http://www.newtheory.org **ISSN: 2149-1402**

Journal of New Theory

Received: 20.02.2019
Published: 02.05.2019

Year: **2019**, *Number*: **27**, *Pages*: **63-73**
Original Article

Prediction of Season-End Point for Football using Pythagorean Expectation[*]

Sezer Baysal[1] <sezer.baysal@ogr.deu.edu.tr>
Engin Yıldıztepe[1,†] <engin.yildiztepe@deu.edu.tr>

[1]*Dokuz Eylül University, Department of Statistics, Izmir, Turkey*

Abstract – The use of data collected on players, teams, and games for performance evaluation, player selection, score-outcome estimation, and strategy development using data mining tools and techniques are defined as sports data mining. Performance measures, unlike the common statistical methods, developed for each sport branch have an important role in sports data mining processes. Performance measures calculated for team sports can be used to predict the expectation of winning. The Pythagorean expectation developed for this objective was originally used in baseball games. The Pythagorean Expectation has also been adapted for other team sports with two results, such as basketball. However, the studies using Pythagorean Expectation for sports which have three possible outcomes are very limited. In this study, a suggestion for the calculation of Pythagorean Expectation for football is presented. In the application section, end-season rankings and points for the 2017/2018 season of the selected fifteen European football leagues are predicted by using the suggested method. The data of the past five seasons of the selected European football leagues is used as the training dataset. All calculations are performed in R.

Keywords – *Sports data mining, Pythagorean Expectation, Point prediction, Soccer, Football*

1 Introduction

Collecting and storing data have been easier and cost-effective in parallel with the progress in technology. Herewith, large amounts of data are generated in many different areas and used for different purposes. Sports data mining is defined as the use of data for performance evaluation, player selection, outcome-point prediction and strategy development by data mining tools and techniques. The decision makers of sports organizations can take more scientific and unbiased decisions by sports data mining compared to traditional methods. Sports data mining is rapidly spread and adopted due to clearly demonstrating team player performance and helping talent scouts to discover new talents. In addition, the popularity of sports data mining has increased due to the studies conducted on predicting the outcomes of sports events.

[*] *The initial version of this study has been presented as an oral presentation at the AB2019 conference.*
[†] *Corresponding Author*

In sports data mining, it is necessary to define sabermetric first. Sabermetric depends on the idea of creating new statistics that better measure individual and team performances compared to the traditional statistical methods in baseball. Although the idea had been proposed earlier, it has been introduced by Bill James at the end of the Seventies in the annual "Baseball Abstracts" booklets published by himself. James rapidly pronounced his name and increased popularity with his unusual ranking methods and new statistical performance measurements called sabermetrics. The transition from traditional statistics to sabermetrics is the result of queries and solutions on the performance criteria introduced by Bill James. James [8] described the sabermetrics of which he developed in his later books. Pythagorean Expectation (PE), a performance measurement metric that predicts the game-winning rate of teams in baseball, was developed by James [7]. The PE has been widely used for baseball in the subsequent years. Lee [9] applied the PE for the 2005-2014 seasons of the Korean Baseball League and compared the expected and actual game winning numbers of clubs. Inconsistency between expected and actual winning numbers, assuming the conditions of the teams originated due to an unusual distribution, has been related to the coefficient of variation and standard deviation in the number of runs allowed. Tung [16] applied the PE to the data set of seasons from 1901 through 2009 and produced a confidence interval for the number of games predicted to be won. Valero [17] predicted the outcomes of the American Baseball League by using sabermetrics, including PE to assess the predictive capabilities of data mining methods. Valero, following the statistical analysis, showed that classification methods resulted in better outcomes. The PE is given in detail in the second section.

Performance measurements in basketball are performed as a team rather than individually since the performances of the players are relatively more dependent on each other compared to baseball. Dean Oliver is the pioneer of performance measurement in basketball. Oliver has developed new statistics for basketball in the Eighties [15]. In 2004, Oliver published the statistical methods for assessment in basketball and calculation tools to evaluate the teams [13].

The statistical techniques used in American football have not yet reached the levels reached in baseball and basketball. Schumaker et al. [15] attributed this to less number of games in American football compared to baseball and basketball and lack of some statistics about the players. A team in the American national football league plays 16 games in a season, while 162 games in baseball and 82 games in basketball [15]. Leung and Joseph (2014) mentioned the Christodoulou algorithm that is used in the prediction of dual matchings and applied this algorithm to the American football data [10]. In the application section of their study, the distances of teams to each other in the American Football League were calculated and revealed similar teams by using the PE, Christodoulou algorithm and other sabermetrics. When the results of the future matches are predicted, according to the results of the matches between similar teams, points are assigned to the teams who have not played.

The Christodoulou algorithm generates five statistics for the competing teams in a league based on the game outcomes. These are the number of points gained per game for a team (NPPG), the number of points scored by opponent per game (NPOPG), the number of points per games recorded in a league (NPRL), offensive strength (OS) and defensive strength (DS). The OS specifies the percentage of points scored by a team against their opponent to the number of points per game typically allowed by this opponent. For example, if NPRL is 40 in a league and a team can score 60 points per game, then the OS of the team would be 60/40. The DS indicates the ratio of points that a team allows to the opponent relative to the NPRL. For example, if the NPRL is 40 in a league, and an average team score per game is 20, then the DS of a team would be 20/40. The Christodoulou algorithm aims to predict the outcomes of games using these statistics [10]. The aforementioned statistics which indicate the performance of a team are calculated for a team-A as follows.

$$OS_A = \frac{NPPG_A}{\mathrm{NPRL}_{LEAGUE}}$$
$$DS_A = \frac{\mathrm{NPOPG}_A}{\mathrm{NPRL}_{LEAGUE}} \qquad (1)$$

$$\mathrm{NPPG}_A = \frac{Points\ Scored_A}{Games\ Played_A}$$
$$\mathrm{NPOPG}_A = \frac{Points\ Scored\ by\ Opponent\ _A}{Games\ Played_A} \qquad (2)$$

The score of the game played by teams A and B can be predicted as follows using the above statistics.

$$Score_A = OS_A \times \mathrm{NPOPG}_B + \mathrm{NPPG}_A \times DS_B$$
$$Score_B = OS_B \times \mathrm{NPOPG}_A + \mathrm{NPPG}_B \times DS_A \qquad (3)$$

Cricket is another sport field where performance measures are applied. Cricket sport is considered utterly rich in terms of statistics [3]. John Buchanan, the coach of the Australian national team, has pioneered many of sabermetrics involving in the cricket sport between 1999 and 2007. The most well known is “Marginal Wins”. The performances of players are evaluated through these statistics according to their positions and can also be compared with the opponent players [15]. Vine [18] determined the “lucky” and “unlucky” teams by comparing the predicted and actual number of winnings of cricket teams. Vine who used the 4-season data set of the Australian Cricket League, assumed the coefficient of γ as 7.41 while adapting the PE to cricket sport. While determining these coefficients, the criterion was defined as the minimum root mean squared error (RMSE).

Several attempts have been carried out to create statistical measures similar to sabermetric in football. However, analysis of game activity and game-based events in football are far better difficult than baseball. Because the performances of the players in football are much more dependent on each other compared to baseball. The roles of the players in baseball have been set sharply; the pitcher hits the ball, the batter meets the coming ball by his bat. In football, teams can attack and defend with various strategies and number of players. Therefore, sabermetric style performance measurement were not generally used in the data mining studies performed in football.

In this study, an approach is presented for adaptation of PE, a sabermetric developed for baseball, to football using past season data. In the application, it is aimed to predict the points and the ranking of the teams with the proposed approach based on the scored and conceded goals at the end of the season. The use of PE in football and the proposed approach are presented in the second section. The application is given in the third section, and the results are shared in the final section.

2 Method

In this section, the details of PE and the adaptations of PE in other sport branches including football are given.

2.1 Pythagorean Expectation

PE has been proposed by Bill James [7] as a performance measurement metric that predicts the team winning rate of baseball teams using runs scored (RS), runs allowed (RA) obtained from past games and the league constant of γ. The PE can be used to determine the teams which performing above and below the expectations by comparing the actual winning rate. PE for baseball is calculated as in Equation 4.

$$PE = \frac{RS^{\gamma}}{RS^{\gamma} + RA^{\gamma}} \tag{4}$$

PE is typically used in the middle of a season to predict the standings for the end of a season. For example, if a team wins more than the predicted in the halfway through a season, analysts claim that the team will complete the remaining half of the season with fewer winnings than the predicted [12]. The value of constant γ in original formula has been set to 2.0 by James. Miller [12], however, has shown that the use of constant γ as 1.82 reduces the standard error.

Various applications have been suggested for also baseball. Davenport and Woolner [4] argued that the γ value should be calculated separately for each team according to the balance of offensive and defensive power, and suggested that the γ coefficient in baseball should be calculated as in Equation 5 to obtain a smaller RMSE value.

$$\gamma = 1.5 \times \log\left(\frac{RS + RA}{NG}\right) + 0.45 \tag{5}$$

Where RS is the number of runs allowed, RA is the number of runs allowed and NG is the number of games played by the team.

PE has attracted the attention in other sport branches due to its impact on baseball. Different γ values have been attained in the studies conducted using PE in sport branches such as American football, cricket, basketball and ice hockey. Some of these studies are summarized in Table 1.

Table 1. Recommended γ Values for Different Sport Branches

Sport	**γ**	**Source**
Baseball	1.82	[12]
American Football	2.37	[14]
Basketball	14 13.91	[13] [19]
Ice Hockey	1.927	[2]
Cricket	7.41	[18]

2.2 Pythagorean Expectation in Football

PE in sports which have two possible outcomes (win - or - loss) which mentioned in the previous section have been applied only with the changes made in the γ coefficient. However, the teams acquire points below the predicted if the original formula is applied directly without any arrangement in football which is a sport that can result in a tie [5, 6].

Hamilton [6], considering the possible tie outcome in football, predicted the points earned per game instead of predicting the winning ratios of teams using the extended Pythagorean method. Hamilton tried to overcome the problem that PB was only applicable to sports with two possible outcomes by calculating the probability of winning and draw for each team. Hamilton calculated the predicted point per game (PPPG) for team X by using Equation 6 where X representing a team playing in the league and Y representing the opponents.

$$\mathrm{PPPG} = 3 \times \mathrm{P(X > Y)} + \mathrm{P(X = Y)} \tag{6}$$

Hamilton [6] used the least squares algorithm to express the scored and conceded goals distributions with a three-parameter Weibull distribution. However, Hamilton's method has not widely used due to intensive mathematical and statistical procedures.

Eastwood [5] took the draw possibility into account and adopted the original PE to a football game which has 3-point for a win, 1-point for a draw, 0-points for a loss. Eastwood, instead of calculating the winning possibility of the teams, calculated the PPPG multiplying the average point per game (APPG) by the probability of gaining points. The equation developed by Eastwood to calculate the PPPG for each team is given below;

$$PPPG = \frac{G^{1.22777}}{G^{1.072388} + CG^{1.127248}} \times 2.499973 \tag{7}$$

In the Equation 7; G is the number of goals scored, CG is the number of goals conceded and the APPG is 2.499973.

Hamilton [6] determined the γ value with a single season data and found RMSE value as 3.81. Eastwood [5] obtained lower RMSE values by using the data collected from ten seasons. The adaptation of Eastwood [5] seems like much straightforward and more practical than the adaptation formula of Hamilton [6]. However, Eastwood developed and implemented the formula only over the English Premier League data.

2.3 Proposed Approach

This section outlines the proposed approach to adapt PE to football. The proposed formula, unlike baseball, is aimed to predict the expected points per game of the teams instead of winning possibilities. In order to calculate the PE, the number of goals scored and conceded by the reference team in the league have to be known. In addition, the exponential coefficient γ and the average points distributed per game in the league (APDG) should also be determined. The most important difference between recommended approach and Eastwood's formula is the usage of the γ coefficient. Eastwood's formula uses three different γ coefficients. The PE equation, which calculates the expected points per game for each team, is written as follows with the determination of the required coefficients γ and APDG:

$$\text{PE} = \frac{{Goal_S}^{\gamma}}{{Goal_S}^{\gamma} + {Goal_C}^{\gamma}} \times APPG \tag{8}$$

$Goal_S$ represents the number of goals scored and $Goal_C$ is the number of goals conceded.

Since football is not a sport with two possible outcomes, the APPG value cannot be taken as 3 points. Considering that there are three possible outcomes in football, the ratios of the draw and win-loss in the leagues must be determined (Equation 9). The APPG is calculated by Equation 10 using the statistics for the total game played in the league (TGP), total win (TW), total draw (TD) and total loss (TL).

$$\begin{gathered} Ratio_{win-loss}: \frac{TW+TL}{TGP} \\ Ratio_{draw}: \frac{TD}{\text{TGP}} = 1 - \frac{TW+TL}{\text{TGP}} \end{gathered} \tag{9}$$

$$APPG = 3 \times (Ratio_{win-loss}) + 2 \times (Ratio_{draw}) \tag{10}$$

The γ coefficient in PE for football can be predicted by simple linear regression method (SLR). The SLR provides a linear function that models the relationship between the dependent and independent variables with the least squares (LS) algorithm.

In the proposed approach, PB is calculated with the Equation 8 by using the various gamma values between 1 and 2 for all teams in the league then PB is multiplied by the number of matches played in order to predict end-season points. A regression model is created by using SLR where the predicted score as the explanatory variable and the actual score as the response variable. Consequently, the SLR models are generated as much as the number of γ tested. The optimum γ coefficient is determined by examining the RMSE obtained in the models and the coefficient of determination R^2.

3 Application

The data of fifteen European football leagues belong to the six seasons between 2012-2013 and 2017-2018 seasons used in the study were compiled from the mackolik.com website [11]. The leagues used in the application belong to countries of Turkey, Italy, Germany, Spain, France, Holland, England, Belgium, Austria, Croatia, Denmark, Czech Republic, Portugal, Romania, and Scotland. The league tables used in the study include the number of games played for each team (G), the number of wins (W), the number of draws (D), the number of losses (L), the goals scored (S), the goals conceded (C) and the end of season points (P). Play-offs, canceled games, and cup games have not been included in the data used. The league tables belong to past five years (2012 – 2017) of fifteen European football leagues were used as training data in the application. The data for 152 teams played during five seasons (2012-17) in the league (never dropped out) were used in the training data set. The data of 244 teams in the 2017-2018 season were separated as test data. All calculations are performed in R statistical programming language. A small excerpt of the data is shown in Table 2.

Table 2. An Example of a League Table

Teams	Country	G	W	D	L	S	C	P
Athletic Bilbao	Spain	190	84	43	63	263	233	295
Pandurii Targu Jiu	Romania	154	64	38	52	222	192	224
Nice	France	190	83	46	61	252	220	295
Zulte Waregem	Belgium	150	68	40	42	241	209	244
AZ Alkmaar	Holland	170	72	40	58	299	265	256
Schalke 04	Germany	170	74	40	56	259	222	262

Firstly, the win-loss and draw ratios were calculated by using Equation 9:

$$Ratio_{win} = 0.7471$$

$$Ratio_{draw} = 0.2528$$

APPG was computed with Equation 10 as follows:

$$APPG = 3 \times (0.7471) + 2 \times (0.2528)$$

$$APPG = 2.7471$$

The most successful results were obtained in the $1 \leq \gamma \leq 2$ range in our preliminary study. Therefore, PE of each team was calculated with Equation 8 using 2.7471 as APPG for eleven different γ coefficients between 1.0 and 2.0. The calculated PE values are multiplied by the number of games played, and eleven distinct points are predicted for the total points of the teams for the five seasons. Eleven simple linear regression models were created to find the most appropriate γ value, where the predicted points were the independent variable (x_i) and the actual points were the dependent variable (y_i). The RMSE and coefficient of determination R^2 values for the obtained models are shown in Table 3 and Figure-1:

Table 3. RMSE and R^2 Values of Models Obtained with Different γ Coefficients

γ	RMSE	R^2
1.0	11.21613	0.974433
1.1	10.55446	0.977361
1.2	10.28528	0.978501
1.3	10.31085	0.978394
1.4	10.54836	0.977387
1.5	10.93303	0.975708
1.6	11.41728	0.973508
1.7	11.96756	0.970893
1.8	12.56064	0.967937
1.9	13.18052	0.964694
2.0	13.81608	0.961207

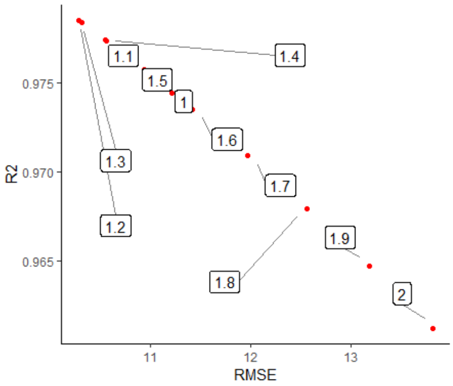

Figure 1. RMSE and R^2 values of models obtained with different γ coefficients

The results revealed that the lowest RMSE and the highest R^2 values were obtained when γ=1.2. The recommended PE of European football leagues can be calculated by the Equation 11 using the specified value of the coefficient. The γ value was recommended as 1.3 in another study using twelve European leagues [1].

$$\text{PE} = \frac{Goal_S^{1.2}}{Goal_S^{1.2} + Goal_C^{1.2}} \times 2.7471 \tag{11}$$

The SLR used with the LS algorithm is a parametric statistical method that requires some assumptions. Thus, initially, the required assumptions must be checked. For this purpose, normality, and independence of the residuals obtained by the model were examined. Secondly, variance homogeneity was investigated. The results of the analyses showed that the assumptions were satisfied. Required tests for the validity of the model and coefficients were also performed. The significance level of all hypothesis tests was accepted as 0.05.

The data of the 2017-2018 season were used for evaluation. In the first stage, the differences between the predicted and actual points were examined to measure the success of the proposed approach (Table 4).

Table 4. An Example for End of Season Actual Points and PE Predictions

Teams	**Actual Point**	**Predicted Point**	**Difference**
Milan	64	59.65	-4.35
Saint-Etienne	55	50.58	-4.42
CFR Cluj	59	54.54	-4.46
RB Leipzig	53	48.40	-4.60
Lazio	72	67.37	-4.63
Real Madrid	76	71.10	-4.90
Utrecht	54	48.80	-5.20

In the evaluation, the margin of error was considered only a match. So, predictions with less than three-point difference from the actual point value were considered successful. The success rates calculated for 15 European leagues are presented in Table 5. The overall success rate for all leagues was 40%.

Table 5. Success Rates of Leagues Obtained in Point Prediction

League	Success Rate
Germany	56%
Czech Republic	56%
Romania	56%
Belgium	50%
England	44%
Denmark	33%
France	33%
Crotia	33%
Spain	33%
Italy	33%
Turkey	33%
Netherland	28%
Austria	22%
Scotland	22%
Portugal	11%

The points gained at the end of the season determine the team standings in the league. The predicted points by PE of the teams in 2017-2018 end-of-season were used to measure the success of the proposed approach based on the standings, and the teams were ranked based on the points predicted among the teams in their leagues. An additional evaluation was performed for the first four teams in the leagues. The first four rankings are considered important for the league success of a team and qualifying to the European cups. When calculating the ranking-based success ratio, predictions that predict the season end ranking of a league exactly or predict the ranking by only one difference are accepted as successful. The ranking-based success ratios are given in Table 6.

Table 6. Success Rates of Leagues by Ranking

League	Success Rate for ranking	Success Rate for only First Four ranking
Crotia	100%	100%
Scotland	100%	100%
Austria	90%	100%
Romania	86%	100%
Denmark	79%	100%
England	75%	100%
Netherland	72%	100%
Portugal	72%	100%
Italy	70%	100%
France	65%	100%
Spain	65%	100%
Czech Republic	63%	100%
Germany	61%	100%
Belgium	44%	50%
Turkey	39%	75%

The success ratios for the ranking were higher than 50%, except for two leagues. The rankings for Croatia and Scotland leagues were predicted exactly.

4 Conclusion

In this study, a PE calculation approach was proposed for European football leagues. The 2017-2018 end-of-season points of 244 teams playing in European leagues were predicted in order to measure the success of the proposed approach. The success of PE, which is proposed with two different approaches according to the ranking and score, was evaluated by using the points predicted. More successful results were obtained with ranking based prediction. In the study, relatively low success ratios were obtained for Turkey and Belgium leagues. However, high success ratios were obtained for Croatia, Scotland, Austria and Romania, where there are fewer teams in the league compared to the other countries. Another noteworthy outcome was the high accuracy rate in the rank-based evaluation. In this study, the γ value in PE formula for football was calculated as 1.2. Specific γ values must be calculated for different leagues in order to make successful predictions. Further studies are planned to determine the γ values for Asian and South American football leagues.

References

[1] S. Baysal, E. Yıldıztepe, *Futbol Takımlarının Sezon Sonu Puanlarının Tahmini için Pisagor Beklentisine Dayalı bir Çalışma*, Akademik Bilişim 2019, Ordu, 2019.

[2] J. Cochran, R. Blackstock, *Pythagoras and the National Hockey League*, Journal of Quantitative Analysis in Sports 5 (2009) 1-13.

[3] J. Croucher, *Player ratings in one-day cricket*, Proceedings of the Fifth Australian Conference on Mathematics and Computers in Sport, Sydney, 2000.

[4] C. Davenport, K. Woolner, *Revisiting the Pythagorean Theorem: Putting Bill James' Pythagorean Theorem to the test'*, https://www.baseballprospectus.com/news/article/342/revisiting-the-pythagorean-theorem-putting-bill-james-pythagorean-theorem-to-the-test/ (January, 2019).

[5] M. Eastwood, *Applying the Pythagorean Expectation to Football: Part Two*, http://pena.lt/y/2012/12/03/applying-the-pythagorean-expectation-to-football-part-two/ (December, 2018).

[6] H. H. Hamilton, *An extension of the pythagorean expectation for association football*, Journal of Quantitative Analysis in Sports 7 (2011).

[7] B. James, *The Bill James Baseball Abstract*, 1980.

[8] B. James, *The Bill James Historical Baseball Abstract*, Villard, New York, 1985.

[9] J. Lee, *Measuring the accuracy of the Pythagorean theorem in Korean pro-baseball*, Journal of the Korean Data and Information Science Society 26 (2015) 653-659.

[10] C. K. Leung, K. W. Joseph, *Sports data mining: predicting results for the college football games*, Procedia Computer Science 35 (2014) 710-719.

[11] Mackolik, *Puan durumu*, http://arsiv.mackolik.com/Puan-Durumu (December, 2018).

[12] S. J. Miller, *A Derivation of the Pythagorean Won-Loss Formula in Baseball*, Chance 20 (2007) 40-48.

[13] D. Oliver, *Basketball on paper : rules and tools for performance analysis*, Potomac, Washington D.C, 2004.

[14] A. Schatz, *Pythagoras on the Gridiron*, https://www.footballoutsiders.com/stat-analysis/2003/pythagoras-gridiron (January, 2019).

[15] R. P. Schumaker, O. K. Solieman, H. Chen, *Sports Data Mining*, Springer, Boston, 2010.

[16] D. D. Tung, *Confidence Intervals for the Pythagorean Formula in Baseball*, http://www.rxiv.org/pdf/1005.0020v1.pdf (November, 2018).

[17] S. C. Valero, *Predicting Win-Loss outcomes in MLB regular*, International Journal of Computer Science in Sport 15 (2016) 91-112.

[18] A. J. Vine, *Using Pythagorean Expectation to Determine Luck in the KFC Big Bash League*, Economic Papers 35 (2016).

[19] D. Zminda, J. Dewan, STATS Inc. Staff, *STATS Basketball Scoreboard, 1993-94*, Harpercollins Publishers, Skokie, 1993.