A parametric deterministic model for estimating football trivariate outcome point

Ismail. A. Adedeji[1], Adewunmi Olaniran Adeyemi [2*] and Eno Emmanuella Akarawak[2]

[1]*Departments of Actuarial Science and Insurance, University of Lagos, Akoka, Lagos*
[2]*Department of Mathematics, University of Lagos, Akoka, Lagos*

This study aimed at using a standardized bivariate Pearson family of distribution to generalize a bivariate function and also develop a Parametric Deterministic Model (PDM) that represents mathematical relationship between football matches having trivariate outcomes and the position of teams in the league table. The study established bivariate function as a useful generalization of the univariate Pareto (Type 1) distribution; it also evaluated the previous five-year performance of teams from four major leagues in Europe based on their end of season points(P_{ik}). The Anderson-Darling goodness of fit test (AD-Test) was employed to measure how well the data fits specified theoretical distribution. Negative Binomial distribution is the model that provides the best fit for approximating the distribution of the over-dispersed estimated points associated with six top teams in each of the four leagues. The result revealed Ligue 1 is the most competitive among the four leagues while English Premier League (EPL) is rated the most competitive in the sub-leagues of six teams at the top of the table. The PDM accurately modelled the potential outcomes of series of football matches with the corresponding estimated points for the teams and also generated the league table for EPL using data of 2016/2017 football season.

Keywords: b*ivariate Pearson family; football trivariate outcomes; football points; parametric deterministic model; Pareto distribution*

1 Introduction

There is growing interest in the field of statistics on how to define new generators or mathematical functions that can be used to generate new models. Among the notable generators in literature is the Beta-G by Eugene *et al*. (2002). In recent years, statisticians have introduced various generalized models and their flexibilities over their baseline distributions were tested using real life data. Marshall and Olkin (1997) proposed a flexible semi-parametric family of distributions by using an additional parameter called the tilt parameter and thereafter defined a new survival function. Various methods for generating new models include the works of Pearson (1895), Johnson (1949). This study considered case III of some solutions to differential equations by VanUven (1947) and used it as a generator to construct a bivariate function and also developed a Parametric Deterministic Model (*PDM*) that estimates points obtained from trivariate outcome of football events by various football teams in the league format.

Analysis of events in the sports related industries continue to attract widespread popularity in the field of statistics. Many articles on statistical methods for analyzing sports data from English Soccer, American Football, Baseball, Basketball and Hockey have been published. In the football segment of the sports industry, there is huge investment of funds by various stakeholders. It is therefore, of paramount importance to review and analyze the performance of

* Corresponding Author; Email: adewunyemi@yahoo.com

football teams and outcomes of football results for effective decision making. The football industry has been very quiet and inactive in terms of data analysis for many years despite its huge followership. However, researchers are now developing keen interest in the industry as the rise and role of data analysis is gaining popularity for effective analysis of opponent, tactical feedback, post-match analysis, post-match statistics and performance analysis of prospective target during the transfer windows. In 2016, a journalist with an America podcast opined that Leicester City Football club insight and data analytics helped the team win the premier league, particularly in area of correctly recruiting players for position in the field (in contrast to Arsenal Football club). Statistical and data analytics efforts are required in both player recruitment and all on-field actions.

A search into literature revealed different statistical techniques have been proposed by researchers for modeling data from number of goals scored by each team in a football match while interest in the study of outcomes (Win, Draw, and Loss) is very poor. Consequently, there is scarce paper on research techniques for analyzing outcomes of football matches played by various teams.

Moroney (1957) refuted claims that Poisson distribution can provide a good model for the number of goals scored, the researcher then suggested a Negative Binomial model for providing a better fit. In the study carried out by Reep and Benjamin (1968), it was concluded that chance dominates the game of football while in the study of association football and statistical inference, Hill (1974) believes football experts were able to predict with some probability of success the final position of the football league table. The researcher concluded that skill dominates the game of football rather than chance. Reep *et al.* (1971) applied the Negative Binomial distribution to the number of goals scored by a team using the English Football League data. The model proved to be a better fit compared to the Poison distribution. Maher (1982) then carried out a further research on the much criticized Poisson model; the researcher adopted an Independent Poisson model which gives a reasonably accurate description of football scores. Dixon and Cole (1997) fitted a parametric model to English Football Leagues by using correction factor to enhance the performance of the Independent Poisson model. Karlis *et al.* (2003) proposed a bivariate Poisson model which allows for correlation between the two scores of opposing teams. Furthermore, instead of directly modeling the number of goals scored, (Karlis and Ntzoufras, 2008) applied Bayesian methods to Skellam's distribution for modeling the differences of the number of goals scored by each team.

Hughes and Franks (2005) analyzed the number of passes responsible for goals scored in two FIFA World Cup finals between 2003 and 2008.Analysis revealed that there were more goals scored from longer passing sequences than from shorter passing sequences and that football teams produced significantly more shots per possession for these longer passing sequences. Bittner *et al*. (2007) analyzed historical football scores data from European leagues and international tournaments using proposed self-affirmation model, the model was found suitable for describing the data in both scenarios. Few years later, Baio and Blangiardo (2010) proposed a Bayesian Hierarchical model for the number of goals scored. The model was applied to football historical data from the Italian Serie A of 1991/1992 season. Firth *et al.* (2013) introduced the Bradley –Terry model for paired comparison data to model results of sporting competitions and applied the model to 2008-2009 Italian Serie A football season. Smith (2014), in presenting his views in PM Sport, opined that big data played some role in Premier League

football clubs. Koopman (2015) developed a statistical model for analyzing and forecasting English Football League. The researchers concluded that the analysis can produce a significant positive return over the bookmakers odd. Giovanni *et al.* (2016) proposed a specification for capturing main characteristics of goal distribution, the work was carried out by making an improvement on the Poisson-based model in the literature, and the proposed methodology was able to produce a good forecast performance. However, when allowing for potential correlation in team level random effects across multiple game-level response from different assumed distributions, Broatch *et al.* (2017) proposed a Multivariate Generalized Linear Mixed Models which can improve accuracy of prediction and inference capability.

Many studies have been carried out on number of goals scores in football matches but much has not been said on data analysis of outcomes (Win, Draw, Loss). This study therefore aimed at proposing a bivariate function for developing a parametric deterministic model for estimating points accumulated by football teams in a League format. A deterministic model is a mathematical representation of a system in which relationships are fixed, e.g., compound interest model, sales forecast model, the inputs of these models can be changed and you get a new answer Wittwer (2004), the author later presented deterministic model as a dictionary of ecology in Wittwer (2014). In section 2, we present a bivariate function and develop a Point Deterministic Model (*PDM*). The bivariate function was studied and used to generalize the Pareto (Type 1) distribution. Section 3 covers analysis of data and presentation of results while sections 4 were devoted for five year comparisons among four Football Leagues in Europe. Section 5 and 6 contain the discussion and conclusion of the study, respectively.

2 Material and Method

2.1 A generalized bivariate function and parametric deterministic model development

The proposed Points Deterministic Model (*PDM)* is based on the possible trivariate outcomes of football matches played and represented as Win, Draw and Loss. A bivariate function was designed in term of the probable outcomes of matches played by football teams in a standard league table. Variable names are then attached to each possible outcome that is used to build the model. The response variable is (*Pi*), which is the expected number of points from total number of matches played. Let x_{ij}, $y_{ij} and\ z_{ij}$ be independent random variables defined as follows:

x_{ij}= number of matches won by team in the league; $i = 1 \dots 20\ teams;\ j = 1 \dots 38\ matches$

y_{ij}= Number of drawn matches by team in the league;

z_{ij}= Number of matches lost by team in the league.

Total number of league matches played during the football season by team i is giving by

$$n_i = x_{ij} + y_{ij} + z_{ij} \tag{2.1}$$

where $n_1 = n_2 \dots = n_{20} = 38$ matches in a football season from $j = 1 \dots n_i$

2.2 Formulation of proposed generalized bivariate function

An expression required to generate the bivariate function using the standardized form of case III suggested by VanUven (1947) in Stuart and Ord (1994, pp. 273) is given as

$f(x_1, x_2) \infty (x_1^2 + x_2^2 + 2bx_1x_2 + 1)^p$, case III (2.2)

We designed eight parameter function, $f(x_{ij}, y_{ij}, z_{ij})$, for the proposed parametric deterministic model (*PDM*) and manipulates the variables into a structured mathematical function using:

$f(x_1, x_2) \infty (x_1^2 + x_2^2 + 2\theta x_1x_2 + \vartheta)^k, \ldots \vartheta = 1$ (2.3)

$$f(x_{ij}, y_{ij}, z_{ij}) = \frac{\alpha^A {x_{ij}}^m}{p} + \frac{\beta^B {y_{ij}}^m}{p} + \frac{\theta^c {z_{ij}}^m}{p}, \quad (2.4)$$

where α = constant coefficient of winning a match;
β = constant coefficient of drawing a match;
θ = constant coefficient of losing a match;
$1/p$ = proportionality constant.
If $m = 2, k = 1, A = B = 0 = \theta = \vartheta$, equations (2.3) and (2.4) are identical. The proposed Parametric Deterministic Model does not depend on the number of lost matches since it has no meaningful statistical contribution to (*Pi*), hence when θ is zero, equation (2.4) reduces to a six parameter bivariate function as in equation (2.5) below.

$$f(x_{ij}, y_{ij}) = \frac{\alpha^A {x_{ij}}^m + \beta^B {y_{ij}}^m}{p} \quad (2.5)$$

where $x_{ij}, y_{ij} \geq 0$, $m = p = 2$, $\alpha > 0$, $\beta > 0$, $A > 0$, $B > 0$. Equation (2.5) is the proposed bivariate function for developing a Parametric Deterministic Linear Model for estimating Total Points and performance of teams in the league.

2.3 Formulation of parametric deterministic model for estimating teams' total points

The technique for developing a fitted model for estimating teams' total points from outcomes of football matches involved taking the sum of partial derivatives of the newly constructed bivariate function, $f(x_{ij}, x_{ij})$, in equation (2.5).

$$f(x_{ij}y_{ij}) = \frac{\alpha^A X_{ij}^2 + \beta^B Y_{ij}^2}{2}$$

$$\frac{\partial f(x_{ij}, y_{ij})}{\partial x_{ij}} = \frac{2\alpha^A x_{ij}}{2} = \alpha^A x_{ij} \quad (2.6)$$

$$\frac{\partial f(x_{ij}, y_{ij})}{\partial y_{ij}} = \frac{2\beta^B y_{ij}}{2} = \beta^B y_{ij} \quad (2.7)$$

$$\frac{\partial f(x_{ij}, y_{ij})}{\partial x_{ij}} + \frac{\partial f(x_{ij}, y_{ij})}{\partial y_{ij}} = \alpha^A x_{ij} + \beta^B y_{ij} \quad (2.8)$$

Let $P_i = \alpha^A x_{ij} + \beta^B y_{ij}$. (2.9)

The possible values of parameters are presented in Table 2.1

Table 2.1: Table of Parameters with different values of A and B

A	B	$\hat{\alpha}$	$\hat{\beta}$
1	1/3	3	1
1/3	1	3	1
1	3	1	1
3	1	1	3

The unique solution where the function has the empirical fixed values as the theoretical values of estimated parameters is at $A = 1$, $B = 1/3$ and $A = 1/3$, $B = 1$, equation (2.9) can be rewritten as;

$$Pi = \hat{\alpha}x_{ij} + \hat{\beta}y_{ij} \quad (2.10)$$

$$Pi = 3x_{ij} + y_{ij}, \quad (2.11)$$

i = team 1, 2,…, n=*20 while* j = matches 1, 2, …, n_i =38.

2.4 Characteristics of the bivariate function

1. If a =1, b=1/3, and a=1/3, b=1, the function has a unique solution for α *and* β
2. The bivariate function can be used to generate the Deterministic Model *iff* $m = p \geq 2$
3. $f(x_{ij}, y_{ij})$ is undefined when $p = 0$.

2.4.1 *Mathematical properties of the proposed PDM,* $Pi = 3x_{ij} + y_{ij}$

1. The model has the capacity of fitting various trivariate outcomes of football matches with the corresponding estimated points.
2. It can be used to determine each football team position in the league table.
3. The total point obtained in a complete football season by any particular football team is given by Pi.
4. The total point obtained in a complete football season by all the football teams in the league table is given by $\sum_{i=1}^{n} Pi$.
5. Pi in the model is a dependent variable while x_{ij} *and* y_{ij} are independent variables
6. The model can be used to analyze the performance of each team in a football league
7. The model estimated points is at the maximum if $x_{ij} = 38$ and $Pi = 3x_{ij} = 114$
8. The model estimated points is at the minimum if $x_{ij} = y_{ij} = 0$ and $Pi = 0$

2.5 Special models from the generalized bivariate function

Let X be a random variable whose cumulative distribution function, CDF is given by F(x), such that

$$F(x) = 1 - \left(\frac{x}{x_0}\right)^{-k} \quad ; \quad x \geq x_0, \; k > 0. \quad (2.12)$$

The corresponding probability density function (pdf) is obtained by differentiating Equation (2.12) with respect to variable x to give

$$f(x) = \frac{k{x_0}^k}{x^{k+1}}; x \geq x_0,\ k > 0 \quad (2.13)$$

Comparing Equation (2.5) with the (pdf) obtained in Equation (2.13)

$f(x_{ij},\ y_{ij}) = \frac{\alpha^a {x_{ij}}^m + \beta^b {y_{ij}}^n}{p}$ and $f(x) = \frac{k{x_0}^k}{x^{k+1}}$, respectively.

1. If $\beta=0$, Equation (2.5) reduces to the event that a team won all its matches:

$$f(x) = \frac{\alpha^a x^m}{p}; \quad (2.14)$$

2. If $a = k = \frac{1}{p}$; Equation (2.14) reduces to :

$$f(x) = k\alpha^k x^m; \quad (2.15)$$

3. If $m = -(k+1)$, Equation (2.15) reduces to :

$$f(x) = \frac{k\alpha^k}{x^{k+1}}. \quad (2.16)$$

Equation (2.16) above is defined by Barry(1983) as the probability density function for a Pareto (Type 1) random variable, x.The continuous random variable,x, with positive support is said to have the Pareto (Type 1) distribution characterized by scale parameter,α, and a shape parameter,k,if its probability density function is given by

$$f(x) = \begin{cases} \frac{k\alpha^k}{x^{k+1}}, & x \geq \alpha,\ k > 0 \\ 0, & x < \alpha \end{cases} \quad (2.17)$$

3. Result, Application and Discussion

The mathematical expression that described the relationship between the dependent and independent variables is the proposed Parametric Deterministic Model (*PDM)* in equation (2.14). The model was applied to historical data of English Premier League *(EPL*) *2*016/2017 football season and the Football League table is produced using the model in Table 3.1. Excel and *R* statistical software were used for data visualization and analysis.

Average Accumulated Point per team is given by$\overline{P_i}=\frac{P_i}{n_i}$, of which the value is 2.4473 for Chelsea football club and 2.2632 for Tottenham. Similar results for all other teams are easily obtained.

Average number of matches won per Team in the *EPL* is $\overline{x_{ij}} = \frac{\sum x_{ij}}{N} = 0.389474$.

Table 3.1 2016/2017 *EPL* Table using *(PDM)* $P_i = \hat{\alpha}x_i + \hat{\beta}y_i$

T_j	Team	n_i	x_i	y_i	z_i	$\hat{\alpha}x_i$	$\hat{\beta}y_i$	P_i
1	Chelsea	38	30	3	5	90	3	93
2	Tottenham	38	26	8	4	78	8	86
3	Man. City	38	23	9	6	69	9	78
4	Liverpool	38	22	10	6	66	10	76
5	Arsenal	38	23	6	9	69	6	75
6	Man. United	38	18	15	5	54	15	69
7	Everton	38	17	10	11	51	10	61
8	Southampton	38	12	10	16	36	10	46
9	AFC Bournemouth	38	12	10	16	36	10	46
10	West Bromwich	38	12	9	17	36	9	45
11	West Ham	38	12	9	17	36	9	45
12	Leicester City	38	12	8	18	36	8	44
13	Stoke City	38	11	11	16	33	11	44
14	Crystal Palace	38	12	5	21	36	5	41
15	Swansea City	38	12	5	21	36	5	41
16	Burnley	38	11	7	20	33	7	40
17	Watford	38	11	7	20	33	7	40
18	Hull City	38	9	7	22	27	7	34
19	Middleborough	38	5	13	20	15	13	28
20	Sunderland	38	6	6	26	18	6	24

Total numbers of 760 matches were played in the League during the season and the estimated League Average Accumulated Point per match

$\overline{P}_i$ is $\frac{\sum P_i}{\Sigma n_i} = 1.389474$,

Variance $(P_i) = \sum \frac{(P_i - \overline{P}_i)^2}{\Sigma n_i} = \sigma_k^2 = 79.40532$, standard deviation $(P_i) = \sigma_k = 8.910966$.

Computation of probability of each trivariate outcome for the (*EPL)* is give in Table 3.2

Table 3.2: Probabilities of total outcomes of football matches in *EPL*

Outcome	**win**	**draw**	**loss**	**total**
Probability	**0.3895**	**0.2210**	**0.3895**	**1.000**

The result from the table reveals that probability of total matches played and won in the league during the football season is the same as probability of total matches played and lost given as 0.3895 while the probability of total drawn matches is 0.2210.

3.1 Validity of the parametric deterministic model

The material for testing the validity is from the historical data of English Premier League of 2016/2017 football season. Applying the model $P_i = \hat{\alpha} + \hat{\beta}y_{ij}$, estimated total points P_i, in the last column for each team in Table 3.1 can be generated. This study provide

example of how the estimates are obtained using two teams; Chelsea and Tottenham football clubs which have the following trivariate outcomes in the premier league during the season.

1. Chelsea $(x_{ij}, y_{ij}, z_{ij}) = (30, 3, 5)$
2. Tottenham $(x_{ij}, y_{ij}, z_{ij}) = (26, 8, 4)$

For the two teams, $i = 1, 2$

$$P_1 = 3(30) + 3 = 90 + 3 = 93 \text{ point} \quad (3.1)$$

$$P_2 = 3(26) + 8 = 78 + 8 = 86 \text{ point} \quad (3.2)$$

By substituting the independent variables x_{ij} and y_{ij} ; estimated P_i which represent the total accumulated points for the two clubs are easily obtained. Table 3.1 presents similar results for all the remaining teams in the *EPL*. The total accumulated Points can be computed at any particular period of the season irrespective of the number of matches played, the model can also be a potential teams' performance analytics tool.

4 Five-Year Comparisons of Four European Leagues

Equation (2.10) given as $P_i = \hat{\alpha}x_i + \hat{\beta}y_i$ for i=1,2,…n where n is total number of football teams can be used to estimate total points for each football team in any particular league. Top leagues in Europe included for this analysis are: English Premier League (*EPL*), La Liga (Spain), Serie A (Italia), and Ligue1 (France). The data set for 5-year comparison contain 100(P_{ik}) for each league while the data frame is made up of 400 (P_{ik}) observations; therefore, comparing different four leagues' performance starts from estimating P_i for each of the leagues such that;

$$P_{ik} = \hat{\alpha}x_{ik} + \hat{\beta}y_{ik}, \quad (4.3)$$

k =1,2…4 football leagues. i =1, 2, …, n football teams

A higher variation in level of competitiveness estimated from the quartiles implies there is less competition among the team because it is obvious that teams with highest level of ability will always be dominating the lesser teams. Smaller value of dispersion implies greater competition because league with this value is made up of teams with similar attributes and tends to cluster or clump together around the median without distinguishable features in strength and performance. Results of data analysis are presented in table 4.3; figure 4.1 is the boxplot for data visualization.

Table 4.3: Rank of League Competitiveness

League	Q3	Q1	D	SD	Mean	Median	Rank
EPL	336.5	197	138.5	86.57	261.95	232	4
LALIGA	316	199	117.0	92.18	262.65	240	2
SERIE A	317	196	121.0	90.75	261.25	245.5	3
LIGUE 1	302.5	206.5	96.0	77.62	260.30	244.5.	1

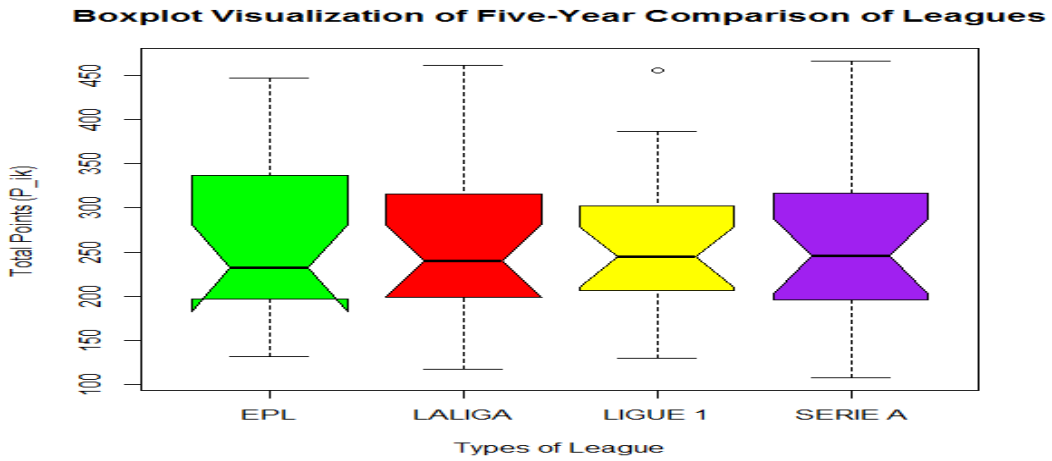

Figure 4.1: Boxplot Visualization of Distributions of each football league

Table 4.3 revealed *EPL* has the highest dispersion from the median compared to other leagues; the overall total points (P_{ik}) for *EPL* are more spread out than the remaining leagues. Data visualization using boxplot in figure 4.1 clearly shows the shape of the distributions and the magnitude of various dispersions of the Leagues from the median.

4.3 Distributions of leagues and model selection

Each football league was subjected to various discrete probability distribution models to ascertain the best model fit, estimate of parameters and statistics of the distribution were also obtained and the result is presented in Table 4.4 and Table 4.5. The Anderson-Darling Statistics was used to select the best probability distribution for each football league. The Anderson-Darling goodness of fit test (AD-Test) is a measure of how well the data fits a specified distribution; that is, the test shows if a sample of data is from a population with specific distribution. Considering the (AD-Test) statistics at 5 percent significant value, the *p*-values for the test are all greater than $\alpha = 0.05$. The distribution of each of the six team at the top of the league table playing either in the Europa or UEFA league follows the specified theoretical distributions in Table 4.4 by analysis obtained using Easy-Fit Distribution Fitting Software. For the six teams at tail end of the leagues, the theoretical distributions for the sample data are contained in Table 4.5.

Table 4.4: Distributions for Teams in European League

League	Distribution	Parameters	K-S	AD
EPL	Negative Binomial	n=9, p= 0.0349	0.1298	0.5159
LALIGA	Negative Binomial	n= 8, p=0.0391	0.1296	0.3561
SERIE A	Negative Binomial	n= 8, p=0.0317	0.1104	0.4671
LIGUE 1	Negative Binomial	n=11, p=0.0432	0.1501	0.6277

Table 4.5: Distributions for Lower Six Teams in Each League

League	Distribution	Parameter	K-S	AD
EPL	Negative Binomial	n= 125, p=0.2489	0.1264	0.1569
LALIGA	Discrete Uniform	a= 24, b= 481	0.1313	0.1461
SERIE A	Discrete Uniform	a= 279, b=469	0.1754	0.3369
LIGUE 1	Negative Binomial	n= 52, p=0.1276	0.1341	0.1985

4.4 Sub-league of top six teams for UEFA competitions

Qualification for UEFA competition involves teams fighting for position in the upper quintile of the league distribution. Top six comparisons were carried out to estimate inferential statistics about the ability of teams in the top six sub-leagues. The result is summarized with table 4.6. There is a very stiff competition for UEFA champions' league and the UEFA Europa league football among the teams in *EPL* compare to other leagues. The variability in the performance of teams in LALIGA is 57.34; the highest revealing that the top six positions of teams in LALIGA is the least competitive among the leagues compared over the 5-year period.

Table 4.6: Rank for Top Six Positions in the four Leagues

League	Q3	Q1	D	SD	Mean	Median	Rank
EPL	401	347	54	38.98	378.17	374	1
LALIGA	438	329	109	57.34	382.33	381.5	4
SERIE A	415	329	86	55.29	374.17	365.0	3
LIGUE 1	387	310	77	53.07	359.33	354.0	2

4.5 Relegation battle among teams in the lower sub-league

Three teams are mostly relegated from the league due to their poor performance at the end of the season. Survival from relegation to the lower division of football league was analyzed among the bottom five teams in the lower quartile of the twenty -team leagues. The numerical result is presented in the Table 4.7.

Table 4.7: Results of Bottom Teams Relegation Battle

League	Distribution	Q3	Q1	D	SD	Mean	Median	Rank
EPL	Discrete uniform	194	160	34	23.32	174.50	180.5	1
LALIGA	Discrete uniform	193	145	48	30.75	170.67	182.0	3
SERIE A	Discrete uniform	191	135	56	32.78	164.67	176.5	4
LIGUE 1	Discrete uniform	204	164	40	27.62	182.50	194.0	2

The table shows teams fighting for survival from relegation to lower football division are more competitive in the premier league followed by the Ligue1 while the least competition for survival is the characteristics among teams in the Serie A. Based on the variability of structure of teams in each league, survival is most competitive in *EPL* and the least competitive in Serie A for the five-year comparison.

5. Summary and Conclusion

The study revealed that a bivariate statistical function can be used to generalize a univariate distribution. Most literatures on football matches were dominated by analysis of goals score in football matches. Moving away from analysis of goals scored, interest in this study is centered on developing a logical mathematical function from the bivariate Pearson family system of distribution for the trivariate outcomes of matches played and to examine the capacity of the mathematical function as a generalization of other useful models. A bivariate mathematical function for developing a statistical Parametric Deterministic Model that can be used to generate accumulated Points ($\bar{\boldsymbol{P}}_{\boldsymbol{i}}$) from outcomes of football matches played was successfully designed. Application of the model to *EPL* data of 2016/2017 season produced accurate estimates of Points for the end of the football season as shown in Table 3.1. Expectation of football data analysts is that either the average of matches won in a more competitive league be very close with less dispersion or that the teams have a heavy cluster around the center of the league compared to what is obtained in a less competitive league, competitiveness is a function of matches won and lost in the league. Four football leagues in Europe (*EPL, LALIGA, ITALIA, LIGUE1*) were selected for performance and model comparisons, analysis of data showed a higher percentage of matches won by teams in the *EPL* compared to other leagues. Estimated values of dispersion for each football league are presented in Table 3.3; dispersion among football teams in *La LIGA* is highest and the lowest in *LIGUE1*. The four leagues were run through model selection from various discrete distributions and result revealed all the European leagues follow a Negative Binomial (*NB*) distribution. Sub-leagues of teams at the lower tail of *EPL* and Ligue1 that are competing for survival from relegation to lower league division is best fitted by the Discrete Uniform distribution while LALIGA and Serie A are well fitted by the Negative Binomial Model.

The proposed generalized bivariate function represents a good framework for statistical analysis of data from outcomes of football matches, the bivariate function as revealed provides a good generalization for the Pareto Type 1 distribution in Barry (1983). The newly proposed model forms an adequate mathematical expression for both the response and explanatory random variables required for estimating football points and for producing the league table. Analysis of goals scores can be incorporated in further studies for proper ranking of teams having equal estimated accumulated points. Five-years historical data from four European leagues up to the season of 2017/2018 shows *LALIGA* is least competitive while *LIGUE1* is most competitive. The analytical results however digressed from the opinion of many football fans who thinks the widely celebrated *EPL* and *LALIGA* are the most competitive leagues during the period covered in this study. When analyzing sub-leagues of top six teams in each of the four leagues, study reveals top six team in the *EPL* is most competitive in Europe.

References

Baio, G. Blangiardo, M. (2010). Bayesian Hierarchical Model for the Prediction of Football Results. *Journal of Applied Statistics*, 37(2), 253-264, DOI: 10.1080/02664760802684177.

Barry C.A. (1983). *Pareto Distribution*, International Cooperative Publ. House.

Bittner, A., Nussbaumer, A., Janke, W. and Weigel, M. (2007). Self –Affirmation Model for Football Goal Distributions, *Europhysics Letters Association*, arXiv:0705-2724v1.1-6. DOI: 10.1209/0295-5075/78/58002.

Broatch, J.E. and Karl, A.T. (2017). Multivariate Generalized Linear Mixed Models for Joint Estimation of Sporting Outcomes, *Italian Journal of Applied Statistics*, 30(2), 189-211.

Cattelan, M., Varin, C. and Firth, D. (2013). Dynamic Bradley – Terry Modeling of Sports Tournaments, *Journal of the Royal Statistical Society*, Series *C*, 62, 135-150.

Dixon, M. and Cole, S. (1997). Modeling Association Football Scores and Inefficiencies in the Football Betting Market, *Journal of the Royal Statistical Society, Series* C, 46, 265 – 280.

Eugene N., Lee, C and Famoye, F. (2002). Beta-Normal Distribution and its Applications, *Communications in Statistics-Theory Methods,* 31, 497-512.

Giovanni, A. and Luca, D.A. (2016). PARX Model for Football Match Predictions, *Journal of Forecasting,* DOI; 1002/for.2471.

Hill, I.D. (1974). Association Football and Statistical Inference, *Applied Statistics*, 23(2), 203-208.

Hughes, M. and Franks, I. (2005*)*. Analysis of Passing Sequences, Shots and Goals in Soccer, *Journal of the Royal Statistical Society*, Series A, Vol. 134, DOI: 10.1080/02640410410001716779.

Johnson N.L (1949). Systems of Frequency Curves Generated by Translation, *Biometrika*, 36, 149-176.

Karlis, D. and Ntzoufras, I. (2003). Analysis of Sports Data Using Bivariate Poisson Models, *Journal of the Royal Statistical Society*, Series D, 52,381-393, DOI:10.1111/1467-9884.00366.

Karlis, D. and Ntzoufras, I. (2008). Bayesian Modeling of Football Outcomes: Using the Skellam's Distribution for the Goals Difference, *IMAJ Management Math*, 20 (2), 133 – 145.

Maher, M.J. (1982). Modeling Association Football Scores, *Statistica Neeerlandica,* 36, 109 – 118, DOI:10.1111/j.1467-9574.1982.tb00782.x.

Marshall, A.N. and Olkin, I. (1997). A New Method of Adding Parameter to a Family of Distributions with Applications to the Exponential and Weibull Families, *Biometrika,* 84, 64 1-652.

Moroney, M.J. (1957). Facts from Figures, *The Journal of the American Statistical Association,* 52 (277), 623 – 629.

Pearson, K. (1895). Contribution to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material, *Philos Trans. R. Soc. Lond. A*, 186, 343-414.

Reep, C. and Benjamin, B. (1968). Skill and Chance in Association Football, *Journal of the Royal Statistical Society*, 131, 581-585.

Reep, C., Pollard, and Benjamin, B. (1971). Skill and Chance in Ball Games, *Journal of the Royal Statistical Society* A, 134, 623-629.

Siem J.K. and Rutger, L. (2015). A Dynamic Bivariate Poisson Model for Analyzing and Forecasting Match Results in the English Premier League, *Journal of the Royal Statistical Society, Series A,* 178 (1), 167-186, DOI:10.1111/rssa.12042.

Smith, M. (2014). How Big Data Gives Premier League Football Clubs and Edge, *PM Sports.*

Stuart A. and Ord K.J. (1994). *Kendall's Advanced Theory of Statistics, Distribution Theory*, Vol. 1 Sixth Edition, Oxford University Press Inc., NY 10016.

VanUven, M.J. (1947). Extension of Pearson's Probability Distribution to Two Variables, I-IV. *Nedelandsche Akademie Van Wetenschappen,* 50(1947), 191-196.

Wittwer J.W. (2014). Deterministic Model, A Dictionary of Ecology, *encyclopedia.com 25. http://www.encyclopodia.com*

Wittwer, J.W. (2004). Deterministic Model Example: Compound Interest, *https://www.Vertex42.com.*