*Brazilian Journal of Probability and Statistics*
2017, Vol. 31, No. 4, 746–764
https://doi.org/10.1214/17-BJPS379


# A Bayesian approach for a zero modified Poisson model to predict match outcomes applied to the 2012–13 La Liga season

**Katiane S. Conceição, Adriano K. Suzuki, Marinho G. Andrade and Francisco Louzada**

*University of São Paulo*

**Abstract.** In any sports competition, strong interest is devoted to the knowledge on the team that will be champion. The result of a match, the chance of a team either qualifying for a specific tournament, or relegating, the best attack and defense are all topics of interest. This paper presents a Bayesian methodology for modeling the number of goals scored by a team based on Zero-Modified Poisson distribution. An important advantage of this distribution is the flexibility in modeling count data without previous knowledge of the sampling characteristic with respect to the frequency of zeros (inflated, standard, deflation). These characteristics are present in the data sets referring to the number of goals scored by different teams. Inference procedures and computational simulation studies are also discussed. The proposed methodology was applied to the 2012–13 La Liga and the results were compared with those of the Poisson model using the De Finetti measure an percentage of correct predictions.

## 1 Introduction

In football and any sports competition, strong interest is devoted to the knowledge on the team (in a collective sport) or player (in an individual sport) that will be the champion. The result of a match, the chance of a team either qualifying for a specific tournament or relegating, the best attack and the best defense are all topics of interest.

Various studies have focused on football prediction. For example, regarding the World Cup Tournament (WCT), Dyte and Clarke (2000) presented a log-linear Poisson regression model which took the FIFA ratings as covariates. They provided some results on the predictive power of the model and simulation results to estimate the probabilities of a team winning the 1998 WCT. Using a counting processes approach, Volf (2009) modeled the development of a match score as two interacting time-dependent random point processes. The interactions between teams were analyzed via a semi-parametric multiplicative regression model of intensity.

*Key words and phrases.* Applied probability, OR in sports, ZMP model, Bayesian approach, simulation.

Received June 2016; accepted September 2017.

The author applied his model to the analysis of the performance of the eight teams that played in the quarter-finals of the 2006 WCT. Suzuki et al. (2009) proposed a Bayesian methodology to predict the match results using 'experts' opinions and the FIFA ratings as prior information. The method was used to calculate the win, draw and loss probabilities for each match and estimate the classification probabilities in a group stage and the winning tournament chances for each team at the 2006 WCT.

Keller (1994) fitted the Poisson distribution to the number of goals scored by England, Ireland, Scotland and Wales in the British International Championship from 1883 to 1980. Lee (1997) considered a Poisson regression to model the number of goals of a football team, whose average reflected the strength of the team, the quality of the opposition and the home advantage (if it is the home team). The independence between the goals scored by the two teams was assumed and the methodology was used for the 1995–96 English Premier League.

In a different approach, Brillinger (2008) directly modeled the win, draw and loss probabilities by applying a trinomial regression model to the Brazilian 2006 Series A championship. The total points, probability of a team winning the championship and probability of ending the season in the top four places were estimated by simulations for each team.

Karlis and Ntzoufras (2009) applied Skellam's distribution to model the difference in the number of goals of each match. The authors argue this approach relies neither on independence nor on the marginal Poisson distribution assumptions for the number of goals scored by the teams. A Bayesian analysis to predict of match outcomes for the English Premiere League (2006–07 season) was performed with a log-linear link function and non-informative prior distributions for the model parameters.

Several studies have been conducted on the Poisson distribution for modeling the number of goals scored. However, a sample formed by the number of goals of a team in different games usually show a variance larger or smaller than the mean, because the observed frequency of zeros differs from the expected frequency when the Poisson distribution is considered. To overcome this problem, a modification in the Poisson distribution must be considered for an adequate modeling of the frequency of zero.

In this study, we assume the number of goals scored by a team follows the Zero Modified Poisson (ZMP) model presented in Dietz and Böhning (2000). Under a Bayesian approach, we present an inferential procedure and a computational simulation study is discussed (see Conceição, Andrade and Louzada, 2013 and Conceição, Andrade and Louzada, 2014). The methodology was used for the 2012–13 La Liga and the De Finetti measure (De Finetti, 1972). The prediction of the results of a given round is done sequentially, characterizing a 1-step ahead prediction. The new data is then incoporated and the parameters are again estimated with the new series and the percentage of correct forecasts were used to quantify the predictive quality of the model.

The paper is organized as follows: Section 3 presents the probabilistic model; Section 4 describes the ZMP Regression Model, the Bayesian inference, the computational simulation procedure and two measures used to quantify our modeling predictive quality; Section 5 provides the results of a simulation study conducted to estimate some probabilities of interest, such as single match (fixed period of time), champion, classification for the 2013–14 UEFA Champions League group phase and relegation (sequential period); finally, Section 6 discusses the results and suggests some future work.

## 2 Real dataset: 2012–13 La Liga season

The Spanish first-division football, known as La Liga season or Liga BBVA, is one of the most popular professional sport leagues worldwide. It is played by 20 teams and the competition format follows the double round-robin format, in which a team plays every other team twice, once at home and once away, in a total of 380 matches. Teams are awarded three points per win, one point per draw and no points per loss and are ranked per total points. The highest-ranked team at the end of the competition is crowned champion. The three lowest ranked teams are relegated to the Second Division. If points are equal between two or more teams, the teams are ranked according to the rules established by the Royal Spanish Football Federation.

In this section, we illustrate the methodology using the total number of goals scored by each team in the 2012–13 La Liga season.

Table 1 shows the number of matches associated with the number of goals scored by each team. For instance, Zaragoza did not score in 18 out of 38 matches while Barcelona scored at least one goal in 100% of their games. After analyzing the data involving the total number of goals scored by each team, we observed that some teams did not score goals in various matches and others scored at least one goal in every game. This fact can indicate that it may not be possible to consider a family of distribution, such as the Poisson distribution, to describe the behavior of the number of goals scored by different teams (see Table 1), and besides, two conflicting situations may occur when we consider the data sets corresponding to the number of goals scored by each team in different matches. For some teams, the data sets have a large frequency of zeros, that is, zero inflated data sets, and for others teams the data sets have under recorded zeros, that is, zero deflated data sets. Thus, in order to consider a distribution that fits the data adequately, the above situations often require previous knowledge of the occurrence of zero inflation or deflation in the sample.

The results presented in Table 1 show for some teams a clear discrepancy between the observed frequency and the expected frequency of zero for some teams when the data is modelled using the Poisson distribution and the maximum likelihood procedure to estimate the model parameter. The observed frequency of zero

**Table 1** *Descriptive analysis of the dataset containing the number of goals scored by a team in the Spanish league* 2012–13

| Teams | Number of goals | | | | | | | Mean | Variance | Zero expected (Poisson Dist.) |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | | | |
| Athletic Bilbao | 9 | 17 | 9 | 3 | 0 | 0 | 0 | 1.158 | 0.785 | 11.938 |
| Atletico Madrid | 8 | 14 | 6 | 4 | 4 | 1 | 1 | 1.711 | 2.373 | 6.869 |
| Barcelona | 0 | 4 | 13 | 7 | 7 | 6 | 1 | 3.026 | 1.864 | 1.843 |
| Betis | 9 | 15 | 6 | 4 | 2 | 2 | 0 | 1.500 | 1.932 | 8.479 |
| Celta | 11 | 18 | 8 | 1 | 0 | 0 | 0 | 0.974 | 0.621 | 14.352 |
| Deportivo La Coruña | 11 | 16 | 4 | 5 | 2 | 0 | 0 | 1.237 | 1.375 | 11.031 |
| Espanyol | 16 | 8 | 8 | 5 | 1 | 0 | 0 | 1.132 | 1.415 | 12.256 |
| Getafe | 10 | 16 | 9 | 3 | 0 | 0 | 0 | 1.132 | 0.820 | 12.256 |
| Granada | 13 | 15 | 8 | 2 | 0 | 0 | 0 | 0.974 | 0.783 | 14.352 |
| Levante | 13 | 15 | 6 | 3 | 1 | 0 | 0 | 1.053 | 1.078 | 13.263 |
| Malaga | 10 | 14 | 6 | 5 | 3 | 0 | 0 | 1.395 | 1.543 | 9.420 |
| Mallorca | 10 | 15 | 12 | 0 | 1 | 0 | 0 | 1.132 | 0.820 | 12.256 |
| Osasuna | 17 | 13 | 5 | 2 | 1 | 0 | 0 | 0.868 | 1.036 | 15.945 |
| Rayo Vallecano | 12 | 8 | 12 | 6 | 0 | 0 | 0 | 1.316 | 1.195 | 10.194 |
| Real Madrid | 4 | 4 | 12 | 5 | 6 | 6 | 1 | 2.711 | 2.698 | 2.527 |
| Real Sociedad | 7 | 9 | 11 | 6 | 4 | 1 | 0 | 1.842 | 1.812 | 6.022 |
| Sevilla | 9 | 14 | 6 | 5 | 3 | 1 | 0 | 1.526 | 1.824 | 8.259 |
| Valencia | 8 | 8 | 12 | 6 | 3 | 1 | 0 | 1.763 | 1.753 | 6.517 |
| Valladolid | 7 | 19 | 9 | 2 | 0 | 0 | 1 | 1.289 | 1.238 | 10.466 |
| Zaragoza | 18 | 9 | 7 | 3 | 0 | 1 | 0 | 0.974 | 1.432 | 14.352 |

is higher than expected frequency for Espanyol and Zaragoza indicating a possible zero inflation while we see the opposite situation, indicating a possible zero deflation for Celta and Valladolid.

Following this context, in this paper we assume the number of goals scored by a team follows the Zero-Modified Poisson (ZMP) model as an alternative to the Poisson model commonly used. The zero-modification consists of including of an additional parameter in the usual Poisson distribution, with the principal role to modify the probability of zero, increasing or decreasing the chance of occurrence of zero. Thus, the ZMP model is flexible enough to be adjusted in both situations (zero-inflation or zero-deflation), without requiring any previous knowledge of the type of modification of the zero frequency present in the dataset. Additionally, the ZMP model has the Poisson model as a particular case. Therefore, the ZMP model is adequate to represent datasets that present any of the situations shown in Table 1.

## 3 Zero-modified Poisson distribution

Let $Y$ be a random variable defined in a set of non-negative integers, $A_0 = \{0, 1, 2, \ldots\}$, and let $P_{\text{ZMP}}(Y = y)$ denote the probability that random variable $Y$

takes a value $y$. The random variable $Y$ is said to have a ZMP distribution with parameters $\mu$ and $p$ if

$$P_{\mathrm{ZMP}}(Y=y)=(1-p)I(y)+pP_P(Y=y), \qquad y\in A_0, \tag{3.1}$$

where $P_P(Y=y)$ is the probability function of a Poisson random variable with mean parameter $\mu$ and $I(y)$ is an indicator function, that is, $I(y)=1$ if $y=0$ and $I(y)=0$ otherwise. Parameter $p$ is subject to the condition (called $p$-condition)

$$0\le p\le \frac{1}{1-P_P(Y=0)}. \tag{3.2}$$

Note that the distribution given in equation (3.1) is not the mixture distribution typically adjusted to zero-inflated data, since parameter $p$ can assume values higher than one to also accommodate the zero-deflation. However, for all values of $p$ between 0 and boundary $1/(1-P_P(Y=0))$, equation (3.1) corresponds to a probability function (for more details, see Dietz and Böhning, 2000, Conceição, Andrade and Louzada, 2013, Conceição, Andrade and Louzada, 2014). The mean and variance of $Y$ are, respectively, $\mu_{\mathrm{ZMP}}=p\mu$ and $\sigma^2_{\mathrm{ZMP}}=p\{\mu+(1-p)\mu^2\}$.

Different values of $p$ lead to a different ZMP distribution, as seen in the evaluation of the proportion of additional or missing zeros given by

$$\begin{aligned} P_{\mathrm{ZMP}}(Y=0)-P_P(Y=0) &= (1-p)+pP_P(Y=0)-P_P(Y=0)\\ &= (1-p)\big(1-P_P(Y=0)\big). \end{aligned} \tag{3.3}$$

According to (3.3), parameter $p$ controls the frequency of zeros. When $p=0$ in (3.3), $P_{\mathrm{ZMP}}(Y=0)=1$. Therefore, (3.1) is the degenerate distribution with all mass at zero. For all $0<p<1$ in (3.3), we have $(1-p)(1-P_P(Y=0))>0$. Therefore, $P_{\mathrm{ZMP}}(Y=0)>P_P(Y=0)$ and (3.1) is the Zero-Inflated Poisson (ZIP) distribution which has a proportion of additional zeros. When $p=1$ in (3.3), $P_{\mathrm{ZMP}}(Y=0)-P_P(Y=0)=0$. Therefore, $P_{\mathrm{ZMP}}(Y=0)=P_P(Y=0)$ and (3.1) is the usual Poisson distribution. For all $1<p<1/(1-P_P(Y=0))$ in (3.3), we have $(1-p)(1-P_P(Y=0))<0$. Therefore, $P_{\mathrm{ZMP}}(Y=0)<P_P(Y=0)$ and (3.1) is the Zero-Deflated Poisson (ZDP) distribution. Finally, $p=1/(1-P_P(Y=0))$ in (3.3) implies $P_{\mathrm{ZMP}}(Y=0)=0$. Therefore, (3.1) is the Zero-Truncated Poisson (ZTP) distribution given by

$$P_{\mathrm{ZTP}}(Y=y)=\frac{P_P(Y=y)}{1-P_P(Y=0)}\big(1-I(y)\big).$$

The ZMP distribution described by equation (3.1) can be written as

$$P_{\mathrm{ZMP}}(Y=y)=\big(1-p\big(1-P_P(Y=0)\big)\big)I(y)+p\big(1-P_P(Y=0)\big)P_{\mathrm{ZTP}}(Y=y),$$

where $P_{\mathrm{ZTP}}(Y=y)$ is the ZTP distribution.

Another parametrization of the ZMP distribution can be obtained by considering $\omega=p(1-P_P(Y=0))$,

$$P_{\mathrm{ZMP}}(Y=y)=(1-\omega)I(y)+\omega P_{\mathrm{ZTP}}(Y=y). \tag{3.4}$$

Its advantage is $\omega$ and $\mu$ are orthogonal, which enables the estimation of $\omega$ independently of $\mu$. However, the parametrization given in (3) enables inferences about parameter $p$ used to identify the type of zero modification (inflated or deflated).

## 4 The ZMP model

The ZMP distribution can be used to model goals scored by teams by specifying the number of goals scored by team $k$ as the response variable. Thus, we can specify

$$Y_k \sim \text{ZMP}(\mu_k, p_k),$$

for $k = 1, 2, \ldots, K$, where $K$ is the number of different teams competing with each other. Concerning the model parameter $\mu_k$, we adopted the following structure:

$$\log(\mu_k) = \beta_0 + \beta_1 I_H(k) + \beta_{A_k} + \beta_{D_k}, \tag{4.1}$$

where $\beta_0$ is a constant parameter, $\beta_1$ is the home effect parameter, $I_H(k)$ is an indicator function that means $I_H(k) = 1$ if team $k$ plays at home and $I_H(k) = 0$ otherwise, $\beta_{A_k}$ is the attacking parameter of team $k$ and $\beta_{D_k}$ is the defensive parameter of the team competing with team $k$ (for more details about equation (4.1), see Lee, 1997 and Saraiva et al., 2016). Note that, in this formulation, a team with a good defense will have a negative defense effect because this will decrease the expected number of goals of the opposing team. On the other hand, a team with a positive defense effect increases the expected number of goals of the opponent.

Alternatively, if we specify

$$Y_k \sim \text{ZMP}(\mu_k, \omega_k),$$

as written in (3.4), we have that:

$$\omega_k = p_k\big(1 - P_P(Y = 0; \mu_k)\big). \tag{4.2}$$

We propose using the constraints that the sum of parameters $\beta_{A_k}$ and the sum of parameters $\beta_{D_k}$ are zero for making the model identifiable:

$$\sum_{k=1}^{K} \beta_{A_k} = 0 \quad \text{and} \quad \sum_{k=1}^{K} \beta_{D_k} = 0.$$

In the ZMP model, the parameters of interest for each team $k$ are represented by vector $\boldsymbol{\beta}_k^\top = (\beta_0 \;\; \beta_1 \;\; \beta_{A_k} \;\; \beta_{D_k})$ and $p_k$ (or $\omega_k$). For inference, we adopted a fully Bayesian approach, which has the advantage of incorporating prior information. The likelihood, prior and posterior densities for the parameters in the model are presented below.

### 4.1 Inference

Let $\boldsymbol{y}_k^\top = (y_{k_1} \;\; y_{k_2} \;\; \cdots \;\; y_{k_n})$ be an observation vector of the independent random variables $Y_{k_i}$ which has ZMP distribution with parameters $\mu_{k_i}$ and $p_k$ (or $\omega_k$), $i = 1, \ldots, n$, where $n$ is the number of games and $y_{k_i}$ corresponds to the number of goals scored by team $k$ in game $i$. Let $\boldsymbol{\mu}_k = (\mu_{k_1} \;\; \mu_{k_2} \;\; \ldots \;\; \mu_{k_n})$ and $p_k$ (or $\omega_k$) be parameters, where $\mu_{k_i}$ and $p_k$ (or $\omega_k$) are related to observation $y_{k_i}$. Consider the parametric vector $\boldsymbol{\beta}_k^\top = (\beta_0 \;\; \beta_1 \;\; \beta_{A_k} \;\; \beta_{D_k})$. For simplifications, we defined matrix $\boldsymbol{X}_k$ of dimensions $n \times 4$ whose rows are composed of vectors $\boldsymbol{x}_{k_i} = (1 \;\; I_H(k_i) \;\; 1 \;\; 1)$, so that $\boldsymbol{x}_{k_i}\boldsymbol{\beta}_k = \beta_0 + \beta_1 I_H(k_i) + \beta_{A_k} + \beta_{D_k}$.

Consider the ZMP model parameterized in $\omega_k$ according to (3.4). The likelihood associated with the observation vector $\boldsymbol{y}_k$ of team $k$ is given by

$$\mathcal{L}_k(\boldsymbol{\mu}_k, \omega_k; \boldsymbol{y}_k) = \prod_{i=1}^{n} \left\{ (1-\omega_k)^{I(y_{k_i})} \left( \frac{\omega_k P_P(Y_{k_i} = y_{k_i})}{1 - P_P(Y_{k_i} = 0)} \right)^{1 - I(y_{k_i})} \right\}. \tag{4.3}$$

Substituting in (4.3) the equation of $\mu_{k_i}$ given by

$$\mu_{k_i} = e^{\beta_0 + \beta_1 I_H(k_i) + \beta_{A_k} + \beta_{D_k}} = e^{\boldsymbol{x}_{k_i}\boldsymbol{\beta}_k},$$

the likelihood associated with the observation vector $\boldsymbol{y}_k$ of team $k$ can be written as

$$\mathcal{L}_k(\boldsymbol{\beta}_k; \boldsymbol{y}_k) = \prod_{i=1}^{n} \left\{ (1-\omega_k)^{I(y_{k_i})} \left( \omega_k \cdot \frac{e^{-e^{\boldsymbol{x}_{k_i}\boldsymbol{\beta}_k}} e^{y_{k_i}\boldsymbol{x}_{k_i}\boldsymbol{\beta}_k}}{(1 - e^{-e^{\boldsymbol{x}_{k_i}\boldsymbol{\beta}_k}}) y_{k_i}!} \right)^{1 - I(y_{k_i})} \right\}.$$

The log-likelihood associated with the observation vector $\boldsymbol{y}_k$ of team $k$ is given by

$$\begin{aligned}
\ell_k(\boldsymbol{\beta}_k, \omega_k; \boldsymbol{y}_k) &= \sum_{i=1}^{n} \left\{ \big(1 - I(y_{k_i})\big) \left[ \log\left( \frac{e^{-e^{\boldsymbol{x}_{k_i}\boldsymbol{\beta}_k}} e^{y_{k_i}\boldsymbol{x}_{k_i}\boldsymbol{\beta}_k}}{(1 - e^{-e^{\boldsymbol{x}_{k_i}\boldsymbol{\beta}_k}}) y_{k_i}!} \right) + \log(\omega_k) \right] \right. \\
&\quad \left. + I(y_{k_i}) \log(1 - \omega_k) \right\} \\
&= \sum_{i=1}^{n} \left\{ \big(1 - I(y_{k_i})\big) \log\left( \frac{e^{-e^{\boldsymbol{x}_{k_i}\boldsymbol{\beta}_k}} e^{y_{k_i}\boldsymbol{x}_{k_i}\boldsymbol{\beta}_k}}{(1 - e^{-e^{\boldsymbol{x}_{k_i}\boldsymbol{\beta}_k}}) y_{k_i}!} \right) \right\} \\
&\quad + \sum_{i=1}^{n} \big\{ \big(1 - I(y_{k_i})\big) \log(\omega_k) + I(y_{k_i}) \log(1 - \omega_k) \big\} \\
&= \ell_k^+(\boldsymbol{\beta}_k; \boldsymbol{y}_k) + \ell_k^0(\omega_k; \boldsymbol{y}_k).
\end{aligned} \tag{4.4}$$

From (4.4), $\ell_k^+(\boldsymbol{\beta}_k; \boldsymbol{y}_k)$ depends only on the positive values of $\boldsymbol{y}_k$. Denoting by $\boldsymbol{y}_k^{+\top} = (y_{k_1}^+ \;\; y_{k_2}^+ \;\; \cdots \;\; y_{k_{n^+}}^+)$ the vector with the $n^+$ positive observations from

$\boldsymbol{y}_k$ and $\boldsymbol{X}_k^+$ the matrix of dimensions $n^+ \times 4$ whose rows are composed of vectors $\boldsymbol{x}_{k_j}^+ = (1 \;\; I_H(k_j) \;\; 1 \;\; 1)$, $j = 1, \ldots, n^+$, the log-likelihood for $\boldsymbol{\beta}_k$ based on the supposition that $\boldsymbol{y}_k^+$ comes from a ZTP distribution is given by

$$\begin{aligned}
\ell_k(\boldsymbol{\beta}_k; \boldsymbol{y}_k^+) &= \sum_{j=1}^{n^+} \left\{ \frac{P_P(Y_{k_j}^+ = y_{k_j}^+)}{1 - P_P(Y_{k_j}^+ = 0)} \right\} \\
&= \sum_{j=1}^{n^+} \left\{ \log\left( \frac{e^{-e^{\boldsymbol{x}_{k_j}^+ \boldsymbol{\beta}_k}} e^{y_{k_j}^+ \boldsymbol{x}_{k_j}^+ \boldsymbol{\beta}_k}}{(1 - e^{-e^{\boldsymbol{x}_{k_j}^+ \boldsymbol{\beta}_k}}) y_{k_j}^+!} \right) \right\} \\
&= \sum_{j=1}^{n^+} \{ y_{k_j}^+ \boldsymbol{x}_{k_j}^+ \boldsymbol{\beta}_k - e^{\boldsymbol{x}_{k_j}^+ \boldsymbol{\beta}_k} - \log(1 - e^{-e^{\boldsymbol{x}_{k_j}^+ \boldsymbol{\beta}_k}}) - \log(y_{k_j}^+!) \},
\end{aligned}$$

for all values of $y_{k_j}^+ > 0$.

Since $\ell_k(\boldsymbol{\beta}_k; \boldsymbol{y}_k^+) = \ell_k^+(\boldsymbol{\beta}_k; \boldsymbol{y}_k)$, the log-likelihood $\ell_k(\boldsymbol{\beta}_k; \boldsymbol{y}_k)$ of the ZMP model is equivalent to the log-likelihood $\ell_k(\boldsymbol{\beta}_k; \boldsymbol{y}_k^+)$ of the ZTP model added of term $\ell_k^0(\boldsymbol{\beta}_k; \boldsymbol{y}_k)$ and given by

$$\begin{aligned}
\ell_k^0(\boldsymbol{\beta}_{2k}; \boldsymbol{y}_k) &= \sum_{i=1}^{n} \{ (1 - I(y_{k_i})) \log(\omega_k) + I(y_{k_i}) \log(1 - \omega_k) \} \\
&= n^+ \log(\omega_k) + (n - n^+) \log(1 - \omega_k).
\end{aligned}$$

Since there are $K$ teams that act independently, the complete log-likelihood is given by

$$\ell(\boldsymbol{\beta}, \boldsymbol{\omega}; \mathcal{D}) = \sum_{k=1}^{K} \ell_k(\boldsymbol{\beta}_k, \omega; \boldsymbol{y}_k),$$

where $\boldsymbol{\beta}^\intercal = (\beta_0 \;\; \beta_1 \;\; \beta_{A_1} \;\; \cdots \;\; \beta_{A_K} \;\; \beta_{D_1} \;\; \cdots \;\; \beta_{D_K})$ and $\mathcal{D} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K\}$ are, respectively, the parametric vector and the dataset formed by observation vectors of all teams.

We shall consider for $\boldsymbol{\beta}$ a multivariate Gaussian prior density with mean vector zero and diagonal precision matrix $10^{-3}\boldsymbol{I}$. Here, $\boldsymbol{I}$ is an indent $(2K + 2) \times (2K + 2)$ matrix, therefore $\boldsymbol{\beta} \sim N(\boldsymbol{0}, 10^3 \boldsymbol{I})$. For each parameter $\omega_k$, $k = 1, \ldots, K$, we consider a Uniform prior density, $U(0, 1)$.

Although we have considered a vague priori densities for the parameters of the model referring to the home, attack and defense factors, more informative priori densities can be elicited considering the specialists' opinion. For example, as it was made by the authors Suzuki et al. (2009), who considered a power priori for the parameters of the model, adjusted to data of the 2006 Football World Cup.

The Bayesian approach for the ZMP model can be constructed by writing the joint posterior density for the vector of parameter $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ as

$$P(\boldsymbol{\beta}, \boldsymbol{\omega}|\mathcal{D}) \propto \exp\{\ell(\boldsymbol{\beta}, \boldsymbol{\omega}; \mathcal{D})\} P(\boldsymbol{\beta}, \boldsymbol{\omega}).$$

From the Bayesian point of view, inferences about the parameters can be based on their marginal posterior densities, which can be obtained by integrating the joint posterior density. In our case, however, analytical solutions for the integrals cannot be obtained. In order to overcome this problem, we use the Metropolis-Hastings algorithm (Chib and Greenberg, 1995), which is an iterative procedure of a broad class of MCMC methods. To implement the algorithm, we consider the full conditional distributions of parameters $\beta_0$, $\beta_1$, $\beta_{A_k}$, $\beta_{D_k}$ and $\omega_k$, for all $k = 1, \ldots, K$. All computational implementations were performed using OpenBUGS and R systems in the R2WinBUGS package. The convergence of the chains was monitored according to the methods recommended by Cowless and Carlin (1996) (package CODA, Plummer et al., 2006). In all cases, the convergence was verified by the Gelman-Rubin diagnosis (Gelman and Rubin, 1992), being very close to 1 (1.01). Estimates are given by the average of the generated MCMC sample. Given the estimates, we use these values to calculate the probability of a win, draw and defeat of each team in the next round.

### 4.2 Deriving the probabilities

For a given match played by teams $Y_1$ and $Y_2$, after the estimation of parameters of the ZMP model, we calculate the probabilities of win ($P_W$), draw ($P_D$) and loss ($P_L$) of team $Y_1$ using the following equations:

$$P_W = P_{\text{ZMP}}(Y_1 > Y_2) = \sum_{i=1}^{\infty} \sum_{j=0}^{i-1} P_{\text{ZMP}}(Y_1 = i) P_{\text{ZMP}}(Y_2 = j), \tag{4.5}$$

$$P_D = P_{\text{ZMP}}(Y_1 = Y_2) = \sum_{i=0}^{\infty} P_{\text{ZMP}}(Y_1 = i) P_{\text{ZMP}}(Y_2 = i), \tag{4.6}$$

and

$$P_L = P_{\text{ZMP}}(Y_1 < Y_2) = \sum_{j=1}^{\infty} \sum_{i=0}^{j-1} P_{\text{ZMP}}(Y_1 = i) P_{\text{ZMP}}(Y_2 = j). \tag{4.7}$$

### 4.3 Algorithm for the simulation

Suppose a tournament is composed of $N$ rounds. For each round $r$, $r = N/2, \ldots, N$, we obtain the final team classification, that is, number of points, victories, draws, defeats, goals scored, goals conceded and goal differences. The final classification is forecasted in a simulation based on the ZMP model involving the following steps:

(a) Fix $n$ as the number of championships to be simulated and $r$ as the number of rounds played. Do $c = 1$ (the counter);
(b) If $c < n$, use the $(r-1) * 10$ observed matches to estimate the parameters of the ZMP model;
(c) For each $M = [N - (r-1)] * 10$ matches to be played, simulate the number of goals scored using the ZMP distribution with the estimated parameters obtained in step (b). Do $c = c + 1$ and return to step (b).

To assemble the final league tables, for each $M$ match predicted, check if there was a victory of team $Y_1$ ($Y_1 > Y_2$), draw ($Y_1 = Y_2$) or victory of team $Y_2$ ($Y_1 < Y_2$). Give 3 points to the winning team and 1 point to both teams in case of a draw. Update the current league table with the simulated results for each $n$ simulated championship. From the final league tables, we can calculate, for example, the chance of a particular team being champion or relegated as follows:

$$P[\text{Team to be champion}] = \#(\text{team ended in the first position in the final league table})/n,$$

$$P[\text{Team to be relegation}] = \#(\text{team ended in the last three positions in the final league table})/n,$$

where # refers to the number of times obtained in the simulation.

### 4.4 Quality of the predictions

As pointed out in Section 1, the De Finetti measure (De Finetti, 1972) was used to quantify our modeling predictive quality. Consider the set of all possible forecasts given by the simplex set $S = \{(P_W, P_D, P_L) \in [0,1]^3 : P_W + P_D + P_L = 1\}$, where $P_W$, $P_D$ and $P_L$ denote the win probability, the draw probability and the loss probability, respectively.

Observe that vertices $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$ of $S$ represent the win, draw and loss outcomes, respectively. Therefore, a way to measure the goodness of a prediction is to calculate the De Finetti distance (De Finetti, 1972), which is the square of the Euclidean distance between the point corresponding to the outcome and the one corresponding to the prediction. For example, if a prediction is $(0.1, 0.6, 0.3)$ and the outcome is a draw $(0,1,0)$, then the De Finetti distance is $(0.1-0)^2 + (0.6-1)^2 + (0.3-0)^2 = 0.26$. We can also associate the average of its De Finetti distances, known as the De Finetti measure, to a set of predictions. Therefore, among some prediction methods, the one with the least De Finetti measure shall be considered the best. Furthermore, the De Finetti value $2/3$ can be considered a reference for comparisons, because an equiprobable predictor which assigns an equal probability for all outcomes ($P_W = P_D = P_L = 1/3$) has $2/3$ as its De Finetti measure.

Before each of the 19 remaining rounds, that is, 20th to 38th rounds, we calculated the win, draw and loss probabilities (see Section 4.2) for all matches and the De Finetti measure (De Finetti, 1972) associated with these predictions.

Another standard way of measuring the goodness of a prediction method is to calculate the percentage of correct forecasts. A forecast $(P_W, P_D, P_L)$ shall be considered correct if the outcome of the highest probability coincides with the observed outcome.

## 5 Data analysis

This section provides the results of the application of the methodology to the 2012–13 La Liga season. The champion of the 2012–13 La Liga was Barcelona, which obtained most overall wins (32 victories), the second best defense (40 goals conceded) and highest number of goals scored (115 goals).

We focused on the single match predictions, as well as the predictions for the whole Tournament. We used the outcomes of the first 190 matches (19 rounds played) of the 2012–13 La Liga season as our data set to predict the following 190 matches (from the 20th to 38th rounds).

### 5.1 Single match prediction

This section provides the forecasts for all the matches of the 28th round, shown in Table 2. We compared the results obtained by ZMP and Poisson models. A forecast $(P_W, P_D, P_L)$ shall be considered correct if the outcome of the highest probability coincides with the observed outcomes. For the 28th round, the ZMP model scored 7 results whereas the Poisson model scored only 4 results. The De Finetti measures obtained were 0.436 and 0.615, respectively.

For all the forecasts for 19 rounds (190 matches predicted), 97 and 83 correct predictions were obtained by the ZMP and Poisson models, respectively, and the associated De Finetti measures were 0.585 and 0.670.

### 5.2 Predictions for the whole tournament—competition simulation

This section addresses the other probabilities of a team being champion, classified for the 2013–14 UEFA Champions League group phase and relegated.

We considered 1000 tournament replications. A tournament replica was obtained by the simulation procedure briefly described above. We calculated the percentage of wins in tournament replicas for each team, the percentage of tournament replica to be qualified for the 2013–14 UEFA Champions League Group stage (ranked among the four best teams) and relegated to the Second Division, known as 2013–14 Liga Adelante (ranked among the three worst teams).

**Table 2** *Forecasts for single matches of the* 28*th round*

| Home Team | Observed Score | Away Team | ZMP Model | | | | | Poisson Model | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | W | D | L | De Finetti | Correct | W | D | L | De Finetti | Correct |
| Deportivo La Coruña | 31 | Celta Vigo | 0.321 | 0.343 | 0.336 | 0.691 | No | 0.497 | 0.253 | 0.25 | 0.38 | Yes |
| Real Sociedad | 41 | Valladolid | 0.484 | 0.256 | 0.260 | 0.399 | Yes | 0.37 | 0.238 | 0.392 | 0.607 | No |
| Getafe | 10 | Athletic Bilbao | 0.515 | 0.256 | 0.228 | 0.353 | Yes | 0.457 | 0.265 | 0.278 | 0.443 | Yes |
| Real Madrid | 52 | Mallorca | 0.739 | 0.128 | 0.133 | 0.102 | Yes | 0.247 | 0.262 | 0.491 | 0.877 | No |
| Valencia | 30 | Betis | 0.482 | 0.206 | 0.313 | 0.409 | Yes | 0.396 | 0.207 | 0.397 | 0.565 | No |
| Málaga | 2 | Espanyol | 0.519 | 0.261 | 0.220 | 0.945 | No | 0.595 | 0.198 | 0.207 | 1 | No |
| Sevilla | 40 | Zaragoza | 0.507 | 0.273 | 0.220 | 0.367 | Yes | 0.609 | 0.191 | 0.2 | 0.229 | Yes |
| Osasuna | 2 | Atlético Madrid | 0.203 | 0.294 | 0.502 | 0.376 | Yes | 0.436 | 0.212 | 0.352 | 0.656 | No |
| Granada | 11 | Levante | 0.322 | 0.318 | 0.360 | 0.699 | No | 0.411 | 0.216 | 0.373 | 0.923 | No |
| Barcelona | 31 | Rayo Vallecano | 0.902 | 0.061 | 0.037 | 0.015 | Yes | 0.452 | 0.244 | 0.304 | 0.452 | Yes |

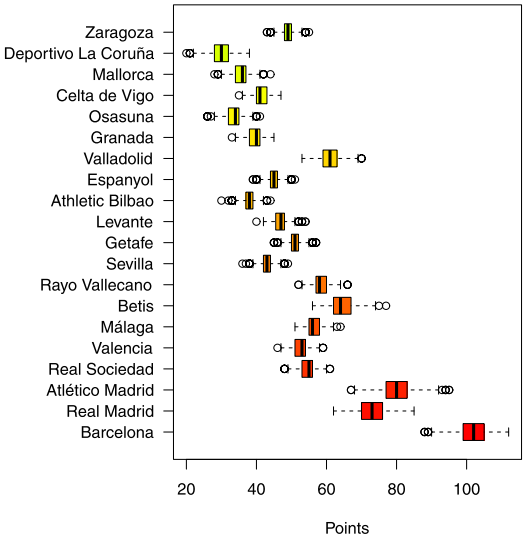


**Figure 1** *Box-plot of the number of points obtained by each team before the* 20*th round.*

5.2.1 *Overall results*. In the 1000 tournament replications, some interesting information can be obtained, such as how many times a team has been champion, how many times a team has ended in the first three positions, the variability in the number of points, goals scored, goals taken, number of victories, losses, draws, etc.

Figures 1 and 2 show, respectively, the box-plots of the predicted numbers of points before the 20th and 35th rounds for each team at the end of the tournament. Figure 1 shows that Barcelona is the favorite to win the tournament and Real Madrid and Atlético Madrid are favorites to be qualified for the 2013–14 UEFA Champions League. A vacancy remains to be played by Real Sociedad, Valencia, Málaga and Betis. In this prediction, Zaragoza was not accredited as a strong candidate for relegation.

According to Figure 2, with only four rounds remaining for the end of the championship, the first three positions were practically defined and a contest was held for the fourth position and the worst teams fought against relegation. In this round, for example, Zaragoza was in the 17th position, tied with Osasuna (16th) and the difference until then with Mallorca, the team ranked in the last position, was only four points. One draw and three defeats in the last four rounds led to the relagation of Zaragoza, which ended in the last position.

Table 3 shows the probabilities of each of the 20 teams reaching each of the 20 positions at the end of the championship. These probabilities were estimated by considering the observed data before the 20th round.

5.2.2 *Some specific results*. Prior to each of the 19 remaining rounds (20th to 38th rounds), a simulation of 1000 whole tournaments was performed so that the

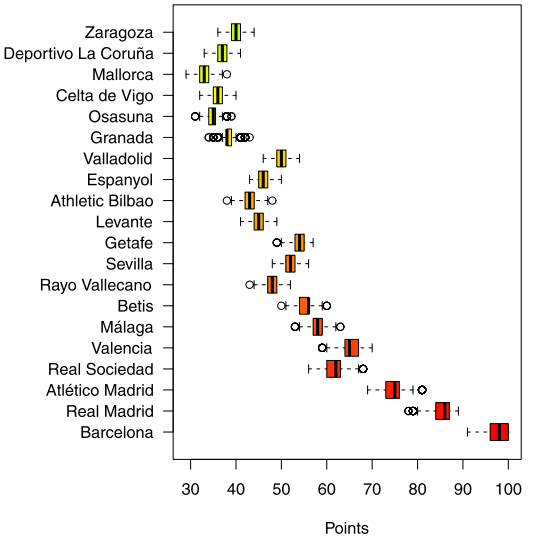


**Figure 2** *Box-plot of the number of points obtained by each team before the* 35*th round.*

mean of probabilities could be obtained. The tables below show the probabilities of tournament wins (Table 4), the probabilities of a team reaching the top four places (Table 5) and the probabilities of a team being relegated (Table 6), for the 19 remaining rounds (20th to 38th rounds) for each team with the highest probabilities.

According to Table 4, for all rounds, Barcelona, the champion, is the team which most frequently ended in the first position in our simulation.

To qualify for the 2013–14 UEFA Champions League Group stage, the teams must be ranked among the four best teams. The probabilities of the six best teams reaching the top four positions are shown in Table 5, where we can observe an intensive dispute between Real Sociedad and Valencia for the fourth position can be observed. Before the last round, Valencia had scored 2 more points than Real Sociedad and both played away in the last round. Taking advantage of the defeat of Valencia to Sevilla by 3–4, Real Sociedad was classified by beating the Deportivo La Coruña by 1–0.

Another probability of interest refers to the relegation of the teams. In a round-robin tournament, the teams extensively dispute to be the champion and qualify for any tournament, but also not to be relegated. The teams are relegated if they are ranked in the three worst positions. The probabilities of the six teams being ranked in the three last positions are shown in Table 6. Four teams in the last round competed to avoid relegation: Celta and the relegated teams Mallorca, Deportivo La Coruña and Zaragoza.

**Table 3** *Mean, standard deviation of final league ranks and probability of the* 20 *positions* (*in %*) *in the* 20*th round*

| Team | Summary | | Probability of each position | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Barcelona | 1.002 | 0.045 | 0.998 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Real Madrid | 2.816 | 0.573 | 0.000 | 0.258 | 0.682 | 0.047 | 0.012 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Atlético Madrid | 2.276 | 0.490 | 0.002 | 0.734 | 0.253 | 0.009 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Real Sociedad | 8.619 | 2.861 | 0.000 | 0.000 | 0.002 | 0.057 | 0.078 | 0.115 | 0.126 | 0.145 | 0.125 | 0.107 | 0.087 | 0.061 | 0.049 | 0.018 | 0.011 | 0.006 | 0.008 | 0.003 | 0.001 | 0.001 |
| Valencia | 7.359 | 2.560 | 0.000 | 0.001 | 0.007 | 0.122 | 0.157 | 0.139 | 0.136 | 0.127 | 0.109 | 0.070 | 0.064 | 0.033 | 0.024 | 0.006 | 0.000 | 0.002 | 0.002 | 0.001 | 0.000 | 0.000 |
| Málaga | 5.627 | 1.922 | 0.000 | 0.002 | 0.026 | 0.340 | 0.213 | 0.152 | 0.110 | 0.066 | 0.044 | 0.023 | 0.015 | 0.004 | 0.001 | 0.002 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Betis | 6.240 | 2.199 | 0.000 | 0.002 | 0.019 | 0.225 | 0.216 | 0.173 | 0.114 | 0.089 | 0.062 | 0.044 | 0.032 | 0.014 | 0.008 | 0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Rayo Vallecano | 8.405 | 2.734 | 0.000 | 0.000 | 0.006 | 0.063 | 0.096 | 0.112 | 0.126 | 0.133 | 0.115 | 0.104 | 0.106 | 0.064 | 0.046 | 0.016 | 0.004 | 0.008 | 0.000 | 0.001 | 0.000 | 0.000 |
| Sevilla | 12.836 | 3.072 | 0.000 | 0.000 | 0.000 | 0.005 | 0.002 | 0.011 | 0.030 | 0.040 | 0.052 | 0.076 | 0.108 | 0.140 | 0.117 | 0.127 | 0.091 | 0.072 | 0.055 | 0.049 | 0.019 | 0.006 |
| Getafe | 10.088 | 2.873 | 0.000 | 0.000 | 0.000 | 0.020 | 0.030 | 0.058 | 0.087 | 0.111 | 0.122 | 0.135 | 0.124 | 0.104 | 0.090 | 0.059 | 0.030 | 0.015 | 0.008 | 0.004 | 0.001 | 0.002 |
| Levante | 8.674 | 2.856 | 0.000 | 0.000 | 0.004 | 0.059 | 0.081 | 0.105 | 0.133 | 0.099 | 0.142 | 0.130 | 0.075 | 0.073 | 0.047 | 0.029 | 0.007 | 0.008 | 0.006 | 0.001 | 0.001 | 0.000 |
| Athletic Bilbao | 14.684 | 2.846 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.004 | 0.005 | 0.013 | 0.016 | 0.041 | 0.058 | 0.072 | 0.115 | 0.139 | 0.129 | 0.130 | 0.101 | 0.088 | 0.056 | 0.032 |
| Espanyol | 16.184 | 2.661 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.000 | 0.000 | 0.008 | 0.010 | 0.007 | 0.021 | 0.045 | 0.064 | 0.098 | 0.111 | 0.136 | 0.140 | 0.134 | 0.138 | 0.086 |
| Valladolid | 8.434 | 2.655 | 0.000 | 0.001 | 0.001 | 0.049 | 0.107 | 0.118 | 0.104 | 0.136 | 0.137 | 0.122 | 0.102 | 0.063 | 0.031 | 0.011 | 0.007 | 0.007 | 0.003 | 0.001 | 0.000 | 0.000 |
| Granada | 17.168 | 2.402 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.002 | 0.008 | 0.016 | 0.029 | 0.034 | 0.057 | 0.071 | 0.120 | 0.128 | 0.173 | 0.197 | 0.164 |
| Osasuna | 18.601 | 1.847 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.002 | 0.001 | 0.006 | 0.007 | 0.025 | 0.037 | 0.055 | 0.079 | 0.121 | 0.217 | 0.448 |
| Celta Vigo | 14.512 | 2.764 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.005 | 0.006 | 0.010 | 0.019 | 0.033 | 0.065 | 0.089 | 0.124 | 0.131 | 0.162 | 0.098 | 0.116 | 0.064 | 0.052 | 0.026 |
| Mallorca | 16.545 | 2.536 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.002 | 0.014 | 0.029 | 0.028 | 0.052 | 0.066 | 0.122 | 0.132 | 0.149 | 0.149 | 0.136 | 0.119 |
| Deportivo La Coruña | 16.486 | 2.580 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.000 | 0.000 | 0.002 | 0.005 | 0.010 | 0.017 | 0.043 | 0.062 | 0.086 | 0.098 | 0.118 | 0.140 | 0.162 | 0.152 | 0.103 |
| Zaragoza | 13.444 | 2.975 | 0.000 | 0.000 | 0.000 | 0.002 | 0.003 | 0.006 | 0.022 | 0.020 | 0.036 | 0.074 | 0.080 | 0.132 | 0.129 | 0.129 | 0.117 | 0.093 | 0.065 | 0.049 | 0.030 | 0.013 |

**Table 4** *Simulation results for the five teams of highest percentages of tournament wins*

| Round | Barcelona | Real Madrid | Atlético Madrid | Real Sociedad | Valencia |
|---|---|---|---|---|---|
| 20 | 0.998 | 0.000 | 0.002 | 0.000 | 0.000 |
| 21 | 0.996 | 0.000 | 0.004 | 0.000 | 0.000 |
| 22 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 23 | 0.998 | 0.001 | 0.001 | 0.000 | 0.000 |
| 24 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 25 | 0.999 | 0.000 | 0.001 | 0.000 | 0.000 |
| 26 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 27 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 28 | 0.999 | 0.001 | 0.000 | 0.000 | 0.000 |
| 29 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 30 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 31 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 32 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 33 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 34 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 35 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 36 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 37 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 38 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |

**Table 5** *Simulation results for the six teams likely to reach the top four positions*

| Round | Barcelona | Real Madrid | Atlético Madrid | Real Sociedad | Valencia | Málaga | Betis |
|---|---|---|---|---|---|---|---|
| 20 | 1.000 | 0.987 | 0.998 | 0.059 | 0.130 | 0.368 | 0.246 |
| 21 | 1.000 | 0.994 | 0.999 | 0.170 | 0.057 | 0.386 | 0.197 |
| 22 | 1.000 | 1.000 | 0.992 | 0.157 | 0.101 | 0.497 | 0.088 |
| 23 | 1.000 | 0.990 | 0.999 | 0.261 | 0.098 | 0.450 | 0.066 |
| 24 | 1.000 | 0.991 | 0.989 | 0.276 | 0.140 | 0.507 | 0.026 |
| 25 | 1.000 | 0.995 | 0.993 | 0.160 | 0.197 | 0.601 | 0.006 |
| 26 | 1.000 | 0.998 | 0.997 | 0.387 | 0.222 | 0.299 | 0.038 |
| 27 | 1.000 | 0.999 | 0.997 | 0.311 | 0.242 | 0.281 | 0.061 |
| 28 | 1.000 | 1.000 | 1.000 | 0.459 | 0.110 | 0.264 | 0.082 |
| 29 | 1.000 | 1.000 | 0.999 | 0.642 | 0.148 | 0.077 | 0.035 |
| 30 | 1.000 | 1.000 | 1.000 | 0.574 | 0.172 | 0.184 | 0.012 |
| 31 | 1.000 | 1.000 | 1.000 | 0.704 | 0.187 | 0.058 | 0.022 |
| 32 | 1.000 | 1.000 | 1.000 | 0.819 | 0.122 | 0.044 | 0.008 |
| 33 | 1.000 | 1.000 | 1.000 | 0.710 | 0.275 | 0.007 | 0.006 |
| 34 | 1.000 | 1.000 | 1.000 | 0.867 | 0.113 | 0.018 | 0.002 |
| 35 | 1.000 | 1.000 | 1.000 | 0.738 | 0.258 | 0.004 | 0.000 |
| 36 | 1.000 | 1.000 | 1.000 | 0.498 | 0.500 | 0.001 | 0.001 |
| 37 | 1.000 | 1.000 | 1.000 | 0.542 | 0.458 | 0.000 | 0.000 |
| 38 | 1.000 | 1.000 | 1.000 | 0.325 | 0.675 | 0.000 | 0.000 |

**Table 6** *Simulation results of the six teams more likely to be ranked in the three last positions*

| Round | Zaragoza | Deportivo La Coruña | Mallorca | Celta de Vigo | Osasuna | Granada |
|---|---|---|---|---|---|---|
| 20 | 0.092 | 0.417 | 0.404 | 0.142 | 0.786 | 0.534 |
| 21 | 0.172 | 0.589 | 0.548 | 0.137 | 0.646 | 0.398 |
| 22 | 0.187 | 0.654 | 0.567 | 0.096 | 0.650 | 0.500 |
| 23 | 0.171 | 0.799 | 0.711 | 0.248 | 0.526 | 0.332 |
| 24 | 0.190 | 0.869 | 0.739 | 0.388 | 0.497 | 0.146 |
| 25 | 0.296 | 0.877 | 0.812 | 0.431 | 0.323 | 0.102 |
| 26 | 0.302 | 0.929 | 0.898 | 0.277 | 0.154 | 0.181 |
| 27 | 0.381 | 0.941 | 0.768 | 0.377 | 0.190 | 0.211 |
| 28 | 0.444 | 0.933 | 0.580 | 0.428 | 0.264 | 0.267 |
| 29 | 0.492 | 0.864 | 0.506 | 0.557 | 0.309 | 0.207 |
| 30 | 0.516 | 0.755 | 0.731 | 0.490 | 0.117 | 0.368 |
| 31 | 0.508 | 0.524 | 0.685 | 0.642 | 0.165 | 0.460 |
| 32 | 0.557 | 0.266 | 0.517 | 0.773 | 0.218 | 0.647 |
| 33 | 0.714 | 0.353 | 0.542 | 0.616 | 0.166 | 0.598 |
| 34 | 0.562 | 0.421 | 0.806 | 0.501 | 0.236 | 0.463 |
| 35 | 0.334 | 0.483 | 0.910 | 0.592 | 0.469 | 0.208 |
| 36 | 0.339 | 0.699 | 0.986 | 0.776 | 0.107 | 0.093 |
| 37 | 0.597 | 0.347 | 0.951 | 0.920 | 0.183 | 0.002 |
| 38 | 0.895 | 0.557 | 0.925 | 0.623 | 0.000 | 0.000 |

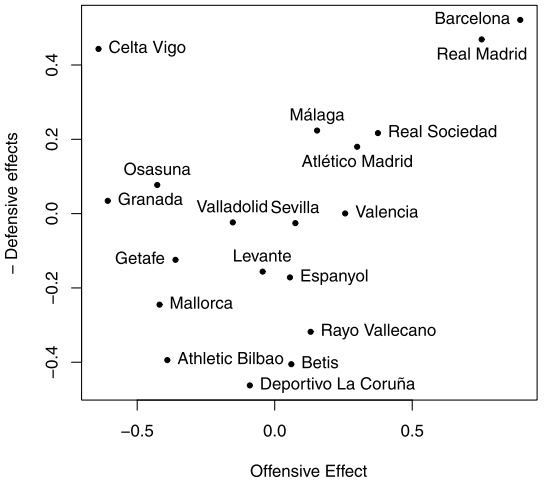


**Figure 3** *Dot plot of the offensive effects versus minus defensive effects for all teams.*

Figure 3 shows the dot plot of the offensive effects versus minus defensive effects for all teams. The teams of strong attack and strong defense ended in the first positions and those of weak defense suffered in the tournament.

## 6 Final remarks

This paper proposed a Bayesian methodology based on the ZMP model that shows good predictive quality, easy implementation and low computational effort for predicting match outcomes. The ZMP model proved very efficient in the predictions in comparison to the widely used Poisson model.

We have reported some probabilities of interest, such as simple match, champion, classification for the 2013–14 UEFA Champions League group phase and relegation. However, other results can be obtained: chance of each team ending a championship in the last position, qualification for the 2013–14 UEFA European League, team of best attack (scoring of goals), team of best defense (taking of few goals), team with most victories, best home team, best away team, etc.

Although our modeling was applied to the 2012–13 La Liga Season, in principle, it is flexible and can be easily adapted to other different tournaments.

Intuitively, if several matches are played under the same conditions, other factors, such as home field advantage, crisis, umpire and atmospheric condition may cause dependence among the datasets. The dependence over time (38 matches), as well as changes that may be suffered by each team throughout the championship are considerations that can make the model more realistic. Approaches considering longitudinal data and random effects can be considered. Although such situations were not considered here, it should be further investigated in the context of our modeling. Particularly, we can also assume a bivariate distribution for $(X, Y)$ to check the presence of dependence.

## References

Brillinger, D. R. (2008). Modeling game outcomes of the Brazilian 2006 series a championship as ordinal-valued. *Brazilian Journal of Probability and Statistics* **22**, 89–104. MR2575392

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. *American Statistician* **49**, 327–335.

Conceição, K. S., Andrade, M. G. and Louzada, F. (2013). Zero-modified Poisson model: Bayesian approach, influence diagnostics and an application to a Brazilian leptospirosis notification data. *Biometrical Journal* **55**, 661–678.

Conceição, K. S., Andrade, M. G. and Louzada, F. (2014). On the zero-modified Poisson model: Bayesian analysis and posterior divergence measure. *Computational Statistics* **29**, 959–980. MR3266043

Cowless, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association* **91**, 883–904.

De Finetti, B. (1972). *Probability, Induction and Statistics*. London: John Wiley. MR0440638

Dietz, E. and Böhning, D. (2000). On estimation of the Poisson parameter in zero-modified Poisson models. *Computational Statistics & Data Analysis* **34**, 441–459.

Dyte, D. and Clarke, S. R. (2000). A ratings based Poisson model for world cup soccer simulation. *Journal of the Operational Research Society* **51**, 993–998.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–511.

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

Keller, J. B. (1994). A characterization of the Poisson distribution and the probability of winning a game. *American Statistician* **48**, 294–298.

Lee, A. (1997). Modeling scores in the Premier League: Is Manchester United really the best? *Chance* **10**, 15–19.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006). Output analysis and diagnostics for MCMC. http://cran.r-project.org/web/packages/coda/index.html.

Saraiva, E. F., Suzuki, A. K., Filho, C. A. O. and Louzada, F. (2016). Predicting football scores via Poisson regression model: Applications to the National Football League. *Communications for Statistical Applications and Methods* **23**, 297–319.

Suzuki, A. K., Salasar, L. E. B., Louzada-Neto, F. and Leite, J. G. (2009). A Bayesian approach for predicting match outcomes: The 2006 (association) football world cup. *Journal of the Operational Research Society* **61**, 1530–1539.

Volf, P. (2009). A random point process model for the score in sport matches. *IMA Journal of Management Mathematics* **20**, 121–131.

Department of Applied Mathematics and Statistics
Institute of Mathematics and Computer Science
University of São Paulo
13566-590, São Carlos, SP
Brazil
E-mail: katiane@icmc.usp.br
suzuki@icmc.usp.br
marinho@icmc.usp.br
louzada@icmc.usp.br