



Comparison among some European football leagues through related count data variables

Valentina Cueva-López, José Rodríguez-Avi, María José Olmo-Jiménez

University of Jaén

1-3 July, 2019

Contents

- 1 Introduction
- 2 Count Data Models
- 3 Description of Data
- 4 Summary of Data
- 5 Data Fits

Introduction

Introduction

- It can be considered that football, as the most popular sport in modern history, spans more than 150 years.
- It began in 1863 in England, when rugby football and association football branched off on their different courses. Thus, the most ancient football association was founded.



Introduction

- In its origins, controlling the ball with the feet was considered a feat, since it raised the admiration of other citizens.



- The first reference of the game is a manual of military exercises that dates back to the China of the Han dynasty in the 2nd and 3rd centuries BC, known as "Ts'uh Kúh".

Introduction

We can also mention the Japanese Kemari which began some 500-600 years later and is still played today.



Introduction

More lively were the "Epislcyros" Greek and the "Harpastum" Roman.



Introduction

- Undoubtedly, football is nowadays the most popular sport in the world.
- It is important not only on the sport level, as a game and pastime, but also on the social level, since it joins people, social groups, clubs or even countries.
- Football is one of the sports which generates more money in Europe, Latin America, Asia and, recently, in United States.

Count Data Models

Count Data Models

- Negative Binomial distribution [3]: $X \sim NB(\theta, \mu)$ with $\theta, \mu > 0$ and probability mass function (pmf) given by

$$P(X = x) = \frac{\Gamma(\theta + x)}{\Gamma(\theta)x!} \left(\frac{\theta}{\theta + \mu} \right)^\theta \left(\frac{\mu}{\theta + \mu} \right)^x, \quad x = 0, 1, \dots$$

Count Data Models

- Generalized Poisson distribution [1]:

$X \sim GP(\lambda, \theta)$, $\lambda > 0$, $\max(-1, -\lambda/m) < \theta < 1$ with pmf given by

$$P(X = x) = \begin{cases} \frac{\lambda(\lambda + \theta x)^{x-1}}{x!} e^{-\lambda - \theta x} & x = 0, 1, \dots \\ 0 & x > m \end{cases}$$

where $m \geq 4$ is the largest positive integer for which $\lambda + m\theta > 0$ when $\theta < 0$.

Count Data Models

- Univariate Generalized Waring distribution [2, 7]:
 $X \sim UGW(a, k, \rho)$ with $a, k > 0$, $\rho > 2$ and pmf given by

$$P(X = x) = \frac{\Gamma(a + \rho)\Gamma(k + \rho)}{\Gamma(a)\Gamma(k)\Gamma(\rho)} \frac{\Gamma(a + x)\Gamma(k + x)}{\Gamma(a + k + \rho + x)\Gamma(x + 1)}, \quad x = 0, 1, \dots$$

Count Data Models

- Complex Biparametric Pearson distribution [4, 5]:
 $X \sim CBP(b, \gamma)$ with $b, \gamma > 0$ and pmf given by

$$P(X = x) = \frac{\Gamma(\gamma + bi)\Gamma(\gamma - bi)}{\Gamma(\gamma)^2} \frac{(bi)_x(-bi)_x}{(\gamma)_x} \frac{1}{x!}, \quad x = 0, 1, \dots$$

where i is the imaginary unit and
 $(\alpha)_r = \Gamma(\alpha + r)/\Gamma(\alpha)$, $\alpha > 0$.

Count Data Models

- Complex Triparametric Pearson distribution [6]:
 $X \sim CTP(a, b, \gamma)$ with $a \in \mathbb{R}$, $b, \gamma > 0$ and pmf given by

$$P(X = x) = f_0 \frac{(a + ib)_x (a - ib)_x}{(\gamma)_x} \frac{1}{x!}, \quad x = 0, 1, \dots$$

where f_0 is the normalizing constant whose expression is

$$f_0 = \frac{\Gamma(\gamma - a - ib)\Gamma(\gamma - a + ib)}{\Gamma(\gamma)\Gamma(\gamma - 2a)}.$$

Description of Data

Description of Data I

We focus on the most famous European football leagues which are the Spanish, German, Italian and English football leagues and we have included the two teams that more titles of the national championship have achieved throughout the history of the league of their country.

- Spanish football league from 1970 to 2018



(a) Real Madrid



(b) Barcelona FC

Description of Data II

- English football league from 1992 to 2014



(c) Manchester United FC



(d) Liverpool FC

Description of Data III

- Italian football league from 1988 to 2018



(e) Juventus de Turin



(f) AC Milan

Description of Data IV

- German football league from 1985 to 2014



(g) Bayern de Múnich



(h) Borussia Dortmund

Description of Data

The variables selected for the study are:

- Number of yellow cards received by a footballer in the corresponding team (along the span time considered).
- Number of goals scored by a footballer in the corresponding team (along the span time considered).

Summary of Data

Summary of Data

Spanish league							
Number of yellow cards							
Statistics	Min.	Q_1	Median	Mean	Q_3	Max.	s.d.
Barcelona	0.00	0.00	2.00	8.20	9.00	81.00	13.92
Real Madrid	0.00	0.00	2.00	8.59	11.00	111.00	15.60
Number of goals scored							
Statistics	Min.	Q_1	Median	Mean	Q_3	Max.	s.d.
Barcelona	0.00	0.00	2.00	10.99	9.75	383.00	27.86
Real Madrid	0.00	0.00	1.50	11.19	8.75	311.00	30.55

Table: Descriptive summary of data

Summary of Data

English league							
Number of yellow cards							
Statistics	Min.	Q_1	Median	Mean	Q_3	Max.	s.d.
Manchester United	0.00	0.00	1.00	7.56	8.00	96.00	14.26
Liverpool	0.00	0.00	2.00	5.89	7.00	66.00	9.18
Number of goals scored							
Statistics	Min.	Q_1	Median	Mean	Q_3	Max.	s.d.
Manchester United	0.00	0.00	2.00	12.46	10.00	158.00	25.88
Liverpool	0.00	0.00	2.00	8.30	7.25	128.00	18.97

Table: Descriptive summary of data

Summary of Data

Italian league							
Number of yellow cards							
Statistics	Min.	Q_1	Median	Mean	Q_3	Max.	s.d.
Juventus	0.00	0.00	3.00	7.15	9.00	59.00	10.94
Milan	0.00	0.00	3.00	7.01	8.00	97.00	12.96
Number of goals scored							
Statistics	Min.	Q_1	Median	Mean	Q_3	Max.	s.d.
Juventus	0.00	0.00	2.00	7.27	7.00	188.00	17.65
Milan	0.00	0.00	1.00	6.51	6.00	127.00	14.44

Table: Descriptive summary of data

Summary of Data

German league							
Number of yellow cards							
Statistics	Min.	Q_1	Median	Mean	Q_3	Max.	s.d.
Bayern Munich	0.00	1.00	4.00	9.36	13.00	59.00	12.50
Borussia Dortmund	0.00	0.00	2.00	7.94	11.75	66.00	12.02
Number of goals scored							
Statistics	Min.	Q_1	Median	Mean	Q_3	Max.	s.d.
Bayern Munich	0.00	0.00	4.00	11.79	12.00	107.00	19.75
Borussia Dortmund	0.00	0.00	2.00	7.89	8.00	116.00	15.78

Table: Descriptive summary of data

Fit of Data

Fit of Data: Yellow Cards

	Football team			
	<i>Bayern Munich</i>	<i>Borussia Dortmund</i>	<i>Manchester United</i>	<i>Liverpool</i>
Model	AIC			
<i>NB</i>	1099.58	1253.97	819.95	1000.25
<i>GP</i>	1112.53	1261.56	829.51	1000.49
<i>CBP</i>	1164.12	1310.62	859.94	1033.83
<i>UGW</i>	1101.60	1256.04	821.96	1001.91
<i>CTP</i>	1129.08	1277.71	841.56	1009.94
Model	<i>p</i> -value			
<i>NB</i>	0.63	0.04	0.34	0.76
<i>GP</i>	0.06	0.11	0.06	0.50
<i>CBP</i>	0.00	0.00	0.00	0.00
<i>UGW</i>	0.56	0.30	0.26	0.70
<i>CTP</i>	0.00	0.00	0.00	0.13

Table: AIC value and *p*-value of χ^2 -goodness of fit test for fits about yellow cards data

Fit of Data: Yellow Cards

	Football team			
	<i>Barcelona</i>	<i>Real Madrid</i>	<i>Juventus</i>	<i>Milan</i>
Model	AIC			
<i>NB</i>	2081.58	2041.53	1492.48	1603.37
<i>GP</i>	2093.52	2061.78	1511.09	1600.09
<i>CBP</i>	2170.39	2143.54	1584.37	1656.69
<i>UGW</i>	2083.60	2043.50	1494.50	1604.00
<i>CTP</i>	2118.49	2087.93	1532.02	1612.32
Model	<i>p</i> —value			
<i>NB</i>	0.63	0.92	0.12	0.23
<i>GP</i>	0.06	0.03	0.00	0.36
<i>CBP</i>	0.00	0.00	0.00	0.00
<i>UGW</i>	0.57	0.89	0.09	0.19
<i>CTP</i>	0.00	0.00	0.00	0.00

Table: AIC value and *p*—value of χ^2 —goodness of fit test for fits about yellow cards data

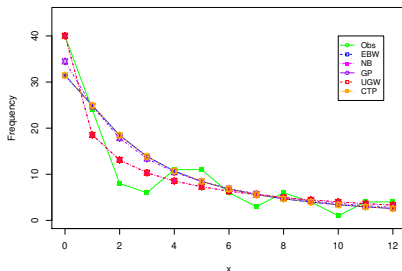
Fit of Data: Yellow Cards

<i>Bayern Munich</i>	<i>Borussia Dortmund</i>	<i>Manchester United</i>	<i>Liverpool</i>
<i>NB</i>			
$\hat{\theta} = 0.485(0.056)$	$\hat{\theta} = 0.436(0.047)$	$\hat{\theta} = 0.280(0.037)$	$\hat{\theta} = 0.518(0.062)$
$\hat{\mu} = 9.358(1.047)$	$\hat{\mu} = 7.943(0.852)$	$\hat{\mu} = 7.560(1.189)$	$\hat{\mu} = 5.894(0.637)$
<i>Barcelona</i>	<i>Real Madrid</i>	<i>Juventus</i>	<i>Milan</i>
<i>NB</i>		<i>GP</i>	
$\hat{\theta} = 0.354(0.029)$	$\hat{\theta} = 0.322(0.027)$	$\hat{\theta} = 0.405(0.041)$	$\hat{\lambda} = 1.355(0.091)$
$\hat{\mu} = 8.204(0.745)$	$\hat{\mu} = 8.590(0.823)$	$\hat{\mu} = 7.154(0.716)$	$\hat{\theta} = 0.806(0.022)$

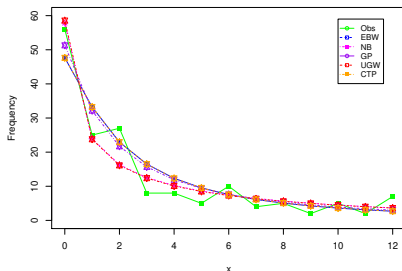
Table: MLEs and standard errors (in brackets) for the best fit of yellow cards data

Fit of Data: Yellow Cards

Figure: Observed and expected frequencies for data about the number of yellow cards



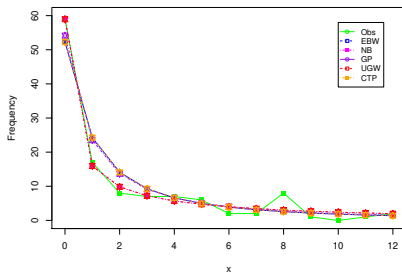
(a) Bayern Munich



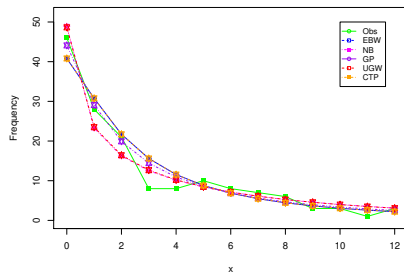
(b) Borussia Dortmund

Fit of Data: Yellow Cards

Figure: Observed and expected frequencies for data about the number of yellow cards



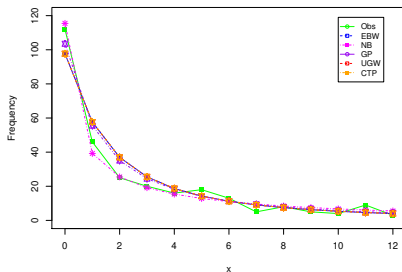
(a) Manchester United



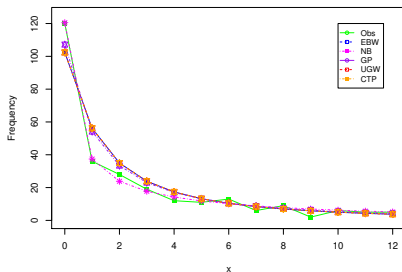
(b) Liverpool

Fit of Data: Yellow Cards

Figure: Observed and expected frequencies for data about the number of yellow cards



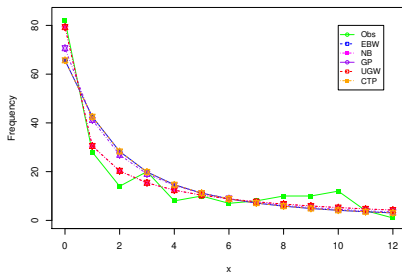
(a) Barcelona



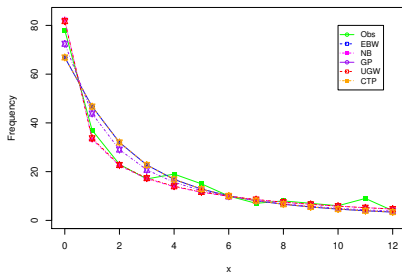
(b) Real Madrid

Fit of Data: Yellow Cards

Figure: Observed and expected frequencies for data about the number of yellow cards



(a) Juventus



(b) Milan

Fit of DataYellow Cards. Conclusion I

- According to the AIC, the best model is the *NB* for all the teams except for the Milan football club which selects the *GP* model.
- This performance is more irregular according to the χ^2 —goodness of fit test, since there are several cases in which it disagrees with the AIC.
 - The Borussia Dortmund team the best fit, according to the test, is that provided by the *UGW* model; however, the fit with lowest AIC is the corresponding to the *NB* model.
 - For the Milan football club both criteria agree on the *GP* model.
 - For the rest of the teams the *NB* distribution is the best model with both criteria.

Fit of DataYellow Cards. Conclusion II

- It should be emphasized that in some teams the yellow cards data are adequately modelled by several distributions, such as the Liverpool football club, in which - in addition to the *NB* distribution - the *UGW* and *CTP* distributions are also suitable.

Fit of Data:Goals Scored

Model	Football team			
	<i>Bayern Munich</i>	<i>Borussia Dortmund</i>	<i>Manchester United</i>	<i>Liverpool</i>
AIC				
<i>EBW</i>	1151.47	1189.71	810.80	924.01
<i>NB</i>	1128.00	1170.98	796.76	924.36
<i>GP</i>	1138.86	1177.41	1177.41	917.88
<i>CBP</i>	1191.13	1218.55	828.55	941.92
<i>UGW</i>	1130.00	1172.29	6003.01	7483.96
<i>CTP</i>	1153.47	1191.71	812.80	926.01
Model	<i>p</i> -value			
<i>EBW</i>	0.00	0.04	0.00	0.42
<i>NB</i>	0.45	0.86	0.33	0.44
<i>GP</i>	0.04	0.33	0.33	0.75
<i>CBP</i>	0.00	0.00	0.00	0.00
<i>UGW</i>	0.38	0.82		
<i>CTP</i>	0.00	0.04	0.01	0.27

Table: AIC value and p -value of χ^2 -goodness of fit test for fits about goals scored data.

Fit of Data:Goals Scored

Model	Football team			
	<i>Barcelona</i>	<i>Real Madrid</i>	<i>Juventus</i>	<i>Milan</i>
AIC				
<i>EBW</i>	1956.44	1863.66	1300.40	1355.69
<i>NB</i>	1932.34	1865.35	1289.98	1345.64
<i>GP</i>	917.88	917.88	917.88	1344.33
<i>CBP</i>	941.92	941.92	941.92	1383.34
<i>UGW</i>	1930.64	1866.53	1289.74	1347.59
<i>CTP</i>	926.01	926.01	926.01	1357.69
<i>p</i> –value				
<i>EBW</i>	0.00	0.00	0.04	0.34
<i>NB</i>	0.46	0.63	0.67	0.67
<i>GP</i>	0.75	0.06	0.75	0.87
<i>CBP</i>	0.00	0.00	0.00	0.00
<i>UGW</i>	0.50	0.57	0.66	0.60
<i>CTP</i>	0.27	0.00	0.27	0.28

Table: AIC value and *p*–value of χ^2 –goodness of fit test for fits about goals scored data.

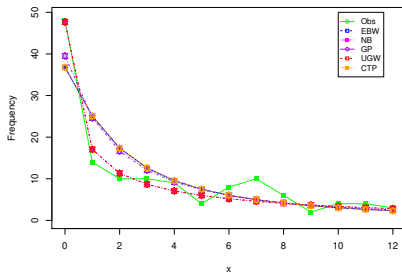
Fit of Data:Goals Scored

<i>Bayern Munich</i>	<i>Borussia Dortmund</i>	<i>Manchester United</i>	<i>Liverpool</i>
<i>NB</i>		<i>GP</i>	
$\hat{\theta} = 0.368(0.033)$	$\hat{\theta} = 0.296(0.042)$	$\hat{\theta} = 0.295(0.033)$	$\hat{\lambda} = 1.410(0.117)$
$\hat{\mu} = 11.792(1.50)$	$\hat{\mu} = 7.586(1.019)$	$\hat{\mu} = 7.886(1.02)$	$\hat{\theta} = 0.761(0.030)$
<i>Barcelona</i>	<i>Real Madrid</i>	<i>Juventus</i>	<i>Milan</i>
<i>GP</i>			
$\hat{\lambda} = 1.089(0.091)$	$\hat{\lambda} = 1.088(0.092)$	$\hat{\lambda} = 1.411(0.117)$	$\hat{\lambda} = 0.956(0.074)$
$\hat{\theta} = 0.861(0.023)$	$\hat{\theta} = 0.862(0.024)$	$\hat{\theta} = 0.861(0.023)$	$\hat{\theta} = 0.853(0.024)$

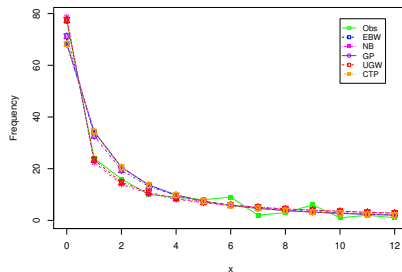
Table: MLEs and standard errors (in brackets) for the best fit of goals scored data

Fit of Data: Goals Scored

Figure: Observed and expected frequencies for data about the number of goals scored



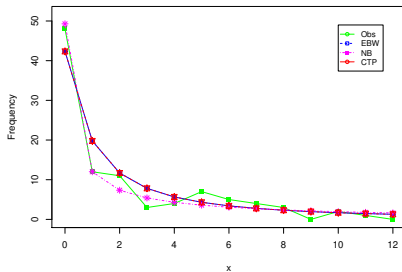
(a) Bayern Munich



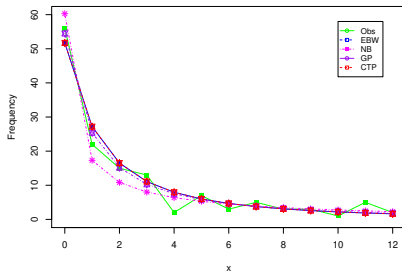
(b) Borussia Dortmund

Fit of Data: Goals Scored

Figure: Observed and expected frequencies for data about the number of goals scored



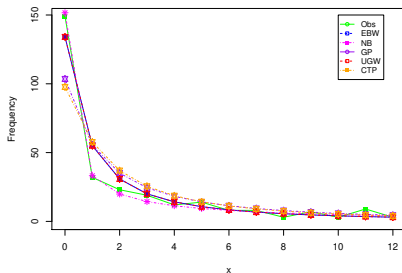
(a) Manchester United



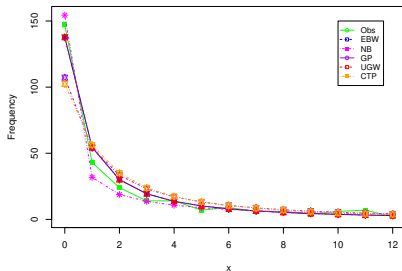
(b) Liverpool

Fit of Data:Goals Scored

Figure: Observed and expected frequencies for data about the number of goals scored



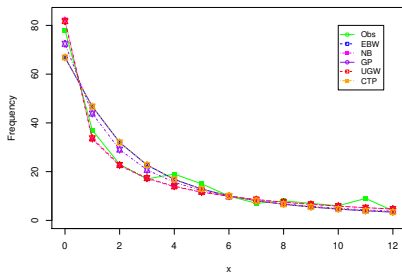
(a) Barcelona



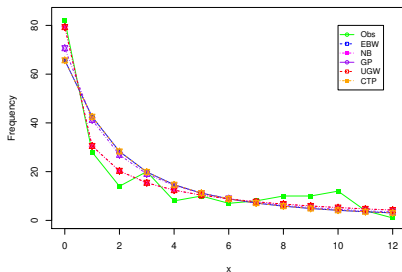
(b) Real Madrid

Fit of Data:Goals Scored

Figure: Observed and expected frequencies for data about the number of goals scored



(a) Juventus



(b) Milan

Fit of Data:Goals Scored. Conclusion I

- The case of Manchester United can be considered special, since according to the p -value of the goodness of fit test, both the GP and NB distributions could be adequate models.
- For the goals scored data corresponding to the Bayern Munich team there are two appropriate models, the NB and the UGW , although the NB fit is the best one according to the AIC.
- Besides these two distributions, the Borussia Dortmund and Real Madrid teams add the GP distribution, although the NB fit continues being the best in terms of the AIC.
- The Barcelona and Juventus teams also add the CTP distribution, although the best fit is that related to the GP distribution.

Fit of Data:Goals Scored. Conclusion II

- Goals data for the Liverpool team can be modelled by the majority of the distributions used (the only ones that do not fit appropriately are the *CBP* and *UGW* distributions). However, in this case, the *NB* fit is not the best, as in the previous cases, but the *GP* fit.
- Finally, we have another particular case, the Milan football club, since for this team only the *CBP* distribution does not provide an appropriate fit.

Bibliography I



P. C. Consul and F. Famoye.

Maximum likelihood estimation for the generalized Poisson distribution when sample mean is larger than sample variance.

Communications in Statistics: Theory and Methods,
17:299–309, 1988.






J.O. Irwin.

The generalized Waring distribution applied to accident theory.

Journal of the Royal Statistical Society. Series A, 131:205–225,
1968.

Bibliography II

-  N. L. Johnson, A. W. Kemp, and S. Kotz.
Univariate discrete distributions.
Wiley, New York, 3rd edition, 2005.
-  J. Rodríguez-Avi, A. Conde-Sánchez, and A. J. Sáez-Castillo.
A new class of discrete distributions with complex parameters.
Statistical Papers, 44:67–88, 2003.
-  J. Rodríguez-Avi and M.J. Olmo-Jiménez.
A regression model for overdispersed data without too many zeros.
Statistical Papers, 58:749–773, 2017.

Bibliography III



J. Rodríguez-Avi, M.J. Olmo-Jiménez, and V. Cueva-López.
A review of the CTP distribution: a comparison with other
over- and underdispersed count data models.
Journal of Statistical Computation and Simulation, 2018.



E. Xelakaki.
The univariate generalized Waring distribution in relation to
accident theory: proneness, spells or contagion?
Biometrics, 39:887–895, 1983.