Yellow fever: investigating referee consistency in the 'Big 5' men's European soccer leagues

Pete Philipson

Newcastle University, UK

NESSIS 2023 - Saturday 23rd September

Overview

1. Some background and exploratory data analysis
2. Modelling the data
3. Results

Some background and exploratory data analysis

Background & motivation

- Interested in a different take on soccer, and working on small count data liable to underdispersion led me to think about yellow cards
- Yellow cards are issued for a moderate disciplinary infraction
- Previous work (using bivariate Poisson and negative binomial models) found evidence of refereeing bias in the English Premier League in the period 1996-2003 [?]
- It was subsequently found that the number of yellow and red cards received reduced the chance of victory in an ordered probit regression model using data from the Bundesliga [?]

The role of the referee

- In soccer the referee has a somewhat thankless task . . .

  "The trouble with referees is that they know the rules but they do not know the game." - Bill Shankly

  "Italian referees can be compared to sheriffs in the Wild West: trying to impose the increasingly flimsy authority of law and order in the face of mistrust, hostility and violence." - From Calcio by John Foot.

- 'Akkiappa!' was a game created by the then-president (!) of Como, Enrico Preziosi - essentially whack-a-mole but with referees - so incensed was he by refereeing decisions against his team

Application: Yellow cards in the 'Big 5' European leagues (2018/19 - 2021/22)

- Consider data on the number of yellow cards shown by referees in the 'Big 5' European leagues between seasons 2018/19 and 2021/22
- This gives $\approx 7.5K$ matches with two observations for each match, played between $k = 129$ teams, overseen by $m = 171$ referees.
- We also have data before, during, and after the impact of Covid-19

Application: Yellow cards in the 'Big 5' European leagues (2018/19 - 2021/22)

- The data on yellow cards (and referees in the case of the EPL) are publicly available at `https://www.football-data.co.uk/`
- Data on referees and crowds are publicly available at `https://fbref.com/en/`
- The two resources can be combined in a fun data-wrangling exercise

Application: Exploratory data analysis

- Overall mean for yellow cards by league is
  1. 1.60 with a variance of 1.21 in the Premier League (EPL);
  2. 1.91 with a variance of 1.28 in Ligue 1 (L1);
  3. 1.82 with a variance of 1.30 in the Bundesliga (BL);
  4. 2.36 with a variance of 1.37 in Serie A (SA);
  5. 2.49 with a variance of 1.50 in La Liga (LL).
- The mean exceeds the variance for each league $\Longrightarrow$ underdispersion
- This ignores the effects of covariates, but suggests underdispersion is a real phenomenon for these data
- Referees are expected to be alike, teams are expected to be different . . .

Application: Exploratory data analysis

- How are the counts dispersed at the referee level?
- The data are made up of 171 (EPL - 27, L1 - 32, BL - 30, SA - 56, LL - 26) - likely heterogeneous - referees

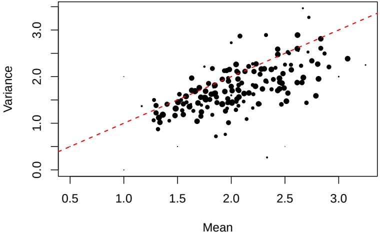


Figure 1: Variance against mean for yellow cards for each referee in the 'Big 5' leagues

Application: Exploratory data analysis

- Are the number of yellow cards given to home and away teams independent?
- Initially explore this using Kendall's τ_b, which can be easily calculated in R; around a quarter of the paired observations are ties, and τ_b adjusts for this

Table 1: Values of τ_b for home and away yellow cards

League	τ_b	p-value
EPL	0.15	< 0.001
L1	0.08	< 0.001
BL	0.18	< 0.001
SA	0.12	< 0.001
LL	0.17	< 0.001

Application: Exploratory data analysis

To explore possible dependency we can look at the ratio of the empirical joint bivariate probabilities to the empirical independent bivariate probabilities:

		Away yellows					
		0	1	2	3	4	≥ 5
Home yellows	0	2.07	1.09	1.01	0.75	0.46	0.41
	1	1.10	1.15	1.05	0.87	0.81	0.70
	2	0.91	0.99	1.03	1.03	1.03	0.90
	3	0.57	0.95	0.93	1.13	1.28	1.41
	4	0.41	0.73	0.97	1.25	1.42	1.78
	≥ 5	0.42	0.60	0.80	1.37	1.71	2.18

Modelling the data

Application: Yellow cards in the 'Big 5' European leagues (2018/19 - 2021/22)

- In order to conduct an analysis that properly accounts for the features of the data (i.e. potential dependence between home and away cards) we desire:
    1. A bivariate count data regression model capable of handling both over and underdispersion;
    2. Standard inferences for the covariates;
    3. Routine implementation - preferably broadly comparable to bivariate Poisson, negative binomial models in terms of CPU time;

Application: Yellow cards in the 'Big 5' European leagues (2018/19 - 2021/22)

- We can also think about the covariates for the model itself
- We would like to have individual (i.e. referee) level effects for the mean and possibly the dispersion
- Also may wish to consider
  1. the effect of being at home;
  2. whether the absence of fans due to the COVID-19 pandemic made a material difference to the number of cards shown;
  3. whether the effect of being behind closed doors differed for the nominal home and away teams;
  4. differences between teams;
  5. differences between leagues;
  6. changes over time (due to rule changes, game focus etc.)

The Conway-Maxwell-Poisson distribution

- Routinely adopted count distributions were developed to handle the more common overdispersion.
- The Conway-Maxwell-Poisson (CMP) distribution generalises the Poisson distribution, with an extra parameter to account for possible over- or underdispersion, or *both* (bidispersion).
- First introduced in 1962 in the context of queueing [**?**].
- Little footprint in the statistical literature prior to 2005, but has gained some traction recently.

Why does handling bidispersion matter?

- Commonly, the Poisson model is too liberal in the presence of overdispersion
- This is typically remedied by a negative binomial (NB) model, or some other Poisson mixture model (or through a zero-altered model)
- In the presence of both types of dispersion the Poisson model will be both conservative and liberal
- An NB model can only help with the overdispersion and is, at best, the same as the Poisson in cases of underdispersion

CMP definition

- The CMP distribution has probability mass function given by

$$\Pr(X = x \mid \lambda, \nu) = \frac{\lambda^x}{(x!)^\nu} \frac{1}{G_\infty(\lambda, \nu)}$$

  with $x = 0, 1, 2, \ldots, \lambda > 0$ and $\nu \geq 0$.
- λ is the rate parameter and ν models the dispersion.
- $G_\infty(\lambda, \nu) = \sum_{r=0}^{\infty} \lambda^r / (r!)^\nu$ ensures that the CMP distribution is proper.
- Several classic discrete distributions are special cases
  1. Poisson ($\nu = 1$)
  2. Geometric ($\nu = 0$)
  3. Bernoulli ($\nu \longrightarrow \infty$)

The mean-parameterised CMP distribution

- CMP distribution does not provide 'nice' inferences.
- No mean parameter so can only talk about general trends.
- Can reparameterise in terms of the mean μ

$$\mu = \sum_{r=0}^{\infty} \frac{r\lambda^r}{(r!)^\nu G(\lambda, \nu)}$$

- Rearranging leads to

$$\sum_{r=0}^{\infty} (r - \mu) \frac{\lambda^r}{(r!)^\nu} = 0.$$

- Hence, λ is an n^{th}-order polynomial, depending on μ and ν, under this reparameterisation, known as a MPCMP distribution (or CMP_μ).

Fitting MPCMP models

- Huang [?] suggested a hybrid bisection and Newton-Raphson approach to find λ.
- This was applied in small sample Bayesian settings [?] but is not scaleable.
- Ribeiro et al. [?] used an asymptotic approximation of $G_\infty(\lambda, \nu)$ to obtain a closed form estimate for λ.
- This approximation is poor for small values of μ and ν (as are often encountered with underdispersion).
- By Descartes' rule of signs there is one positive, real solution for λ (when $r > \mu$).
- This allows us to solve the high-order polynomial, but we have to do so many times at each iteration ...

The MPCMP parameterisation in practice

- Can pre-compute look-up tables of values of λ for combinations of μ, ν and then look up $\lambda(\mu, \nu)$
- This is done in conjunction with bilinear interpolation and reduces computational times greatly
- Implemented using a step size of 0.01 for both μ and ν, with $\mu \in [0, 19]^*$ and $\nu \in [0, 10]$.
- Using multiple cores the look-up table can be created in ≈ 20 seconds
- This approach performs well in simulation studies [**?**]

The model

Denoting the yellow cards issued by referee i to team j at home ($k = 1$) or away ($k = 2$), the linear predictor is modelled by

$$\begin{aligned}\log(\mu_{ijk}) &= \beta_{\text{League}_i} + \beta_6 I(\text{Home}_{ijk} = 1) + \beta_7 I(\text{NoFans}_{ijk} = 1) \\ &+ \beta_8 I(\text{Home}_{ijk} = 1) \times I(\text{NoFans}_{ijk} = 1) + \theta_i + \gamma_j\end{aligned}$$

where

- $\boldsymbol{\beta}$ is the vector of parameters for the league effects, home advantage, no fans and the interaction of interest
- the individual referee effects are captured through $\theta_i, i = 1, \dots 171$
- the individual team effects are captured via $\gamma_j, j = 1, \dots, 129$
- dispersion at the league level is modelled via $\psi_\ell = \log(\nu_\ell), \ell = 1, \dots, 5$

The model (continued)

- The dependence between the number of home and away yellow cards is modelled via a copula
- A similar approach has been taken to model disciplinary points [?], international soccer matches [?], and English domestic soccer matches [?]
- A Frank copula was either the only, or preferred, copula in each case - the association is governed by parameter κ

The model (continued)

- Suppose Y_1 and Y_2 are a pair of random variables.
- From Sklar [**?**], the joint distribution function F may be written in the form

$$F(y_1, y_2) = C\left\{F_1(y_1), F_2(y_2)\right\}$$

where $C(\cdot)$ is a copula, of which there are many choices.
- Frank's copula is adopted here, whereby

$$C(u, v) = \frac{1}{\kappa}\log\left(1 + \frac{(e^{\kappa u} - 1)(e^{\kappa v} - 1)}{e^{\kappa} - 1}\right)$$

- We can also add zero (or other) inflation at the marginal level or the joint level

The model (continued)

- The association is governed by κ
- This association is also allowed to vary be league, in keeping with our earlier values of τ_b, via parameters $\kappa_\ell, \ell = 1, \ldots, 5$
- We adopt mean-parameterised CMP distributions for each marginal cumulative distribution function
- Happily, we can utilise our look-up table & bilinear interpolation approach once more - now just for the cumulative probabilities

Choice of priors and computational details

- Semi-informative priors are adopted for the league and referee effects
- Vaguer priors are adopted for the dispersions and copula parameters
- Sum-to-zero constraints are placed on the referees and teams *within* each league
- 10K MCMC iterations takes $\approx$ 20 minutes

Results

Results: home and pandemic effects

- The posterior mean effect for being at home is $0.90\ (0.88, 0.94)$, i.e. a 10% reduction in the rate
- The posterior mean effect for playing behind closed doors is $0.87\ (0.83, 0.92) \implies$ a 13% reduction in the rate
- The interaction has a posterior mean of $1.13\ (1.03, 1.22)$, essentially removing the home effect completely
- The posterior HDIs for the dispersions, ν_ℓ, all lie above one $\implies$ there is significant underdispersion at play within each league
- The posterior HDIs for κ_ℓ all lie above zero $\implies$ positive dependence between the home and away yellow cards

Posterior league effects

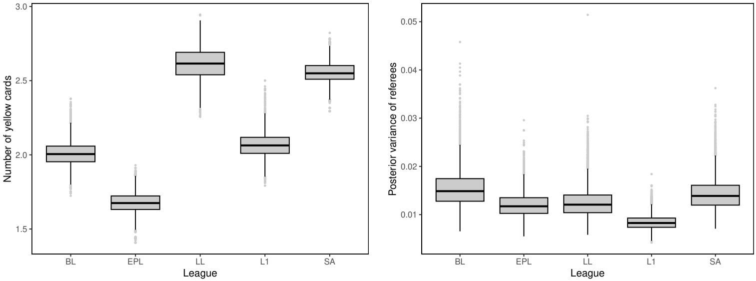


Figure 2: Boxplots of posterior distributions for the mean league effects (left) and for the variance of the referees (right)

Posterior mean referee effects

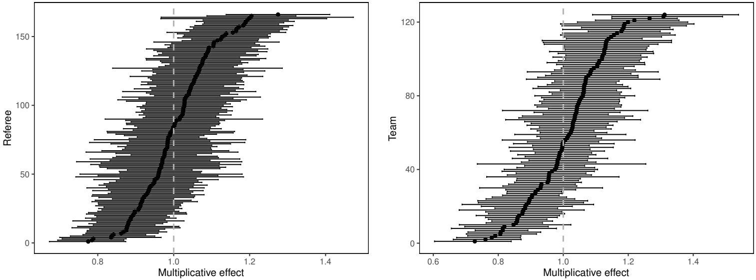


Figure 3: Posterior HDIs of θ_i for each referee (left) and team (right)

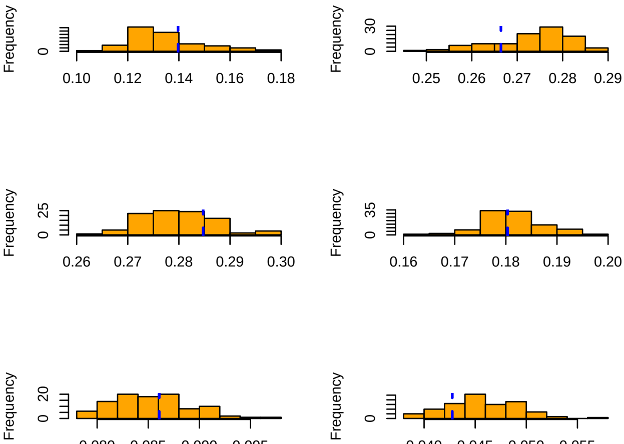
Summary table of ‘top’ and ‘tail’ referees

Rank	Referee	Country	$E(e^{\theta})$	$SD(e^{\theta})$
1	Alejandro Hernández	LaLiga	1.27	0.07
2	M Dean	EPL	1.20	0.06
3	J Brooks	EPL	1.20	0.13
4	R East	EPL	1.20	0.11
5	Gianluca Rocchi	SerieA	1.19	0.06
6	S Attwell	EPL	1.19	0.07
7	Fabio Maresca	SerieA	1.19	0.05
8	Daniele Orsato	SerieA	1.19	0.05
9	C Pawson	EPL	1.18	0.07
10	Javier Estrada	LaLiga	1.17	0.07
⋮	⋮	⋮	⋮	⋮
162	Harm Osmers	Bundesliga	0.88	0.05
163	Florent Batta	Ligue1	0.88	0.05
164	G Scott	EPL	0.87	0.05
165	José Luis Munuera	LaLiga	0.86	0.05
166	Fabrizio Pasqua	SerieA	0.84	0.04
167	Olivier Thual	Ligue1	0.84	0.06
168	Tobias Reichel	Bundesliga	0.84	0.07
169	Alberola Rojas	LaLiga	0.79	0.04
170	Eric Wattellier	Ligue1	0.78	0.04
171	Manuel Gräfe	Bundesliga	0.77	0.05

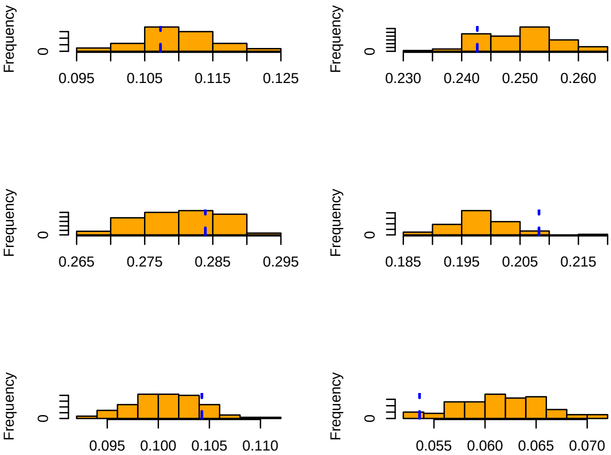
Summary table of 'top' and 'tail' teams

Rank	Team	Country	$E(e^{\gamma})$	$SD(e^{\gamma})$
1	Paderborn	Bundesliga	1.31	0.12
2	Leeds	EPL	1.31	0.09
3	Getafe	LaLiga	1.27	0.05
4	Leganes	LaLiga	1.22	0.06
5	Sheffield United	EPL	1.20	0.09
6	Chievo	SerieA	1.19	0.11
7	Venezia	SerieA	1.19	0.10
8	Metz	Ligue1	1.19	0.07
9	Fulham	EPL	1.18	0.09
10	Monaco	Ligue1	1.18	0.06
⋮	⋮	⋮	⋮	⋮
113	Juventus	SerieA	0.86	0.05
114	Inter	SerieA	0.85	0.04
115	Girona	LaLiga	0.85	0.09
116	Man City	EPL	0.82	0.07
117	Guingamp	Ligue1	0.81	0.08
118	Real Madrid	LaLiga	0.81	0.04
119	Dortmund	Bundesliga	0.81	0.06
120	Freiburg	Bundesliga	0.81	0.06
121	Atalanta	SerieA	0.79	0.05
122	Napoli	SerieA	0.78	0.05
123	Bayern Munich	Bundesliga	0.76	0.05
124	Liverpool	EPL	0.73	0.06

Model checking: home yellow cards



Model checking: away yellow cards



Other potential applications of the CMP_μ distribution

- Ice-hockey goals at the team level (e.g. Boston Bruins 22/23 season)
  1. Goals scored 3.72 (2.35)
  2. Goals conceded 2.16 (2.09)
- Baseball at the player level (at bats, runs, hits, home runs) features a host of small counts, liable to underdispersion
- NFL touchdowns (team level, QB level, . . .)
- Rushing yards per carry by NFL running backs

Bibliography I