Are the shots predictive for the football results?

I tiri sono predittivi per modellare il numero di reti nel calcio?

Leonardo Egidi, Francesco Pauli, Nicola Torelli

Abstract In modelling football outcomes, scores' data are regularly used for the estimation of the attack and the defence strength of each team. However, these teams' abilities are quite complex and are correlated with many quantities inherent to the game. Additional available information, relevant for their estimation, are shots, both made and conceded. For such a reason, we propose a hierarchical model that incorporates this information in three stages for each game and each team: number of scores, number of shots on target and number of total shots. We fit the model on English Premier League data and obtained predictions for future matches.

Abstract *Nel modellare i risultati calcistici, i dati sui goal sono solitamente utilizzati per stimare le abilità di attacco e difesa di ogni squadra. Tuttavia, queste abilità hanno una natura complessa e sono correlate con molte quantità inerenti al gioco. Un'ulteriore informazione disponibile, rilevante per stimare questi parametri, è data dai tiri, sia quelli realizzati che quelli concessi. A tale scopo proponiamo un modello gerarchico che incorpora questa informazione in tre stadi per ogni partita e ogni squadra: numero di goal, numero di tiri nello specchio e numero di tiri totali. Abbiamo applicato il modello sui dati della Premier League inglese e ottenuto previsioni per partite future.*

Key words: modelling football outcomes, hierarchical model, shot, prediction

1 Introduction

Modelling the outcome of a football match is the subject of much debate, and various models based on different assumptions have been proposed. The basic assump-

Leonardo Egidi, Francesco Pauli, Nicola Torelli
Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche, 'Bruno de Finetti', Università degli Studi di Trieste, Via Tigor 22, 34124 Trieste, Italy, e-mail: legidi@units.it, francesco.pauli@deams.units.it, nicola.torelli@deams.units.it

tion is that the number of goals scored by the two teams follow two Poisson distributions (Maher, 1982; Baio and Blangiardo, 2010)—possibly with rates parameters accounting for different sources of information, such as bookmakers odds (Egidi et al., 2018)—but many researchers investigated the correlation between them by proposing a more complicated bivariate Poisson distribution (Karlis and Ntzoufras, 2003).

Another typical assumption is the inclusion in the models of some teams' effects to describe the attack and the defence strengths of the competing teams. For this aim, the advent of some dynamic structures (Owen, 2011; Koopman and Lit, 2015) allowed these parameters to vary over the time, in order to specify an intuitive temporal evolution of these teams' skills along the match days and the seasons. The historical match results, possibly along with a set of further covariates, are usually the only data used for the estimation of these abilities. However, the scoring and the defence abilities are strongly correlated with the shots and the shots conceded respectively. For such a reason, including this information into a model designed for predicting the scores could provide relevant benefits both in terms of the realistic description of the game and the prediction of future matches outcomes. As an example of the relevance of the shots on target and the total shots on the statistical prediction of match results for the Italian football league Serie A, see Carpita et al. (2015).

In this paper, we propose a Bayesian hierarchical model consisting of a data-hierarchy in three stages for each game and each team, where the nested quantities are: number of scores, number of shots on target and number of total shots. The number of scores and the number of shots on target follow two binomial distributions respectively, with probability and population treated as further parameters. Intuitively, the total shots also consist of all those attempts—e.g., long distance shots, last minute shots—which may represent a noisy proxy for the attack skills, and for such a reason they represent the last level of the assumed hierarchy.

As far as we know from reviewing the current literature, this proposal represents a novelty also in terms of parameters' interpretation: binomial probabilities associated to the first two levels reflect the conversion rate of the shots on target in goals and the precision rate of all the shot attempts, respectively. In Sect. 2 we introduce the entire model, and we focus on the Gaussian process for the attack and the defence abilities. Moreover, we assess the binomial assumption for the scores through a Pearson's chi-squared test. We present the application on the English Premier League in Sect. 3, along with parameters' estimates and predictions for the test set season. Sect. 4 concludes.

2 A joint model for the shots and the scores

Here, $\boldsymbol{y}_m = (y_{m1}, y_{m2})$ denotes the vector of observed scores, where y_{m1} and y_{m2} are the number of goals scored by the home team and by the away team in the m-th match of the dataset, respectively. Let $\boldsymbol{s}_m = (s_{m1}, s_{m2})$ denote the shots on the

target and $\boldsymbol{w}_m = (w_{m1}, w_{m2})$ the total number of shots, respectively. For each m, the data information is represented by the joint vector $(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{w})$. The total number of teams considered across the seasons is $T = 34$. In what follows, the nested indexes $h(m), a(m) = 1, \ldots, T$ and $\tau(m)$ identify the home team, the away team and the season τ associated with the m-th game, respectively. The three-stages hierarchical model for the scores and the shots is then specified as follows:

$$\begin{aligned} y_{m1} &\sim \mathsf{Binomial}(s_{m1}, p_{h(m),\tau(m)}) \\ y_{m2} &\sim \mathsf{Binomial}(s_{m2}, q_{a(m),\tau(m)}) \\ s_{m1} &\sim \mathsf{Binomial}(w_{m1}, u_{h(m)}) \\ s_{m2} &\sim \mathsf{Binomial}(w_{m2}, v_{a(m)}) \\ w_{mj}|\theta_{mj} &\sim \mathsf{NegBinomial}(\theta_{mj}, \phi), \quad j = 1, 2. \end{aligned} \tag{1}$$

The *conversion probabilities* p and q are modelled with two inverse logit, depending on the attack and the defence strengths of the competing teams:

$$\begin{aligned} p_{h(m),\tau(m)} &= \text{logit}^{-1}(\mu + att_{h(m),\tau(m)} + def_{a(m),\tau(m)}) \\ p_{a(m),\tau(m)} &= \text{logit}^{-1}(att_{a(m),\tau(m)} + def_{h(m),\tau(m)}). \end{aligned} \tag{2}$$

The attack and defence parameters are assumed to follow two Gaussian processes:

$$\begin{aligned} att_{,\tau} &\sim \mathsf{GP}(\mu_{att}(\tau), k(\tau)), \\ def_{,\tau} &\sim \mathsf{GP}(\mu_{def}(\tau), k(\tau)), \end{aligned} \tag{3}$$

with mean functions $\mu_{att}(\tau) = att_{\cdot,\tau-1}$, $\mu_{def}(\tau) = def_{\cdot,\tau-1}$, and covariance function with generic element $k(\tau)_{i,j} = \exp\{-(\tau_i - \tau_j)^2\} + 0.1$. As outlined in the literature, a 'zero-sum' identifiability constraint within each season is required: for T teams we assume: $\sum_{t=1}^{T} att_{t,\tau} = 0$, $\sum_{t=1}^{T} def_{t,\tau} = 0, \tau = 1, \ldots \mathscr{T}$.

The *shots' precision probabilities* u and v and the *shooting rates* θ_{m1}, θ_{m2} are given a Beta distribution with hyperparameters δ, ε and a Gamma distribution with hyperparameters α, β, respectively:

$$u_{h(m)} \sim \mathsf{Beta}(\delta_{h(m)}, \varepsilon_{h(m)}), \; v_{a(m)} \sim \mathsf{Beta}(\delta_{a(m)}, \varepsilon_{a(m)}) \tag{4}$$

$$\theta_{m1} \sim \mathsf{Gamma}(\alpha_{h(m)}, \beta_{h(m)}), \; \theta_{m2} \sim \mathsf{Gamma}(\alpha_{a(m)}, \beta_{a(m)}). \tag{5}$$

The model is completed by the specification of weakly informative priors for the home effect parameter μ, the overdispersion parameter ϕ, and the hyperparameters $\alpha, \beta, \delta, \varepsilon$.

It is of interest to assess the legitimacy of the binomial distribution for the model above. For this purpose, we consider the empirical distribution of the scores conditioned on a given number of shots on target, $y_{\cdot j}|s_{\cdot j} = z$, $z \in \mathbb{N}$, and we check whether this sample may be thought as drawn from a Binomial(n, p), with n and p fixed. For each z, we performed some Pearson χ^2-tests comparing the empirical distribution

of the scores conditioned on the z-th shot on target and the hypothesized binomial distribution, with n and p estimated from the data. For both the home and the away scores, for each z the Pearson χ^2 test suggests to not reject the null hypothesis of binomial distribution (all the p-values are always greater than the threshold $\alpha = 0.05$).

3 Application: Premier League from 2007/2008 to 2016/2017

We collected the historical data arising from 10 seasons of the English Premier League (EPL), from 2007/2008 to 2016/2017. The data structure of the model is presented in Table 1 with respect to the first match day in EPL 2007/2008. The goal is fitting the model and deriving the parameters' estimates. Secondly, we make predictions for a set of future matches. Model coding has been written using Stan (Carpenter et al., 2017), precisely the Rstan interface. We strictly followed the software guidelines for monitoring the chains' convergence and speeding up the computational times. The chosen number of Hamiltonian Markov Chain iterations is 2000, with a burnin period of 500.

Table 1 Data structure for the first match day, EPL, 2007/2008 season. Each column reports: match, season, home team, away team, home goals, away goals, home shots, away shots , home shots on target, away shots on target.

Match	Season	$h[m]$	$a[m]$	y_{m1}	y_{m2}	w_{m1}	w_{m2}	s_{m1}	s_{m2}
1	07/08	Aston Villa	Liverpool	1	2	10	17	6	7
2	07/08	Bolton	Newcastle	1	3	13	7	9	5
3	07/08	Derby	Portsmouth	2	2	12	12	5	6
4	07/08	Everton	Wigan	2	1	12	14	8	4
5	07/08	Middlesbrough	Blackburn	1	2	10	4	6	4
6	07/08	Sunderland	Tottenham	1	0	9	6	4	3
7	07/08	West Ham	Man City	0	2	9	14	2	5
8	07/08	Arsenal	Fulham	2	1	19	12	13	9
9	07/08	Chelsea	Birmingham	3	2	19	6	11	4
10	07/08	Man United	Reading	0	0	22	3	9	2

3.1 Parameters' estimates

As usual in Bayesian inference, posterior means $\pm$ posterior standard deviations or posterior quantiles are the summaries for describing and visualizing the posterior distribution of the parameters.

The attack and defence abilities are directly connected with the scoring probabilities in (2), and modelled as Gaussian processes in (3) in terms of seasonal evolu-

tion. Fig. 1 displays the 50% posterior intervals for the attack (solid red line) and the defence (solid black line) in formula (3) across the seasons. Higher values for the attack are associated with a greater propensity to convert the shots on target in goals; conversely, lower values for the defence correspond to a better ability to not concede goals. These plots may explain something unexpected even for football experts and help in revealing new instances behind events thought as completely unpredictable. For instance, Leicester won the Premier League 2015/2016 with associated initial odds such as 1:5000, and no one, at the beginning of the season, could have predicted that performance. However, there is a surprising trend that emerges clearly: starting from 2007/2008 season, Leicester dramatically improved the propensity to convert the shots in goal and, at the same time, reinforced its defence. The values registered for the attack and the defence are among the highest and the lowest in the EPL respectively. Maybe, the victory of Leicester, despite highly surprising, was less unpredictable than what the experts had thought.

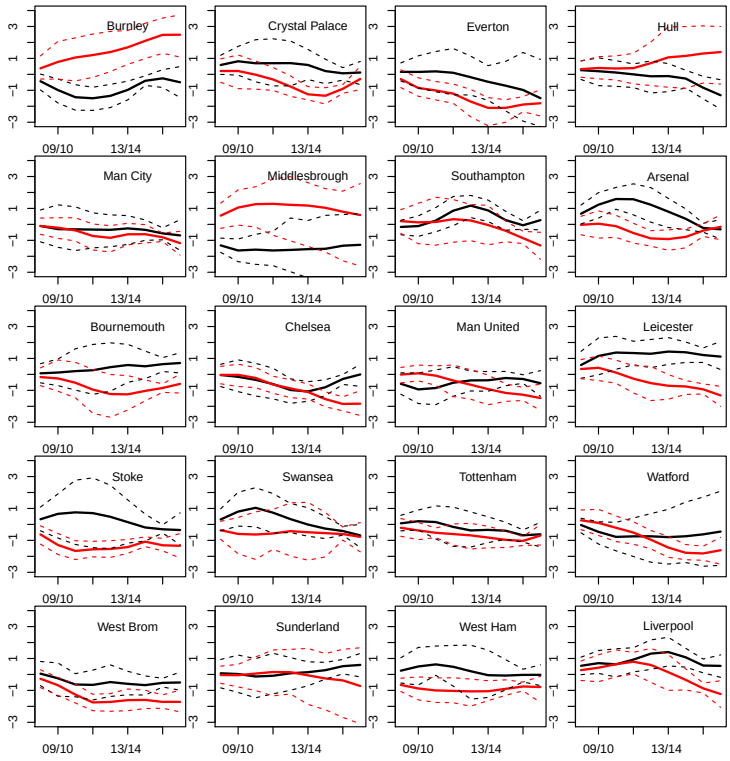

Fig. 1 Posterior estimates for the attack (solid black lines) and defence (solid red lines) across the nine seasons considered, from 2008/2009 to 2016/2017, for the twenty teams belonging to the EPL in the 2016/2017 season.

The power of model (1) is to represent a sort of scores' genesis, able to approximately reproduce the features of the real game. The scores represent the final level, depending on the population of the shots on target, which in turn depends on the population of the total shots. The posterior means $\pm$ standard error for the average conversion probabilities p, q and the home precision probabilities u are displayed in Fig. 2 (left panel). For the majority of the teams, the precision probabilities (black bars) are higher than the conversion probabilities (red and blue bars), and this is intuitive in terms of football features: usually, the ratio between the shots on target and the total shots tends to be higher than the ratio between the goals and the shots on target. However, Middlesbrough, one of the three relegated teams at the end of the 2016/2017 season, is associated with the highest precision probability—half of its shots are on the target—but with the lowest conversion—only about one attempt over ten in the targets corresponds to a score. Conversely, the precision probabilities for Leicester almost overlap the conversion probabilities. For what concerns the total shots, Fig. 2 (right panel) displays the average shots rates, where Chelsea, Manchester City, Tottenham and Liverpool register the highest values. For each team, the trend is to kick more when playing at home.

Although a broad analysis of these statistics should benefit from other comparisons and covariates, we imagine these kinds of plots and summaries could be beneficial for football managers or tactic experts, at least in a naive perspective summarized by the quote 'kick less, kick better'.

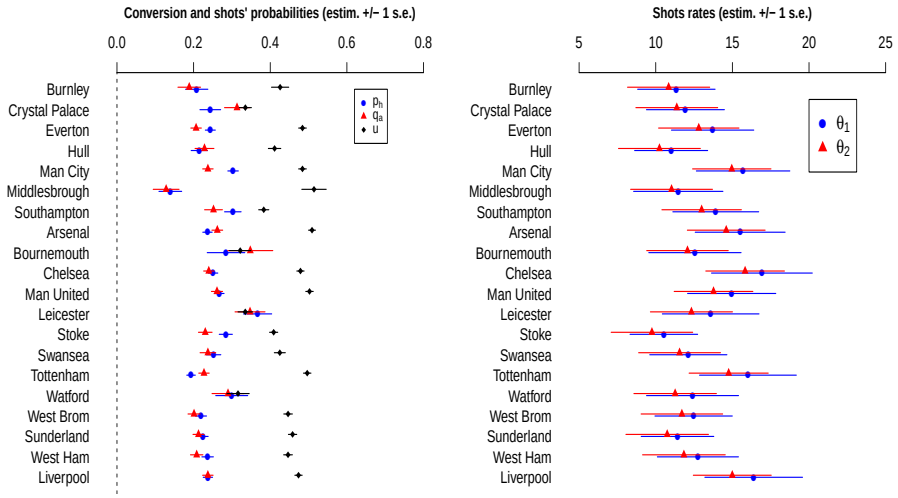

Fig. 2 Average of the posterior estimates for the conversion probabilities p_h and q_a and for the home precision probabilities u (left panel); average of the posterior estimates for the shots rates $\theta_{\cdot 1}, \theta_{\cdot 2}$ (right panel) for the twenty teams belonging to the EPL 2016/2017.

3.2 Prediction and posterior probabilities

Making predictions for future games and seasons is of great appeal for sport statisticians. We used historical data arising from the past seasons for making predictions about the tenth season, the EPL 2016/2017. As usual in a Bayesian framework, the prediction for a new dataset may be performed directly via the posterior predictive distribution for our unknown set of observable values. Fig. 3 displays the posterior 50% credible bars (grey ribbons) for the predicted achieved points for each team for the season 2016/2017, together with the observed final ranks. At a first glance, the model correctly detects Chelsea as EPL champion at the end of the 2016/2017 season, and Middlesbrough and Hull City relegated in Championship. Manchester City, Tottehnham and Arsenal appear to be definitely underestimated, whereas Manchester United and Leicester are quite overestimated. Globally, the predictions reflect the observed pattern.

Table 2 reports the model posterior probabilities being the first, the second and the third relegated team; as may be noticed, Sunderland was pretty unlikely to be relegated in Championship. Conversely, Burnley had an high probability to be relegated, but it performed better than the predictions.

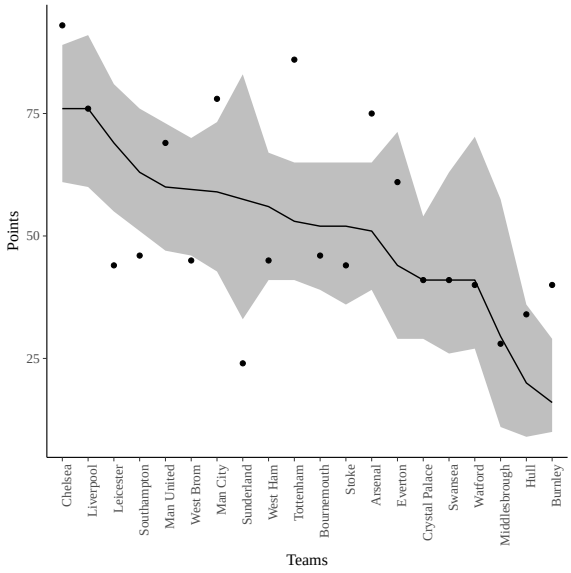

Fig. 3 Posterior 50% credible bars (grey ribbons) for the achieved final points of English Premier League 2016/2017. Black dots are the observed points. Black lines are the posterior medians.

Table 2 Estimated posterior probabilities for each team being the first, the second, and the third relegated team in the Premier League, 2016/2017, together with the observed rank and the number of points achieved (relegated predicted teams by the model are emphasized).

Team	P(1st rel)	P(2nd rel)	P(3d rel)	Actual rank	Points
Burnley	0.096	0.161	0.245	16	40
Hull	0.103	0.198	0.254	18	34
Middlesbrough	0.063	0.117	0.226	19	28
Sunderland	0.055	0.045	0.027	20	24

4 Discussion

We have proposed a Bayesian hierarchical model consisting of a three-stage hierarchy for the scores, the shots on target and the number of total shots. The main novelty is the inclusion of an important latent football feature represented by the kicking ability of each team, modelled both in terms of intensity and precision. Preliminary results on future matches seem to be promising in terms of predictive accuracy. Model comparisons and goodness of fit tools are issues of future interest.

References

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

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell: Stan: A probabilistic programming language. *Journal of Statistical Software 76*(1) (2017)

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

Egidi, L., F. Pauli, and N. Torelli: Combining historical data and bookmakers' odds in modelling football scores. *arXiv preprint arXiv:1802.08848* (2018)

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

Koopman, S. J. and R. Lit: A dynamic bivariate poisson model for analysing and forecasting match results in the english premier league. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 178*(1), 167–186 (2015)

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

Owen, A.: Dynamic bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA Journal of Management Mathematics 22*(2), 99–113 (2011)