabilities are converted to league wins and draws, which are then used to tabulate the esti-

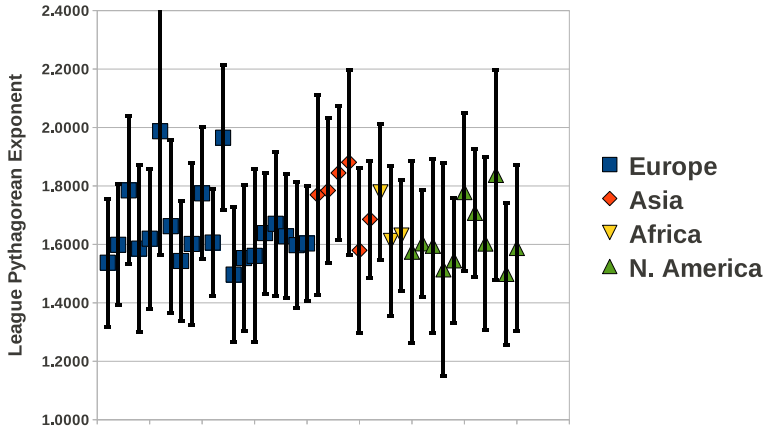

Figure 3: League Pythagorean exponent of national soccer leagues during the 2009/2009-10 season.

mated point totals. These totals are compared to observed point totals which do not account for administrative rulings (e.g. point deductions).

Table 2 presents the final Pythagorean table of the English Premier League during the 2009-10 season, ordered by observed point totals, then goal differential and goals scored. The top seven clubs in the table performed in line with their expectations. Toward the middle and lower ends of the table one encounters the sides that significantly over- or under-achieved during the season. Liverpool was not able to achieve the league results commensurate with their goal-scoring record, and finished outside of the top places which ensure participation in the European Champions League – an outcome that is disastrous for such an aspirational club. At the other end of the table, West Ham's goal statistics should have been sufficient to allow them to finish in mid-table, yet they failed to meet such expectations and nearly finished in the bottom three, which would have demoted them to the lower division of English soccer (Football League Championship). The reverse situation occurred to Wigan Athletic, which was expected to score 29 points yet achieved 36, ensuring their Premier League status in the process. A look at their match results showed that Wigan suffered a number of very heavy defeats yet scored narrow home wins against top-tier clubs such as Arsenal, Chelsea, and Liverpool.

The estimated win-draw-loss records are presented in order to determine the characteristics of the estimator. The extended Pythagorean does a better job of predicting the number of wins for teams at the extremes of the league table. The formula tends to overpredict the number of draws and losses, especially toward the

end of the table. However, the overprediction of draws is sometimes compensated by the slight underprediction of wins, which results in a smaller Pythagorean residual. The teams with the larger Pythagorean residuals have won or lost at least three more matches than estimated.

Table 2: Final Pythagorean table for English Premier League, 2009-10 season.

	League							Pythagorean				
Team	GP	W	D	L	GF	GA	P	$\hat{W}$	$\hat{D}$	$\hat{L}$	$\hat{P}$	ΔP
Chelsea	38	27	5	6	103	32	86	27	6	5	87	**-1**
Manchester United	38	27	4	7	86	28	85	26	7	5	85	**0**
Arsenal	38	23	6	9	83	41	75	23	8	7	77	**-2**
Tottenham Hotspur	38	21	7	10	67	41	70	20	9	9	69	**1**
Manchester City	38	18	13	7	73	45	67	20	9	9	69	**-2**
Aston Villa	38	17	13	8	52	39	64	17	11	10	62	**2**
Liverpool	38	18	9	11	61	35	63	20	10	8	70	**-7**
Everton	38	16	13	9	60	49	61	17	9	12	60	**1**
Birmingham City	38	13	11	14	38	47	50	11	11	16	44	**6**
Blackburn Rovers	38	13	11	14	41	55	50	11	10	17	43	**7**
Stoke City	38	11	14	13	34	48	47	10	11	17	41	**6**
Fulham	38	12	10	16	39	46	46	12	11	15	47	**-1**
Sunderland	38	11	11	16	48	56	44	12	10	16	46	**-2**
Bolton Wanderers	38	10	9	19	42	67	39	9	9	20	36	**3**
Wolverhampton Wanderers	38	9	11	18	32	56	38	8	10	20	34	**4**
Wigan Athletic	38	9	9	20	37	79	36	7	8	23	29	**7**
West Ham United	38	8	11	19	47	66	35	11	9	18	42	**-7**
Burnley	38	8	6	24	42	82	30	8	8	22	32	**-2**
Hull City	38	6	12	20	34	75	30	7	8	23	29	**1**
Portsmouth	38	7	7	24	34	66	28	8	9	21	33	**-5**

The Dutch Eredivisie is a domestic league with different characteristics from the English Premier League in its higher scorelines and the dominance by a small minority of clubs, albeit until very recently. Table 3 displays the league Pythagorean table for the 2009-10 season. The season featured a record-breaking season by Ajax Amsterdam, which scored 106 goals and allowed only 20, yet lost the league championship to FC Twente by a single point. Ajax's observed point total was exactly in-line with its statistical expectations, but Twente outperformed its expectations by twelve points – a margin of four games. One explanation could lie in the variance in the goals allowed of the two sides; Twente's variance of 0.68 was slightly smaller than Ajax's 0.77, but it is not clear whether the differences in

variance explain Twente's overachievement. The other overachiever in the Eredivisie was Heracles who performed two games better than expected, but they would have earned their final place regardless. At the other end of the table, Waalwijk, which finished in the direct relegation place, and Willem II, which finished in the relegation playoff place, had poor seasons, but perhaps Sparta Rotterdam had a performance that was expected of them. Overall, teams at the top of the table win more matches than might have been draws, while teams at the very bottom lose more matches that could have been draws.

The estimated win-draw-loss records of the teams illustrate how much Twente overachieved in their title-winning season. Twente scored fewer goals than third-placed PSV, and it appears that this created an estimated record that was much closer to PSV's. However, their defensive goal record might have made the difference in a number of matches and could have been a contributing factor to their outsized Pythagorean residual. It is worth mentioning that Ajax's win-draw-loss record was predicted exactly by the extended Pythagorean, as well as the records of Vitesse and Sparta Rotterdam. As in the English Premier League, the teams with large Pythagorean residuals won more matches than predicted. An exception can be found at the bottom of the table, where Waalwijk and Willem II exhibited large negative Pythagorean residuals due to having fewer drawn matches. Unless the goalscoring record is highly lopsided, the estimated number of draws in the league will vary between 20-30%.

USA's Major League Soccer has a regular season followed by playoffs, as is typical in other North American sports leagues. Table 4 displays the Pythagorean table for the 2010 regular season. The Pythagorean expectation reveals some differences between Major League Soccer and the two other European leagues examined. All of the MLS sides perform roughly in line with the statistical expectations. Only the Los Angeles Galaxy and the Columbus Crew overachieved by more than five points in the league; Chicago Fire was the most underachieving team with a Pythagorean residual of -5. In European leagues, by comparison, it is not uncommon to observe residuals of seven or more points, which would indicate an outlying performance. MLS teams also score a reduced number of goals and exhibit much tighter goal differences than the English or Dutch leagues, which might also account for the smaller Pythagorean residuals and the more uniformly predicted number of draws.

In general, teams that finish at the top or bottom of leagues deserve to be there by virtue of the fact that their point total is in line with the expectations from their goal statistics. There are some exceptions, like FC Twente in the Netherlands or Wigan Athletic in England. Teams at the very bottom not only deserve to be relegated but also play much worse than their statistics would indicate, which might indicate some sort of on-field breakdown – an event that is very typical of the worst-performing teams in a league.

Table 3: Final Pythagorean table for Dutch Eredivisie, 2009-10 season.

	League							Pythagorean				
Team	GP	W	D	L	GF	GA	P	$\hat{W}$	$\hat{D}$	$\hat{L}$	$\hat{P}$	ΔP
Twente	34	27	5	2	63	23	86	22	8	4	74	**12**
Ajax	34	27	4	3	106	20	85	27	4	3	85	**0**
PSV	34	23	9	2	72	29	78	22	7	5	73	**5**
Feyenoord	34	17	12	5	54	31	63	18	9	7	63	**0**
AZ	34	19	5	10	64	34	62	19	8	7	65	**-3**
Heracles	34	17	5	12	54	49	56	14	8	12	50	**6**
Utrecht	34	14	11	9	39	33	53	13	10	11	49	**4**
Groningen	34	14	7	13	48	47	49	13	9	12	48	**1**
Roda JC	34	14	5	15	56	60	47	13	8	13	47	**0**
NAC Breda	34	12	10	12	42	49	46	11	9	14	42	**4**
Heerenveen	34	11	4	19	44	64	37	9	8	17	35	**2**
VVV-Venlo	34	8	11	15	43	57	35	10	8	16	38	**-3**
NEC	34	8	9	17	35	59	33	8	8	18	32	**1**
Vitesse	34	8	8	18	38	62	32	8	8	18	32	**0**
ADO Den Haag	34	7	9	18	38	59	30	9	8	17	35	**-5**
Sparta Rotterdam	34	6	8	20	30	66	26	6	8	20	26	**0**
Willem II	34	7	2	25	36	70	23	7	7	20	28	**-5**
RKC Waalwijk	34	5	0	29	30	80	15	5	6	23	21	**-6**

5 Conclusion

This publication has presented an extension of the Pythagorean expectation to association football and other sports that permit draws. The formula is considerably more complicated than the original because of the presence of the exponential and Gamma function terms, as well as the summation terms in order to compute win and draw probabilities. The other major contribution of the paper is the establishment of a ”universal” Pythagorean exponent that can be used to develop expectations for domestic leagues around the world over many seasons. The extended Pythagorean has been applied to a number of soccer leagues from around the world and has been demonstrated to give good estimates of team performance as a function of their goal scoring records and designate potential outliers in the competition; that is, those teams that are significantly over- or underperforming with respect to their statistical expectations.

The soccer Pythagorean is a team-centered metric in that it assesses performance using the most important metric of all – goals. In the process, the Pythagorean answers the question, ”How is the team performing relative to expectations?”

Table 4: Final Pythagorean table for USA Major League Soccer, 2010 regular season.

	League							Pythagorean				
Team	GP	W	D	L	GF	GA	P	$\hat{W}$	$\hat{D}$	$\hat{L}$	$\hat{P}$	ΔP
Los Angeles Galaxy	30	18	5	7	44	26	59	15	8	7	53	**6**
Real Salt Lake	30	15	11	4	45	20	56	17	8	5	59	**-3**
Columbus Crew	30	15	8	7	42	32	53	13	8	9	47	**6**
Red Bull New York	30	15	6	9	38	29	51	13	9	8	48	**3**
FC Dallas	30	12	14	4	42	28	50	14	8	8	50	**0**
Seattle Sounders FC	30	14	6	10	39	35	48	12	8	10	44	**4**
Colorado Rapids	30	12	10	8	44	32	46	14	8	8	50	**-4**
San Jose Earthquakes	30	13	7	10	34	33	46	11	9	10	42	**4**
Kansas City Wizards	30	11	6	13	36	35	39	11	9	10	42	**-3**
Toronto FC	30	9	8	13	33	41	35	9	8	13	35	**0**
Chicago Fire	30	8	9	13	35	40	33	10	8	12	38	**-5**
Houston Dynamo	30	9	6	15	40	49	33	10	7	13	37	**-4**
New England Revolution	30	9	5	16	32	50	32	7	7	16	28	**4**
Philadelphia Union	30	8	7	15	35	49	31	8	7	15	31	**0**
CD Chivas USA	30	8	4	18	31	45	28	8	8	14	32	**-4**
DC United	30	6	4	20	21	47	22	5	8	17	23	**-1**

but not "How will the team perform in the future given its current form?" The soccer Pythagorean points out which teams might be performing well outside the 3-5 point differential, which would motivate further study of those teams. In the same vein, the Pythagorean could form part of the package of coaching metrics.

References

Anonymous (2006): "The Limits of Statistical Determinism, the Failure of Pythagorean Expectation," http://dcunitedblog.blogspot.com/2006/04/limits-of-statistical-determinism.html.

Cochran, J. J. and R. Blackstock (2009): "Pythagoras and the National Hockey League," *Journal of Quantitative Analysis in Sports*, 5, Article 11.

Davenport, C. and K. Woolner (1999): "Revisiting the Pythagorean Theorem: Putting Bill James' Pythagorean Theorem to the Test," http://www.baseballprospectus.com/article.php?articleid=342.

Dixon, M. J. and M. E. Robinson (1998): "A birth process model for association football matches," *The Statistician*, 47, 523–558.

James, B. (1980): *The Bill James Abstract*, self-published.
Miller, S. J. (2007): "A Derivation of the Pythagorean Won-Loss Formula in Baseball," *Chance Magazine*, 20, 40–48.
Oliver, L. D. (2004): *Basketball on Paper: Rules and Tools for Performance Analysis*, Potomac Books.
Schatz, A. (2003): "Pythagoras on the Gridiron," http://www.footballoutsiders.com/2003/07/14/ramblings/stat-analysis/4/.