Comparing Goal-Based and Result-Based Approaches in Modelling Football Outcomes

Leonardo Egidi[1] · **Nicola Torelli**[1]

Abstract
Two main approaches are considered when building statistical models for football outcomes: (1) the goal-based approach, where the number of goals scored by two competing teams is modelled, and (2) the result-based approach, where a three-category outcome (win–draw–loss) is modelled. The debate about which approach is preferable is still ongoing, although the general agreement is that any direct comparison between the forecasting abilities of the two approaches should be based on the quality of the forecasts. Alternatively, a greater emphasis can be given to diagnostic measures in order to judge the quality of model specifications, as is more customary in statistical modelling. In this paper, we develop a broad comparison of four possible Bayesian models, focusing on model checking and calibration and then using Markov chain Monte Carlo replications to explore the predictive performance over future matches. Although inconclusive, we believe our set of comparison tools may be beneficial for future scholars in differentiating the two approaches.

Keywords Football results · Forecasting · Goodness of fit · Predictive performance · MCMC

1 Introduction

The outcome of a football match may be modelled according to two distinct perspectives.

The *goal-based* approach implies modelling via suitable count distribution, the number of the goals scored and conceded by the teams in each match. In the literature, we mainly recognize three types of goal-based Poisson models: double Poisson (Maher 1982; Baio and Blangiardo 2010; Groll and Abedieh 2013; Egidi et al. 2018), bivariate Poisson (Dixon and Coles 1997; Karlis and Ntzoufras 2003), and Poisson difference/Skellam (Karlis and Ntzoufras 2009). Once a model has been estimated, the derivation of the so-called

✉ Leonardo Egidi
legidi@units.it

Nicola Torelli
nicola.torelli@deams.units.it

[1] Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche Bruno de Finetti, Università degli Studi di Trieste, Via Tigor 22, Trieste, Italy

three-way process (home win, draw, away win) can be obtained by aggregating the estimated probabilities.

The *result-based* approach consists of modelling directly the three-way process, by use of ordered probit (Koning 2000) or logit (Carpita et al. 2015, 2019) regression models. This second framework is nested within the first one: the result of a football match is established from the goals scored and conceded, while knowledge of the simple three-way result says nothing about the number of the goals scored by the two teams.

Result-based models have a simpler structure, require fewer parameters, and avoid any assumption about the goals' interdependence; however, they could dramatically underestimate/overestimate the actual strength of a team, since matches concluded on scores of 1–0 and 5–0 are of equal value. After extensive debate, Goddard (2005) asserted that any direct comparison between the forecasting abilities of the two types of models must be based on forecasts of match results. However, few studies have compared these two perspectives in terms of goodness of fit and calibration in addition to predictive performance on a test set.

From a predictive viewpoint, we should always choose the model that yields the best predictions, according to some well-chosen metrics. Although it sounds appealing, this cannot be the sole preference: can we trust a model that gives poor predictions of the *group stages* of the World Cup but, surprisingly, gives accurate predictions of the *knockout* stages? Perhaps not. Rather, we should select a model, or a class of models, after extensive analysis of its performance, both in terms of goodness of fit and forecasting abilities.

The number of goals in a match represents an example of paired count data which are, in fact, also used in social sciences to build social rankings and to measure relative preferences assigned to certain objects or items, with the aim of ordering objects on a preference scale according to an attribute. The attributes are usually based on subjective evaluations of properties of the objects (e.g., tastiness of food, beauty of owers, perceived risk of portfolios) or on "objective" outcomes under some predefined rules (e.g., strength of football teams, quality of scientific journals, etc.). The `R` package `prefmod` (Hatzinger and Dittrich 2012) presents some preference models, by extending the Bradley–Terry (Bradley and Terry 1952) and Thurstone–Mosteller (Thurstone 1927; Mosteller 2006) models, to predict the outcome of pairwise comparisons by using continuous distributions for the pairwise difference. Responses to Likert-type items, often called ratings, are another form of data collection to obtain preference orderings (Dittrich et al. 2007).

The issue of modelling paired count data is relevant for many other fields connected to social sciences. In our opinion a broad comparison of these discrete models, based on predictive accuracy and diagnostic measures, may be beneficial not only for analysing football data but also for a broader audience consisting of epidemiologists, biologists, and psychologists. Examples of models similar to those here considered are in Karlis and Ntzoufras (2006), Böhning et al. (1999) and Davison (1992) and, with specific application to predicting football results, in Ley et al. (2019). Whenever paired comparisons are required, the finer the set of tools to discern among the distinct models, the better is the choice for the analyst.

In this paper, we develop a broad comparison of four possible Bayesian models using the data from the FIFA World Cup 2018 hosted in Russia; we focus on the model posterior checking and calibration (Gelman et al. 2013), and then use Markov chain Monte Carlo (MCMC) (Robert and Casella 2013) replications to explore predictive performance for future matches. Although inconclusive, we believe our comparison review may be beneficial for future scholars to differentiate between the goal-based and result-based models. The answer, as emerges in this paper, may not be unique, perhaps even controversial, and the choice of the final model is left to the analyst's expertise. This paper does not

emphasize on selecting the best model, but on the 'bag of tools' required to select a good model, consisting of posterior predictive checks, predictive information criteria, probability scores, and, more generally, predictive accuracy diagnostics.

The rest of the paper proceeds as follows. In Sect. 2, we propose four distinct Bayesian models, two of them are goal-based, and two are result-based. Posterior predictive checking is introduced in Sect. 3, whereas a variety of predictive accuracy measures is presented in Sect. 4 along with graphical visualization. Section 5 concludes.

2 Models

2.1 Multinomial Models

Let $z_n \in \{1, \mathrm{X}, 2\}$ denote the observed categorical result for the n-th match, $n = 1, \dots, N$, where $\{1, \mathrm{X}, 2\}$ hereafter denotes the three-way process for the home team win, the draw, and the away team win, respectively. Within the World Cup framework detailed in Sect. 3, 'Home' and 'Away' do not have particular meanings attached to them, they simply distinguish the two competing teams and maintain consistency with the statistical football literature. A multinomial model for the categorical random variable Z_n is assumed:

$$\begin{aligned} Z_n|\boldsymbol{\pi}_n &\sim \mathsf{Multinom}(1, \boldsymbol{\pi}_n),\ n = 1, \dots, N \\ \pi_{n1} &= \frac{\exp\{\eta_{n1}\}}{1 + \sum_{j=1}^{2} \exp\{\eta_{nj}\}} \\ \pi_{n2} &= \frac{\exp\{\eta_{n2}\}}{1 + \sum_{j=1}^{2} \exp\{\eta_{nj}\}} \\ \pi_{nX} &= \frac{1}{1 + \sum_{j=1}^{2} \exp\{\eta_{nj}\}}, \end{aligned} \tag{1}$$

where $\boldsymbol{\pi}_n = (\pi_{n1}, \pi_{nX}, \pi_{n2})$ is the vector of match probabilities, and $\boldsymbol{\eta}_n = (\eta_{n1}, \eta_{n2})$ is the vector of linear predictors for the home and the away team, respectively, defined as:

$$\begin{aligned} \eta_{n1} &= \theta + \mathrm{att}_{h_n} + \mathrm{def}_{a_n} + v_1 \frac{\gamma}{2} f(w_n) \\ \eta_{n2} &= \theta + \mathrm{att}_{a_n} + \mathrm{def}_{h_n} - \frac{\gamma}{2} f(w_n), \end{aligned} \tag{2}$$

where θ is the common baseline parameter; the parameters att_T and def_T represent the attack and the defence abilities, respectively, for each team T, $T = 1, \dots, N_T$; the nested indexes $h_n, a_n = 1, \dots, N_T$ denote the home and the away team playing in the n-th game, respectively; the predictor $w_n = (\mathrm{rank}_{h_n} - \mathrm{rank}_{a_n})$ is the difference of the FIFA World Rankings[1]—expressed in FIFA ranking points divided by 10^3—between the home and the away team in the n-th game, evaluated here through a function $f(\cdot)$, and multiplied by a parameter $\gamma/2$. This last term tries to correct for the ranking difference occurring between two competing teams. To allow for this, possible functions are $f_1(w_n) = |w_n|^{-1}$ along with $v_1 = -1$ or the identity function $f_2(w_n) = w_n$ along with $v_1 = 1$. f_1 increases (decreases) as

[1] https://www.fifa.com/fifa-world-ranking/.

the ranking difference w_n decreases (increases), and the factor $\frac{\gamma}{2} f_1(w_n)$ is then subtracted from both the equations in Eq. (2) to increase the draw probability as the teams are closer. This first function tries to correct for the well-known phenomenon of *draw inflation* [see, for instance, Karlis and Ntzoufras (2003)], favouring the draw occurrence when teams are close to each other in the FIFA rankings. Instead, $f_2(w_n)$ increases as the ranking difference w_n increases, and the factor $\frac{\gamma}{2} f_2(w_n)$ is added to the "home" team and subtracted from the "away" team. This second function aims at giving more weight to the marginal winning probabilities for the first or second team by adding or subtracting a positive factor, respectively. As explained in Sect. 2.3, γ will be assigned a weakly informative Gaussian distribution to account for any possible real value. In the rest of the paper, the multinomial models will be referred as *Multinomial* and *Multinomial 2* depending on the function f.

2.2 Poisson Models

Let (x_n, y_n) denote the observed number of goals scored by the home and the away team in the n-th game, respectively. A simple double Poisson model implies for each match $n = 1, \dots, N$ the following specification:

$$\begin{aligned} X_n | \lambda_{1n} &\sim \text{Poisson}(\lambda_{1n}) \\ Y_n | \lambda_{2n} &\sim \text{Poisson}(\lambda_{2n}) \\ \log(\lambda_{1n}) &= \theta + \text{att}_{h_n} + \text{def}_{a_n} + \frac{\gamma}{2} w_n \\ \log(\lambda_{2n}) &= \theta + \text{att}_{a_n} + \text{def}_{h_n} - \frac{\gamma}{2} w_n, \end{aligned} \tag{3}$$

where $\lambda_{1n}, \lambda_{2n}$ represent the scoring rates for the home and away team, respectively; all the other parameters have the same interpretation as in Sect. 2.1. Including positive parametric goals' dependence is made possible by using a bivariate Poisson distribution (Karlis and Ntzoufras 2003). In such a framework, the numbers of goals are jointly modelled:

$$\begin{aligned} (X_n, Y_n | \lambda_{1n}, \lambda_{2n}, \lambda_{3n}) &\sim \text{BivPoisson}(\lambda_{1n}, \lambda_{2n}, \lambda_{3n}) \\ \log(\lambda_{1n}) &= \theta + \text{att}_{h_n} + \text{def}_{a_n} + \frac{\gamma}{2} w_n \\ \log(\lambda_{2n}) &= \theta + \text{att}_{a_n} + \text{def}_{h_n} - \frac{\gamma}{2} w_n \\ \log(\lambda_{3n}) &= \beta_0, \end{aligned} \tag{4}$$

where $\text{E}(X_n) = \lambda_{1n} + \lambda_{3n}, \text{E}(Y_n) = \lambda_{2n} + \lambda_{3n}$, and $\text{cov}(X_n, Y_n) = \lambda_{3n}$; λ_{3n} acts as a measure of dependence between the goals scored by the two competing teams.

2.3 Priors and Constraints

As a matter of parameters' interpretation, once the models have been estimated the larger is the team-attack parameter, the greater is the attacking quality for that team; conversely, the lower is the team-defence parameter, the better is the defence power for that team.

For each team $T = 1, \dots, N_T$, attack and defence parameters are assigned weakly informative priors (Gelman et al. 2008):

$$
\begin{aligned}
\text{att}_T &\sim \mathcal{N}(\mu_{\text{att}}, \sigma_{\text{att}}) \\
\text{def}_T &\sim \mathcal{N}(\mu_{\text{def}}, \sigma_{\text{def}}) \\
\sigma_{\text{att}}, \sigma_{\text{def}} &\sim \text{Cauchy}^{+}(0, \tau) \\
\mu_{\text{att}}, \mu_{\text{def}} &\sim \mathcal{N}(0, 10),
\end{aligned}
\tag{5}
$$

where $\mathcal{N}(\mu, \sigma)$ is the notation adopted for a Gaussian distribution with mean μ and standard deviation σ, whereas $\text{Cauchy}^{+}(0, \tau)$ is the half-Cauchy distribution with location zero and scale τ. As explained by Gelman (2006), the half-Cauchy distribution for the scale parameter in a hierarchical model is likely to not affect the posterior estimates, is flexible, and has a better behaviour near 0. The value of 10 for the standard deviation of $\mu_{\text{att}}, \mu_{\text{def}}$ has been chosen to account for possible broad values for the global attack/defence mean parameters. Different values have been tested and nothing changed in terms of posterior results. To achieve identifiability, these parameters are imposed a "sum-to-zero" constraint (Baio and Blangiardo 2010):

$$
\sum_{T=1}^{N_T} \text{att}_T = 0, \quad \sum_{T=1}^{N_T} \text{def}_T = 0.
$$

The parameter γ associated with the ranking difference is assigned a weakly informative Gaussian prior, whereas the parameter β_0 modelling the goals' correlation in Eq. (4), defined on $\mathbb{R}^{+}$ (only positive correlation), is assigned a half-Gaussian prior allowing for extreme values:

$$
\gamma \sim \mathcal{N}(0, 2) \tag{6}
$$

$$
\beta_0 \sim \mathcal{N}^{+}(0, 5). \tag{7}
$$

The standard deviations set to 2 and 5 for γ and β_0, respectively, have been chosen in accordance with weakly informative criteria and only upon some sensitivity tests.

3 Posterior Estimates and Model Checking

The models introduced in Sect. 2 were fitted on the dataset containing the results of all the 64 tournament's matches (48 of the group stages, and 16 of the knockout stage) for the FIFA World Cup 2018. The value of the FIFA ranking difference w incuded in the models was considered on June 7th, only a few days before the tournament commenced. Model fit has been obtained by use of the `R` (R Core Team 2018) package `rstan` (Stan Development Team 2018) relying on Hamiltonian Monte Carlo sampling, whereas chains' convergence was monitored via the Gelman–Rubin statistics (Gelman and Rubin 1992), as suggested by Gelman et al. (2013). The main advantage of using Stan over the traditional MCMC automatic tools relying on Gibbs sampling, such as JAGS (Plummer 2003) and WinBUGS (Lunn et al. 2000), is its efficiency in exploring the joint posterior distribution, especially in cases where the posterior has some irregularities. Moreover, the main computational appeal is that Stan is very well connected with other tools, such as the packages `bayesplot` (Gabry and Mahr 2019) and `loo` (Vehtari et al. 2019), which may be used to depict posterior plots and to retrieve predictive information criteria, respectively. We fitted the models at each World Cup stage, using: the first group stage matches as the training set

to predict the matches of the second group stage, then the first and the second group stage matches to predict the matches of the third group stage, and so on until the finals.

Once the models have been estimated, the next step is to provide some goodness of fit measures. Posterior predictive checking (Gelman et al. 2013) is the main tool to check whether a Bayesian model is able to produce replications as closely as possible to the observed data. The idea is to generate hypothetical replications y^{rep} from the posterior predictive distribution $p(y^{rep}|y) = \int \pi(\theta|y)p(y^{rep}|\theta)d\theta$, where $\pi(\theta|y)$ is the posterior distribution and $p(y^{rep}|\theta)$ is the likelihood function for hypothetical values. Rarely is this distribution analytically tractable, for such a reason we need a two-steps simulation at each MCMC iteration to (a) generate θ from $\pi(\theta|y)$; (b) generate y^{rep} from $p(y^{rep}|\theta)$. Figure 1 displays the true data distribution (dark pink) plotted against the replicated MCMC distributions (light pink) for the four models considered in Sect. 2. For the multinomial models [panel (a) and (b)], the dependent variable is the categorical random variable taking values in $\{1, \mathrm{X}, 2\}$, whereas in the Poisson models [panel (c) and (d)], the variable considered in the plot is the goal difference $X-Y$. The latter models seem to suggest a better agreement of the replications to the observed values, displaying a non-negligible peak of probability mass around zero, occurring for the draws; the replicated distributions for the multinomial class of model appear slightly noisier.

4 Predictive Performance

4.1 Posterior Matches Probabilities

Bayesian models easily retrieve posterior probabilities for future matches by using MCMC simulations from the posterior predictive distribution for future values. Denoting with $\tilde{y}$ a future observable value, the posterior predictive distribution is $p(\tilde{y}|y) = \int \pi(\theta|y)p(\tilde{y}|\theta)d\theta$. Table 1 shows posterior probabilities under each model for the three-way process $\{1, \mathrm{X}, 2\}$ of the 16 matches of the knockout World Cup stage (round of 16, quarter of finals, semifinals, finals). For each model, red cells denote the highest probabilities when not corresponding to the observed results (considering that there will be a result in the regular 90 min), whereas light green cells denote the highest probabilities when corresponding to the observed results. The best overall probabilities for the observed results across all the four models are marked in dark green. The qualified teams are marked in bold characters. Distinct knockout stages (round of 16, quarter of finals, semifinals, and finals) are separated by a solid horizontal line. The table shows that there is no model that clearly dominates the remaining ones: rather, the models tend to behave similarly for many matches, such as Sweden–Switzerland, Colombia–England, France–Belgium, and the two finals France–Croatia and Belgium–England.

However, these three-way probabilities may suffer from some inefficiencies when two teams have high scoring abilities: in such situations, the exact result in terms of the goals scored by the two teams may be preferred to the three-way process $\{1, \mathrm{X}, 2\}$. Poisson models may be used to depict a grid plot of the posterior probabilities for the exact results as in Fig. 2. For each model, darker regions are associated with higher posterior probabilities and red squares correspond to the observed results. It is worth noting that black regions do not correspond to the same probability scale values across the two models, they just denote the most likely results. For the first match, France versus Croatia (top row), the observed final result (4–2) was unlikely under both the double and the bivariate Poisson

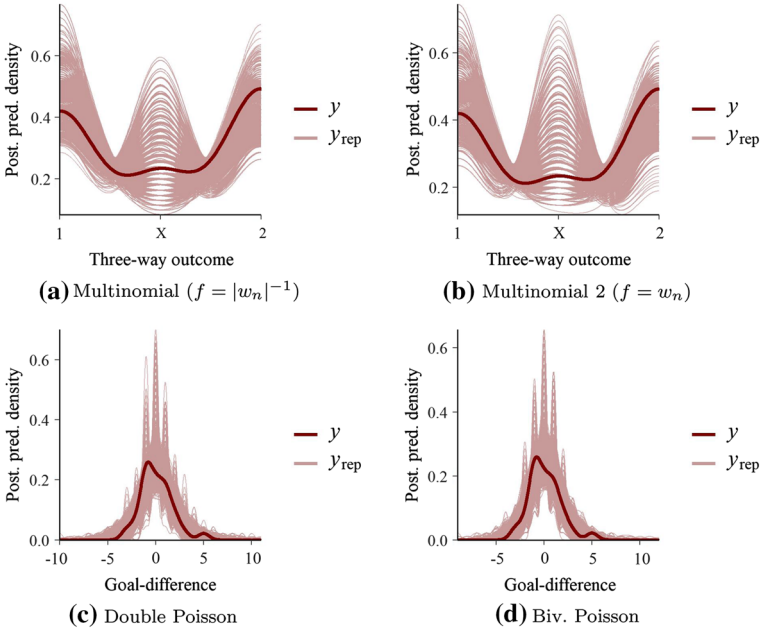


Fig. 1 Posterior predictive checks for the true data distribution (dark pink) plotted against the replicated MCMC distributions (light pink), for the four considered models: multinomial with $f(w_n) = |w_n|^{-1}$, multinomial with $f(w_n) = w_n$, double Poisson and bivariate Poisson. In the top-row plots, y is the categorical three-way process, the x-axis denotes the three possible results $\{1, X, 2\}$ only, and the y-axis depicts the values for the posterior predictive distribution. In the bottom-row plots, y is the goal-difference between the home and away team, in the x-axis, there are the discrete values for the goal-difference, whereas the y-axis depicts the values for the posterior predictive distribution. The plots have been obtained by running 2000 Hamiltonian Monte Carlo (HMC) iterations with `rstan` package, and then using the `bayesplot` package, which always provides a continuous approximation for discrete distributions. (Color figure online)

model; however, taking the lower (upper) triangular matrix of results and summing all the cells, we may obtain the posterior probability of a France (Croatia) win. For the second final, Belgium versus England, the observed result (2–0) had a non-negligible posterior probability, approximately 0.06, under both the Poisson models. However, some further considerations deserve a quick attention. The plots above, even when the observed result is unlikely under the posterior predictive distribution, have the advantage to depict the whole posterior uncertainty and, then, to provide a glimpse about the win–draw–loss process starting from the single exact results. The result 4–2 between France and Croatia having a posterior probability to happen approximately equal to zero is not that relevant, since this issue is well-known among football modellers: rather, we could have many concerns in a model assigned to the 4–2 result a very high probability. In order to be fully transparent, we feel these plots should be displayed for each of the out-of-sample matches.

Table 1 Posterior match probabilities for the 16 matches of the knockout stage: round of 16, quarter of finals, semifinals, finals

		Multin			Multin 2			double Pois.			biv. Pois			
Team1	Team2	1	X	2	1	X	2	1	X	2	1	X	2	Obs.
France	Argentina	0.61	0.05	0.34	0.42	0.36	0.22	0.41	0.30	0.29	0.40	0.29	0.31	4-3 (1)
Uruguay	Portugal	0.43	0.27	0.30	0.28	0.38	0.34	0.30	0.28	0.40	0.31	0.29	0.40	2-1 (1)
Spain	**Russia**	0.38	0.33	0.29	0.54	0.31	0.15	0.49	0.23	0.28	0.47	0.25	0.28	1-1 (X)
Croatia	Denmark	0.50	0.18	0.32	0.36	0.38	0.26	0.39	0.31	0.30	0.38	0.30	0.32	1-1 (X)
Brazil	Mexico	0.46	0.24	0.30	0.54	0.28	0.18	0.53	0.27	0.20	0.51	0.28	0.21	2-0 (1)
Belgium	Japan	0.50	0.27	0.23	0.70	0.23	0.07	0.63	0.23	0.14	0.62	0.22	0.16	3-2 (1)
Sweden	Switz.	0.32	0.28	0.40	0.23	0.36	0.41	0.32	0.28	0.40	0.32	0.28	0.40	1-0 (1)
Colombia	**England**	0.45	0.07	0.48	0.34	0.24	0.42	0.32	0.27	0.41	0.32	0.28	0.40	1-1 (X)
Uruguay	**France**	0.41	0.28	0.31	0.32	0.32	0.36	0.32	0.31	0.38	0.32	0.31	0.37	0-2 (2)
Brazil	**Belgium**	0.33	0.26	0.41	0.34	0.31	0.35	0.39	0.29	0.32	0.39	0.30	0.31	1-2 (2)
Sweden	**England**	0.42	0.23	0.35	0.35	0.30	0.35	0.32	0.29	0.39	0.31	0.30	0.39	0-2 (2)
Russia	**Croatia**	0.24	0.33	0.43	0.14	0.30	0.56	0.25	0.28	0.47	0.24	0.29	0.46	2-2 (X)
France	Belgium	0.36	0.18	0.44	0.30	0.25	0.45	0.31	0.26	0.43	0.30	0.29	0.41	1-0 (1)
England	**Croatia**	0.31	0.21	0.48	0.32	0.29	0.39	0.40	0.27	0.33	0.39	0.29	0.32	1-1 (X)
France	Croatia	0.38	0.32	0.30	0.51	0.23	0.26	0.41	0.28	0.31	0.40	0.28	0.32	4-2 (1)
Belgium	England	0.56	0.20	0.24	0.62	0.20	0.18	0.44	0.26	0.30	0.42	0.28	0.30	2-0 (1)

For each model, red cells denote the highest probabilities when not corresponding to the observed results, whereas light green cells denote the highest probabilities when corresponding to the observed results. The best overall probabilities for the observed results across all the four models are marked in dark green. The qualified teams are marked in bold characters. Distinct knockout stages (round of 16, quarter of finals, semifinals and finals) are separated by a solid horizontal line

1, X, and 2 denote the home team win, the draw, and the away team win, respectively

[1]Observed results are considered within the 90 regular min

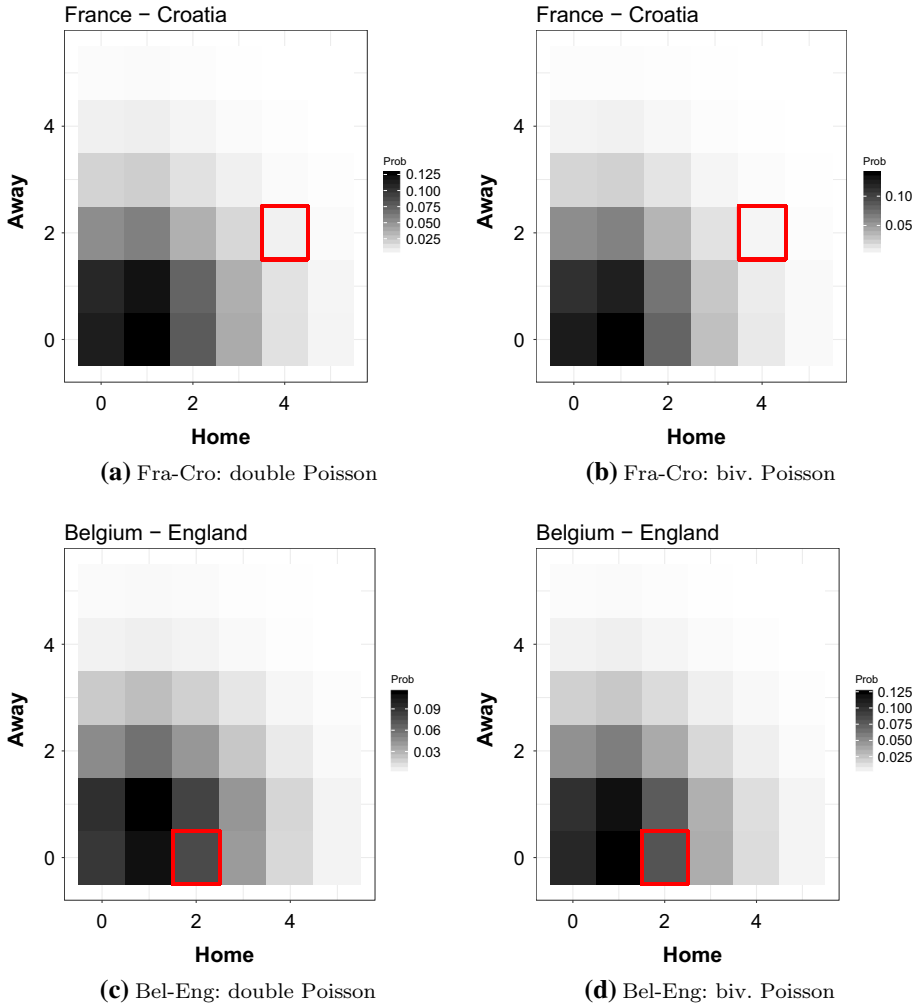
[2]Model fitting: `rstan` package, 2000 iterations

4.2 Pseudo-R^2

One way to assess predictive performance over any number of matches is the pseudo-R^2, defined as the geometric mean of the probabilities assigned to the actual result of each match played during the forecast period (Dobson et al. 2001):

$$\text{pseudo-R}^2 = (p_1 p_2 \dots p_M)^{1/M}, \tag{8}$$

where M is the number of future matches for which forecasts were generated, and p_m is the estimated probability to observe the actual result of the match m. The left plot in Fig. 3 shows the pseudo-R^2 computed for each World Cup stage under the four considered models: multinomial models perform quite poorly in the first phases, but their performance starts to improve after the 3rd group stage. This may be because usually, at later stages, each match involves teams with closer scoring abilities (or FIFA rankings) than in former stages. Therefore, it is safer to predict only the overall outcome. Conversely, Poisson models perform better in the first group stage than in the knockout stage. All four models perform well for the final (France vs. Croatia), giving greater probability for a France win.



(a) Fra-Cro: double Poisson **(b)** Fra-Cro: biv. Poisson

(c) Bel-Eng: double Poisson **(d)** Bel-Eng: biv. Poisson

Fig. 2 Posterior predictive distribution of the possible results for the two finals, France versus Croatia and Belgium versus England, according to the double Poisson and the bivariate Poisson models. The four plots report the posterior uncertainty for the spectrum of all the possible results. Darker regions are associated with higher posterior probabilities and red square corresponds to the observed result

4.3 Brier Score

The Brier score (Brier 1950), used by (Spiegelhalter and Ng 2009), is a type of mean-squared error of the forecasts, ranging from zero to two:

$$\text{Brier score} = \frac{1}{M}\sum_{m=1}^{M}\sum_{i\in\{1,\text{X},2\}}(p_{im} - o_{im})^2, \tag{9}$$

where p_{im} is the forecast probability of the outcome i, $i \in \{1, \text{X}, 2\}$, in the m-th match, and o_{im} is a dummy coding for the actual outcome in the m-th match, equals one if event

Table 2 Average pseudo-R^2 and Brier score across the knockout stage (round of 16, quarter, semifinals, and finals) of the FIFA World Cup 2018: the best probabilities in terms of pseudo-R^2 are offered by the bookies, followed by the multinomial model with $f(w_n) = w_n$ (Multinomial 2)

	Multin.	Multin. 2	D. Pois.	Biv. Pois	Bookies
Pseudo-R^2	0.346	0.367	0.351	0.353	0.378
Brier score	0.647	0.642	0.610	0.612	0.656

Best probabilities in terms of Brier score are offered by the double Poisson model

i happened, zero otherwise. The lower the Brier score, the better is the model's predictive accuracy. Table 2 displays the pseudo-R^2 and the Brier score computed for the knockout stage of the World Cup: Multinomial 2 yields the highest pseudo-R^2 among the other models, but the value is slightly lower than that obtained by using the bookmakers probabilities, taken from the website https://www.oddsportal.com. In this application, we transformed the betting odds into probabilities by adopting the so-called normalization procedure, by dividing each odds for the sum of the odds. Although not unique, this is the most common method to derive exact probabilities from the bookmakers odds (see Egidi et al. (2018) for other details). Quite surprisingly, the Brier score for the double Poisson model is the lowest, and all the models yield values lower than those by bookies. For double and bivariate Poisson, we computed the win–draw–loss probabilities by aggregating over all the possible results coming from the MCMC sampling. For instance, posterior probabilities for the exact results 1–0, 2–0, 2–1, 3–1, etc., all contribute to the overall home-win probability, whereas posterior probabilities for the exact results 0–0, 1–1, 2–2, etc., all contribute to the overall draw probability, and so on for the the away-win probabilities as well. In such a sense, the finer partition implied by Poisson models—win–draw–loss results are nested within the goals realized and conceded in each match—is acknowledged by better predictive results in terms of the Brier score.

4.4 Leave-One-Out Cross Validation

Another way to assess and compute predictive accuracy is leave-one-out cross-validation (LOO) (Vehtari et al. 2017), a method for estimating pointwise out-of-sample prediction accuracy from a fitted Bayesian model using the log-likelihood evaluated at the posterior simulations of the parameter values. The Bayesian LOO estimate of out-of-sample predictive fit is:

$$\text{LOOIC} = -2 \sum_{n=1}^{N} \log p(y_n | y_{-n}), \tag{10}$$

where $p(y_n|y_{-n})$ is the leave-one-out predictive density given the data without the n-th data point. Analogously as with the other predictive information criteria such as AIC, DIC and WAIC, the lower is the LOOIC, the better is the model predictive accuracy. The right plot of Fig. 3 displays the LOOIC for the four considered models along the distinct World Cup stages: as may be seen, multinomial models yield lower LOOIC values in each tournament stage, possibly due to have fewer and fewer parameters than the Poisson models. For each

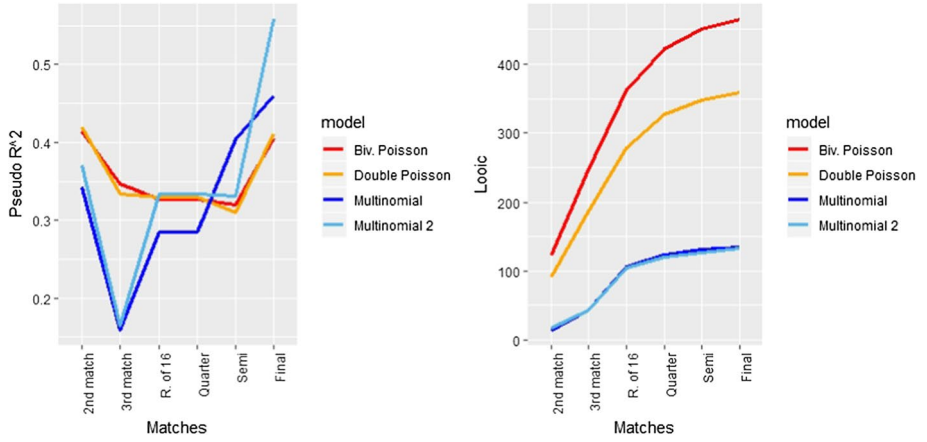


Fig. 3 Pseudo-R^2 and LOOIC for all the considered stages of the FIFA World Cup

model, LOOIC increases as the World Cup progresses, and this happens because at each stage the number of the games increases and the sum in Eq. (10) increases with N.

5 Discussion

We used the FIFA World Cup 2018 results to extensively compare and discern between result-based (multinomial) and goal-based (Poisson-based) models in predicting football outcomes. As we suspected, we do not have a unique answer; this is mainly because a statistical model cannot be validated only based on its predictive performance on a test set, but should be extensively verified and monitored from a broader viewpoint, going from posterior predictive checking to cross-validation tools.

Rather than adding new models to the existing literature, the main aim of this paper is to warn the reader and the football fan about trusting a particular model only for its eventual appealing performance and to push him to continually verify it in a larger sense. As a matter of practice, we understand that statistical modellers and data scientists often need a unique efficient procedure to find inferential conclusions and to attempt out-of-sample predictions. Being faced with the constraint of choosing a model, we would end up selecting the multinomial models for tournaments such as the World Cup, where, usually, predicting the final three-way outcome is easier than trying to predict the exact number of goals. Moreover, model complexity is lower than that of Poisson models, as the leave-one-out cross validation suggests.

Despite a great statistical interest in a knockout tournament structure such as the World Cup, future research should focus on applying the proposed bag of comparison tools to seasonal leagues such as the English Premier League or the Italian Serie A. With a larger number of matches and, consequently, larger amount of information, we expect much more robustness from those analyses and less controversial findings. In such a way, we could claim which class of models is actually better designed to provide suitable goodness of fit and good predictive accuracy measures for long-run seasonal leagues.

References

Baio, G., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, *37*(2), 253–264.

Böhning, D., Dietz, E., Schlattmann, P., Mendonca, L., & Kirchner, U. (1999). The zero-inflated poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*(2), 195–209.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324–345.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthey Weather Review*, *78*(1), 1–3.

Carpita, M., Ciavolino, E., & Pasca, P. (2019). Exploring and modelling team performances of the kaggle european soccer database. *Statistical Modelling*, *19*(1), 74–101.

Carpita, M., Sandri, M., Simonetto, A., & Zuccolotto, P. (2015). Discovering the drivers of football match outcomes with data mining. *Quality Technology and Quantitative Management*, *12*(4), 561–577.

Davison, A. (1992). Treatment effect heterogeneity in paired data. *Biometrika*, *79*(3), 463–474.

Dittrich, R., Francis, B., Hatzinger, R., & Katzenbeisser, W. (2007). A paired comparison approach for the analysis of sets of likert-scale responses. *Statistical Modelling*, *7*(1), 3–28.

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *46*(2), 265–280.

Dobson, S., Goddard, J. A., & Dobson, S. (2001). *The economics of football*. Cambridge: University Press Cambridge.

Egidi, L., Pauli, F., & Torelli, N. (2018). Combining historical data and bookmakers' odds in modelling football scores. *Statistical Modelling*, *18*(5–6), 436–459.

Gabry, J., & Mahr, T. (2019). *bayesplot: Plotting for Bayesian models*. R package version 1.7.0.

Gelman, A., et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, *1*(3), 515–534.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). London: Chapman & Hall.

Gelman, A., Jakulin, A., Pittau, M. G., Su, Y.-S., et al. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, *21*(2), 331–340.

Groll, A., & Abedieh, J. (2013). Spain retains its title and sets a new record-generalized linear mixed models on European football championships. *Journal of Quantitative Analysis in Sports*, *9*(1), 51–66.

Hatzinger, R., Dittrich, R., et al. (2012). Prefmod: An R package for modeling preferences based on paired comparisons, rankings, or ratings. *Journal of Statistical Software*, *48*(10), 1–31.

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*(3), 381–393.

Karlis, D., & Ntzoufras, I. (2006). Bayesian analysis of the differences of count data. *Statistics in Medicine*, *25*(11), 1885–1905.

Karlis, D., & Ntzoufras, I. (2009). Bayesian modelling of football outcomes: Using the Skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, *20*(2), 133–145.

Koning, R. H. (2000). Balance in competition in dutch soccer. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *49*(3), 419–431.

Ley, C., Wiele, T. V. d, & Eetvelde, H. V. (2019). Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches. *Statistical Modelling*, *19*(1), 55–77.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). Winbugs—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, *36*(3), 109–118.

Mosteller, F. (2006). Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. In *Selected papers of Frederick Mosteller* (pp. 157–162). Berlin: Springer.

Plummer, M. et al. (2003). Jags: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, p. 10). Vienna.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Robert, C., & Casella, G. (2013). *Monte Carlo statistical methods*. Berlin: Springer.
Spiegelhalter, D., & Ng, Y.-L. (2009). One match to go!. *Significance*, *6*(4), 151–153.
Stan Development Team. (2018). RStan: The R interface to Stan. R package version 2.18.2.
Thurstone, L. L. (1927). Psychophysical analysis. *The American Journal of Psychology*, *38*(3), 368–389.
Vehtari, A., Gabry, J., Yao, Y., & Gelman, A. (2019). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. R package version 2.1.0.
Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.