Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Predicting probabilities for the 2010 FIFA World Cup games using a Poisson-Gamma model

Leonardo Soares Bastos [a b] & Joel Mauricio Correa da Rosa [a c]
[a] Department of Statistics , Fluminense Federal University , Niteroi , RJ , Brazil
[b] Scientific Computing Program , Oswaldo Cruz Foundation , Rio de Janeiro , RJ , Brazil
[c] Laboratory of Investigative Dermatology , Rockefeller University , New York , NY , USA
Published online: 11 Apr 2013.

To cite this article: Leonardo Soares Bastos & Joel Mauricio Correa da Rosa (2013) Predicting probabilities for the 2010 FIFA World Cup games using a Poisson-Gamma model, Journal of Applied Statistics, 40:7, 1533-1544, DOI: 10.1080/02664763.2013.788619

To link to this article: http://dx.doi.org/10.1080/02664763.2013.788619

Journal of Applied Statistics, 2013
Vol. 40, No. 7, 1533–1544, http://dx.doi.org/10.1080/02664763.2013.788619

Taylor & Francis
Taylor & Francis Group

Predicting probabilities for the 2010 FIFA World Cup games using a Poisson-Gamma model

Leonardo Soares Bastos[a,b]* and Joel Mauricio Correa da Rosa[a,c]

[a]*Department of Statistics, Fluminense Federal University, Niteroi, RJ, Brazil;* [b]*Scientific Computing Program, Oswaldo Cruz Foundation, Rio de Janeiro, RJ, Brazil;* [c]*Laboratory of Investigative Dermatology, Rockefeller University, New York, NY, USA*

(*Received 18 October 2012; final version received 19 March 2013*)

In this paper, we provide probabilistic predictions for soccer games of the 2010 FIFA World Cup modelling the number of goals scored in a game by each team. We use a Poisson distribution for the number of goals for each team in a game, where the scoring rate is considered unknown. We use a Gamma distribution for the scoring rate and the Gamma parameters are chosen using historical data and difference among teams defined by a strength factor for each team. The strength factor is a measure of discrimination among the national teams obtained from their memberships to fuzzy clusters. The clusters are obtained with the use of the Fuzzy C-means algorithm applied to a vector of variables, most of them available on the official FIFA website. Static and dynamic models were used to predict the World Cup outcomes and the performance of our predictions was evaluated using two comparison methods.

Keywords: Poisson-Gamma distribution; Fuzzy C-means clustering; dynamic model; World Cup predictions; Bayesian inference

1. Introduction

Football, or soccer, is the most popular sport in the world. There are currently 208 country members of the main football association, the *Fédération Internationale de Football Association* (FIFA). There are more country members in FIFA than the United Nations (UN). There are 193 countries UN member states. The FIFA World Cup took place in South Africa from 11 June to 11 July 2010, provided an opportunity to examine statistical methods to predict football outcomes. A review of statistical and related methods and models that have been employed in studies of soccer is presented [2].

We would like to emphasize some predictive methods used to predict football outcomes. An independent Poisson model was fitted to English league and cup soccer data from 1992 to 1995 [6]. A Bayesian dynamic generalized linear model was developed to estimate the time-dependent

*Corresponding author. Email: lbastos@est.uff.br

skills of all teams in a league, and to predict next week-end's soccer matches [9]. Another full dynamic Bayesian model was proposed [5], where the authors refine the independent model [6]. The use of a Poisson process in the number of goals in a World Cup game was illustrated by Chu [4]. A fully subjective Poisson model for the 2006 FIFA World Cup games was developed, where the priors were elicited from soccer experts [11]. A Bayesian negative binomial model for the number of goals scored by players in the Spanish football league was fitted by Sáez Castillo *et al.* [10].

We present a Bayesian Poisson-Gamma model where the priors are chosen using historical information. We decided to build clusters of national teams by taking advantages from recent and historical statistics from the national teams. In order to deal with the inherent uncertainty about soccer games outcomes, we decided to use the fuzzy clustering that allows clusters to overlap [1]. The fuzzy clustering method also provides a measure that helps to build a strength factor used in the prior information for the number of goals. In Section 2, we describe the static and dynamic predictive models for the number of goals. In Section 3, we present the fuzzy clustering method and how it is associated with the strength factor. In Section 4, we present the probabilistic forecast of the 2010 FIFA World Cup games, and compare them with the actual results of the World Cup games. In Section 5, we present the conclusion and future work.

2. Predictive model based on the number of goals

In soccer games, when team A plays against team B, denoted by $A \times B$, the winner of this game is the team that scores more goals. If the number of goals is the same, we call the result a draw. Our model is based on modelling the uncertainty on number of goals scored by each team.

Let N_A be the number of goals scored by team A in the game $A \times B$, and analogously N_B be the number of goals scored by team B. If the joint distribution of the number of goals of both teams is known, the probabilities of the possible outcomes of the game team A against team B can be computed as follows:

$$\begin{aligned} P(\text{`Team A wins'}) &= P(N_A > N_B) \\ &= \sum_{a=1}^{\infty} \sum_{b=0}^{a} P(N_A = a, N_B = b), \end{aligned} \tag{1}$$

$$\begin{aligned} P(\text{`Draw'}) &= P(N_A = N_B) \\ &= \sum_{z=0}^{\infty} P(N_A = z, N_B = z), \end{aligned} \tag{2}$$

$$P(\text{`Team B wins'}) = 1 - P(\text{`Team A wins'}) - P(\text{`Draw'}). \tag{3}$$

2.1 *Static predictive model*

We shall assume that in a game $A \times B$, the number of goals scored by team A depends on a scoring rate λ_a, which is the expected number of goals of team A in the game. Therefore, when two teams play against each other, if we know their scoring rates then we assume that the number of goals scored by each team is conditionally independent, i.e.

$$p(n_a, n_b \mid \lambda_a, \lambda_b) = p(n_a \mid \lambda_a, \lambda_b) p(n_b \mid \lambda_a, \lambda_b). \tag{4}$$

The scoring rates associated with the game $A \times B$, λ_a and λ_b, are unknown. Therefore, under a Bayesian perspective we can set a probability distribution for the scoring rates using some prior

information D_0, $p(\lambda_a, \lambda_b \mid D_0)$. Hence, the predictive distribution of the number of goals of both teams conditioned on the initial information is obtained by integrating out the scoring rates

$$p(n_a, n_b \mid D_0) = \int_0^\infty \int_0^\infty p(n_a, n_b \mid \lambda_a, \lambda_b) p(\lambda_a, \lambda_b \mid D_0)\, \mathrm{d}\lambda_a\, \mathrm{d}\lambda_b. \tag{5}$$

We assume that the number of goals scored by team A on team B depends only on the scoring rate of team A, therefore $p(n_a \mid \lambda_a, \lambda_b) = p(n_a \mid \lambda_a)$ and, analogously, $p(n_b \mid \lambda_a, \lambda_b) = p(n_b \mid \lambda_b)$. Chu [4], from data evidence, suggests that probability distribution of the number of goals scored by team K, $K = \{A, B\}$, can be represented by a Poisson distribution with scoring rate given by λ_k with the probability function given by

$$p(n_k \mid \lambda_k) = \frac{\lambda_k^{n_k} \mathrm{e}^{\lambda_k}}{n_k!}. \tag{6}$$

If our uncertainty about the scoring rates of team A and B can be described by two independent distributions, then the predictive model (5) reduces to a product of predictive marginal distributions

$$p(n_a, n_b \mid D_0) = p(n_a \mid D_0) p(n_b \mid D_0). \tag{7}$$

The uncertainty about the scoring rate of team K is represented in the prior distribution $p(\lambda_k \mid D_0)$. Assuming a Gamma distribution with parameters $\alpha_k^{(0)}$ and $\beta_k^{(0)}$ for λ_k, we obtain a negative binomial predictive distribution for the number of goals of team K, i.e.

$$N_K \mid D_0 \sim \mathrm{NegBin}\left(\alpha_k^{(0)}, \frac{\beta_k^{(0)}}{\beta_k^{(0)} + 1}\right), \tag{8}$$

where the target number of successes is $\alpha_k^{(0)}$ and probability of success is given by $\beta_k^{(0)}/(\beta_k^{(0)} + 1)$.

Note that the joint predictive distribution of (N_A, N_B), Equation (7), depends on the prior parameters $\alpha_a^{(0)}$, $\beta_a^{(0)}$, $\alpha_b^{(0)}$ and $\beta_b^{(0)}$.

2.1.1 *Eliciting the prior distributions*

In a game team $A \times B$, it is not a simple task to elicit values for $\alpha_a^{(0)}$, $\beta_a^{(0)}$, $\alpha_b^{(0)}$ and $\beta_b^{(0)}$. However, we can set reasonable values for the mean and variance of the scoring rates λ_a and λ_b using some prior information D_0. In other words, if λ_a follows a Gamma distribution with parameters $\alpha_a^{(0)}$ and $\beta_a^{(0)}$, its mean and variance are, respectively, given by

$$\mathbb{E}(\lambda_a) = \frac{\alpha_a^{(0)}}{\beta_a^{(0)}} \quad \text{and} \quad \mathbb{V}\mathrm{ar}(\lambda_a) = \frac{\alpha_a^{(0)}}{(\beta_a^{(0)})^2}.$$

Therefore, if we can set values $m_a^{(0)}$ and $v_a^{(0)}$ for the mean and variance of the scoring rates, respectively. Hence, we can obtain the following values for $\alpha_a^{(0)}$ and $\beta_a^{(0)}$:

$$\alpha_a^{(0)} = \frac{(m_a^{(0)})^2}{v_b^{(0)}}, \quad \beta_a^{(0)} = \frac{m_a^{(0)}}{v_a^{(0)}}. \tag{9}$$

The parameters $m_a^{(0)}$ and $v_a^{(0)}$ can be elicited by using, for instance, historical data and FIFA official statistics for team A. Suzuki *et al.* [11] build informative priors for the number of goals based on soccer experts' opinion.

Table 1. Hypothetical statistics for the number of goals scored by teams A and B playing against each other.

	Sample mean	Standard deviation
Team A	3.0	1.0
Team B	1.0	1.0

2.1.2 *Example*

Suppose team A will play against team B, and the initial information suggests that on average team A scores more goals than team B with the same variability, suggesting that team A is a more efficient team. Table 1 was built based on the initial information D_0.

If the initial information, D_0, is our only source of information, then from Equation (9) we obtain that the prior parameters of team A are $\alpha_a^{(0)} = 9.0$, $\beta_a^{(0)} = 3.0$, and for team B, $\alpha_b^{(0)} = 1.0$, and $\beta_b^{(0)} = 1.0$. The joint distribution of the results of the match is given by Equations (7) and (8). According to the joint predictive distribution, the most likely result for this hypothetical match is Team A 2 $\times$ 0 Team B. The probabilities of the outcomes of the game are calculated using Equations (1)–(3), the probability that team A wins the game is 0.7503, the probability of team B wins the game is 0.1249 and the probability of a draw is 0.1249.

2.2 *Dynamic model*

One advantage of using independent scoring rates is that we can dynamically update the distribution of the number of goals of each team during a tournament. Let D_t represent our information for all teams until the round t, where D_0 represents our initial information for the teams. After the first round, all teams have played and we have the numbers of goals scored by each team, $N_K^{(1)}$, $K = \{A, B, C, \ldots\}$. Then the information is updated as follows $D_1 = \{D_0, N_A^{(1)} = n_A^{(1)}, N_B^{(1)} = n_B^{(1)}, N_C^{(1)} = n_C^{(1)}, \ldots\}$. Analogously, on round t, the accumulated information is given by $D_t = \{D_{t-1}, N_A^{(t-1)} = n_A^{(t-1)}, N_B^{(t-1)} = n_B^{(t-1)}, \ldots\}$, where $N_K^{(t-1)}$ is the number of goals scored by team K on round $(t-1)$.

In round t, the joint probability of the number of goals scored by teams A and B is given by

$$p(n_a^{(t)}, n_b^{(t)} \mid D_{t-1}) = \int_0^\infty \int_0^\infty p(n_a^{(t)}, n_b^{(t)} \mid \lambda_a, \lambda_b, D_{t-1}) p(\lambda_a, \lambda_b \mid D_{t-1}) \, \mathrm{d}\lambda_a \, \mathrm{d}\lambda_b. \tag{10}$$

Assuming that the number of goals are conditionally independent given their respective scoring rates, and that the scoring rates are independent, Equation (10) reduces to

$$p(n_a^{(t)}, n_b^{(t)} \mid D_{t-1}) = p(n_a^{(t)} \mid D_{t-1}) p(n_b^{(t)} \mid D_{t-1}),$$

where

$$p(n_k^{(t)} \mid D_{t-1}) = \int_0^\infty p(n_k^{(t)} \mid \lambda_k) p(\lambda_k \mid D_{t-1}) \, \mathrm{d}\lambda_k, \quad K = \{A, B\}.$$

The distribution of the number of goal of team K conditional on the scoring rate λ_k is given in Equation (6), and the distribution of the scoring rate of team K, λ_k, conditional on the information D_{t-1} is a Gamma distribution with parameters $\alpha_k^{(t-1)} = \alpha_k^{(0)} + \sum_{i=1}^{t-1} n_k^{(i)}$ and $\beta_k^{(t-1)} = \beta_k^{(0)} + (t-1)$.

Therefore, for round t after combining a Poisson model for the number of goals of team K and the updated prior distribution for the scoring rate of team K, λ_k, the predictive distribution is the

following negative binomial distribution:

$$N_K^{(t)} \mid D_{t-1} \sim \text{NegBin}\left(\alpha_k^{(t-1)}, \frac{\beta_k^{(t-1)}}{\beta_k^{(t-1)}+1}\right). \tag{11}$$

3. The strength factor

In order to obtain the probabilistic forecasts for the 2010 FIFA World Cup games, we model the number of goals of team in each match using the negative binomial distributions (8) and (11), and the probabilities of the possible outcomes of the game are given in Equations (1) and (2).

However, the parameters $\alpha_k^{(0)}$ and $\beta_k^{(0)}$ for each team have to be elicited. Our initial thought was to learn about the parameters using the average score of each team while qualifying for the World Cup. However, there are some confederations stronger than others, as we can notice from Figure 1. The teams represented by CONMEBOL, on average, have large values of points of FIFA ranking, whereas the average score is relatively low. On the other hand, New Zealand, the only team from OFC, has a low value of points for the FIFA ranking, but it has a relatively large average score for qualifying for the World Cup. These two examples illustrate that the average scores for qualifying are not comparable between teams from different confederations. Therefore, we avoid using the average score of each team for qualifying to obtain the parameters $\alpha_k^{(0)}$ and $\beta_k^{(0)}$.

We propose a different metric to obtain the parameters $\alpha_k^{(0)}$ and $\beta_k^{(0)}$ for team k. Using Equation (9), we need to set values for $m_k^{(0)}$ and $v_k^{(0)}$. We assume that the variance $v_k^{(0)}$ is the same for all teams, and is set as the sample variance of the average scores of all qualified teams for qualifying for the World Cup, S_y^2. The mean $m_k^{(0)}$ must depend on the team. We propose to use the sample mean of the average scores of all qualified teams, $\bar{y}$, weighted by a strength factor, δ_k, which varies from team to team. Therefore,

$$m_k^{(0)} = \bar{y}\frac{\delta_k}{\delta_{(32)}}, \quad v_k = S_y^2, \quad k = 1, 2, \ldots, 32, \tag{12}$$

where $\delta_{(32)}$ is the maximum strength factor among the 32 national teams. The values for $\bar{y}$ and S_y^2 were calculated from the data available on the official FIFA website in June 2010, $\bar{y} = 1.85$ and $S_y^2 = 0.2529$. Note that $m_k^{(0)} \leq \bar{y}$, the reason for this assumption is that the average scores during the qualifying tend to be larger than the average during the World Cup. For instance, the average

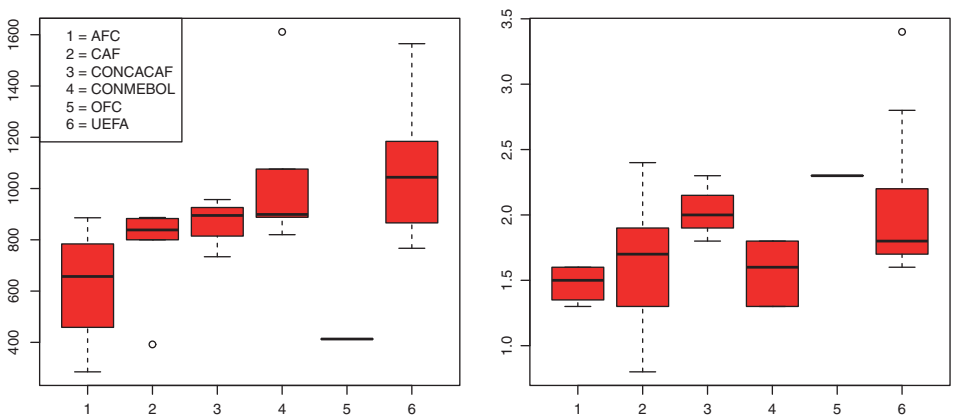

Figure 1. Boxplots for all qualified teams by confederation of (a) the number of points in FIFA ranking in June 2010 and (b) the average scores of each team during the qualifies for the 2010 World Cup.

Table 2. Variables used to build the strength factor of each national team.

Variable	Description
X_1	Total points obtained in World Cups
X_2	Lowest rank position obtained since the creation of FIFA ranking
X_3	Range between the lowest and highest rank
X_4	Number of points in FIFA ranking in June 2010
X_5	Number of players who participated in 2009-20-10 Champions League Quarter-finals
X_6	Performance in World Cup Qualifying games (% of points)

scores of all qualifying games for the FIFA World Cup 2006 were 2.91 goals, whereas the average scores for the FIFA World Cup 2006 were 2.3 goals.

The strength factor, δ_k, is calculated using some variables that we believe to influence the strength of each team. These variables are presented in Table 2. The first three variables reflect the national team tradition and the others reflect the team performance close to the FIFA World Cup 2010. The fifth variable makes it possible to include the individual quality of the players by considering their participation in the most important football tournament. The data set for each national team was collected from the official FIFA website.

The strength factor is normalized and its use serves as a kind of penalty to national teams that scored a large number of goals during the qualifying stage but have played against considerably weaker opponents. The fact that these national teams are more likely to play against weaker national teams suggests a smaller value for the strength factor reducing therefore the prior expected number of goals. For instance, this is the case of the USA and New Zealand. They have qualified to the World Cup playing on confederations whose technical quality is not at the same level as Europe and South America.

3.1 *Fuzzy C-means clustering*

Since we need to discriminate the national teams according to their strengths, and this information is highly subjective and imprecise, we decided to use some elements from fuzzy set theory [13]. Thus, we decided to build fuzzy clusters of national teams according to their similarities in the set of variables described in Table 2 and then we explored the membership degree as a measure of discrimination among the teams. The chosen algorithm used to group the national teams was the Fuzzy C-means [1].

The Fuzzy C-means algorithm yields as a result of the clustering procedure pairs $(g, u_g(\mathbf{x}_k))$, in which g represents the gth cluster, $u_g(\mathbf{x}_k)$ is a membership degree of the kth observation to the gth cluster and $\mathbf{x}_k \subset \mathbb{R}^p$ are covariates of the kth observation, for $k = 1, 2, \ldots, K$

$$J_m = \sum_{k=1}^{K} \sum_{g=1}^{G} [u_g(\mathbf{x}_k)]^m \|x_k - \mathbf{v}_g\|^2. \tag{13}$$

The degree of fuzziness is determined by $m \geq 1$ and the particular case in which $m = 1$ the algorithm creates the so-called hard partition where the entities does or does not belong to a cluster.

The algorithm is based on the minimization of the c-means objective function in Equation (13), where $\mathbf{v}_g$ is the cluster centroid, also called prototype. With the restriction $\sum_{g=1}^{G} u_g(\mathbf{x}_k) = 1 \ \forall k$, to minimize Equation (13) represents a constrained and nonlinear problem of optimization. If there are at least G different $\mathbf{x}_k$ points, J_m is minimized only if

$$u_g(\mathbf{x}_k) = \frac{1}{\sum_{l=1}^{G} (\|\mathbf{x}_k - \mathbf{v}_g\| / \|\mathbf{x}_k - \mathbf{v}_l\|)^{2/(m+1)}} \tag{14}$$

Table 3. Clusters obtained using the Fuzzy C-means clustering with $m = 1.25$ and $G = 4$.

	National teams
Cluster 1	South Africa, Algeria, Korea DPR, Slovakia, Slovenia, Honduras, New Zealand, Paraguay
Cluster 2	Germany, Spain, Holland, England, Italy
Cluster 3	Argentina, Brazil, France, Portugal
Cluster 4	Australia, Cameron, Chile, Korea Republic, Ivory Coast, Denmark, the USA, Ghana, Greece, Japan, Mexico, Nigeria, Serbia, Switzerland, Uruguay

and

$$\mathbf{v}_g = \frac{\sum_{k=1}^{K} \mathbf{x}_k [u_g(\mathbf{x}_k)]^m}{\sum_{k=1}^{K} [u_g(\mathbf{x}_k)]^m}. \tag{15}$$

For more details, see [8].

The fuzziness degree, m, and the number of clusters, G, are unknown quantities. Although most applications suggest $m = 2$ as a good compromise between hard and very fuzzy clustering. A theoretical basis for selecting this parameter was developed [12]. The authors have found an upper bound for the fuzziness degree given by the following rule:

$$m \leq \frac{\min(p, K-1)}{\min(p, K-1) - 2} \quad \text{if } \min(p, K-1) \geq 3. \tag{16}$$

In our application, we have $K = 32$ national teams and $p = 6$ covariates; therefore, the upper bound for the fuzziness degree is 1.5. Since lower bound for the fuzziness degree is 1, we are going to set the fuzziness degree as the middle point of the interval 1.25. After, cluster stability measures: silhouette and connectivity [3], have indicated 4 as the number of clusters.

3.2 *National teams clustering*

We have used the Fuzzy C-means clustering algorithm implemented in the R function `fanny` and the four clusters are presented in Table 3. Cluster 2 is the one containing the so-called favourite teams to win the World Cup, which is therefore the strongest group.

The strength factor δ_k of the kth team is a weighted mean of the degree of membership with weights given by the average of FIFA ranking points for each cluster. That is,

$$\delta_k = \sum_{j=1}^{4} w_j u_j(\mathbf{x}_k), \tag{17}$$

where the weights w_j are the average of FIFA ranking points for each cluster. Note that w_j is a way of summarizing the information about the jth cluster prototype and, as commented in the previous section, all data have influence on all prototypes. Thus, it is important to emphasize that the strength factor for a national team somehow takes into account information about all opponents. Figure 2 presents the boxplot of the number of points in FIFA ranking by cluster.

4. Probabilities for World Cup 2010 games

The 2010 FIFA World Cup took place in South Africa from 11 June to 11 July 2010. We predict the outcome of each game using models presented in Section 2. The predictive joint distribution of the number of goals by each team in a World Cup game is derived by using the static model (5) and the dynamic model (10).

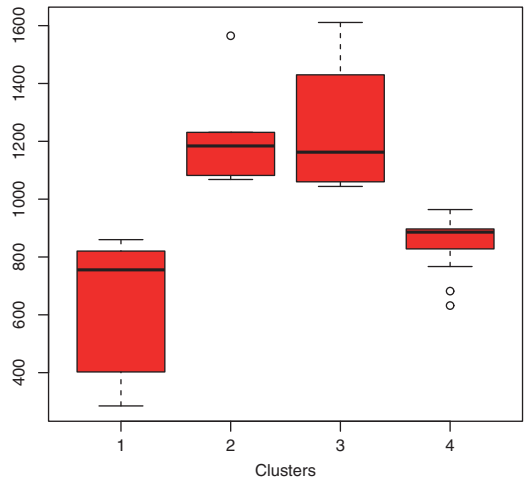

Figure 2. Boxplot of the number of points in FIFA ranking in June 2010 by cluster obtained using the Fuzzy C-means clustering with $m = 1.25$.

Notice that a match with a large number of goals, for instance more than 10 goals per team, is unexpected. In fact, the highest number of goals scored by a team in World Cup games is 10 (Hungary (10) versus El Salvador (1), 1982). Therefore, the probabilities (1) and (2) can be approximated by

$$P(\text{'Team A wins'}) \approx \sum_{a=1}^{M} \sum_{b=0}^{a} P(N_A = a, N_B = b), \tag{18}$$

$$P(\text{'Draw'}) \approx \sum_{z=0}^{M} P(N_A = z, N_B = z), \tag{19}$$

where $M = 20$ provides a good approximation for these probabilities.

Using the strength factor, Equation (17), we obtain the parameters for the negative binomial distribution (8) using the mean and variance elicited in Equation (12). Therefore, we can calculate the outcome probabilities (1) and (2). As an illustration, the predictions for the first World Cup game were South Africa would win with probability 0.2733, Mexico would win with probability 0.4552, and obviously the probability of a draw is given by the complementary 0.2715. Note that the predictions for the static and the dynamic models for the first round games must be the same. The actual result was South Africa 1×1 Mexico. Our predictions for all World Cup games are presented in the appendix.

4.1 *Comparison*

In this section, we evaluate the performance of our predictive models comparing the probability predictions with the final 2010 World Cup results. In case of a draw in the knockout stage, we considered the result as a draw regardless of the result of the penalty shootout. The first comparison method is simply to compute the proportion of correctly predicted games. A correct prediction happens if actual outcome (draw, A wins or B wins) of the game is the one with the highest predictive probability of happening.

The performance of the static and dynamic models was measured by the proportion of correct predictions. For the group stage, the correct predictions for both models are the same: 0.4792.

Table 4. DeFinetti measure for the static and dynamic predictive models by the World Cup stage.

	Static	Dynamic
Group stage	0.6258	0.6323
Knockout stage	0.5668	0.5768
Overall	0.6110	0.6185

On the knockout stage, the static model seems to perform better as the computed proportions are 0.5 and 0.4375 for the static and dynamic models, respectively. The overall performance of the static and dynamic models is 0.4844 and 0.4688, respectively. If we compare the predictions with a naive model by attributing the same chance of outcome for every game, our predictive models are clearly better since in the naive approach only a third of the outcomes are expected to be right.

Another performance measure is the DeFinetti measure [7]. Let the vector $x = (x_A, x_{\text{Draw}}, x_B)$ represent the outcome of the game team $A \times B$. If team A wins $x = (1, 0, 0)$, if team B wins $x = (0, 0, 1)$ and if the outcome is a draw $x = (0, 1, 0)$. The DeFinetti measure is defined by the Euclidean distance between x and the outcome probabilities of the game. The range of the DeFinetti measure is 0 and 2, where 0 corresponds to a perfect prediction and 2 to a completely wrong prediction.

The average of the DeFinetti measure by World Cup stage for the static and dynamic predictive models is presented in Table 4. The static model performs slightly better. And both models are better than the naive model whose DeFinetti measure would be approximately 0.6667. On the knockout stage, both models improved with respect to the group stage. Both comparison measures suggest that the static model performed better for predicting the 2010 FIFA World Cup results than the dynamic model. However, the probabilities were similar on both modelling approaches. Each national team can play at most eight games in a FIFA World Cup, so we believe that the dynamic modelling would outperform the static modelling for tournaments which have a larger number of games, for instance in the 2011 Brazilian football championship there were 20 teams playing against each other (home and away games) resulting in 38 games for each team.

5. Discussion

We presented a static and a dynamic Poisson-Gamma model to predict the outcome of football results based on the number of goals scored by each team. The prior distributions for the Poisson-Gamma model depend on a set of historical data from which we build the strength factor using a Fuzzy c-means clustering method. This automatic procedure for building priors can be equivalent to prior distributions based on the expert's opinions used on the fully subjective model [11]. For instance, if the experts build their predictions based on historical data, our automatic procedure can lead to similar priors.

Both models are better than a naive approach, which sets equal probability to all possible outcomes. And the static model performs slightly better than the dynamic model for the FIFA World Cup results. We believe that dynamic models would outperform static models for larger tournaments where the performance of the teams in the competition may also depend on time-dependent variables such as injuries or coach stability.

We are aware that the independence assumption between the two teams may be a strong assumption. For instance, the number of goals of team A could depend on the defence power or the tactic adopted by team B. Therefore, we should improve our predictions including variables associated with the tactic, attack and defence power of each national team. Another extension should be using a bivariate distribution for each game, where a correlation parameter could be induced into

the modelling. But it is important to remember that the strength factor is a measure that takes into account information about all the opponents and by doing that we can indirectly reduce the effects of assuming independence between any two teams.

A natural extension of our model includes the fuzzy clustering using time-dependent variables. In this case in every time change, the strength factor is updated as well as the clusters and priors leading to improved predictions for the probabilities of the game outcomes. Another extension is to consider a strength factor that is also a function of the opponent, this would be $\delta_{k,l}$ in which k represents the team itself and l its opponent. This procedure might be simpler than trying to model a joint distribution for the number of goals including a correlation parameter.

References

[1] J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.

[2] D. Brillinger, *Soccer/world football*, Tech. Rep. 777, University of California, Berkley, CA, 2009.

[3] G. Brock, V. Pihur, S. Datta, and S. Datta, *clvalid: An r package for cluster validation*, J. Stat. Softw. 25 (2008), pp. 1–22. Available at http://www.jstatsoft.org/v25/i04.

[4] S. Chu, *Using soccer goals to motivate the poisson process*, INFORMS Trans. Education 3 (2003), pp. 62–68. Available at http://ite.pubs.informs.org/Vol3No2/Chu/.

[5] M. Crowder, M. Dixon, A. Ledford, and M. Robinson, *Dynamic modelling and prediction of English football league matches for betting*, J. R. Stat. Soc. Ser. D 51(2) (2002), pp. 157–168.

[6] M.J. Dixon and S.G. Coles, *Modelling association football scores and inefficiencies in the football betting market*, J. R. Stat. Soc. Ser. C 46 (1997), pp. 265–280.

[7] B. de Finetti, *Probability, Induction and Statistics: The Art of Guessing*, Wiley, London, 1972.

[8] N.R. Pal, K. Pal, J.M. Keller, and J.C. Bezdek, *A possibilistic fuzzy c-means clustering algorithm*, IEEE Trans. Fuzzy Syst. 13 (2005), pp. 517–530.

[9] H. Rue and O. Salvesen, *Prediction and retrospective analysis of soccer matches in a league*, J. R. Stat. Soc. Ser. D 49 (2000), pp. 399–418.

[10] A. Sáez Castillo, J. Rodríguez Avi, and J.M. Pérez Sánchez, *Expected number of goals depending on intrinsic and extrinsic factors of a football player. An application to professional spanish football league*, Eur. J. Sport Sci. 13(2) (2011), pp. 127–138.

[11] A.K. Suzuki, L.E.B. Salasar, J.G. Leite, and F. Louzada-Neto, *A Bayesian approach for predicting match outcomes: The 2006 (association) football world cup*, J. Oper. Res. Soc. (2009), pp. 1–10.

[12] J. Yu, Q. Cheng, and H. Huang, *Analysis of the weighting exponent in the fcm*, IEEE Trans. Syst. Man Cybern. Part B Cybern. 34(1) (2004), pp. 634–639.

[13] L. Zadeh, *Fuzzy sets*, Inform. Control 8 (1965), pp. 338–352.

Appendix

Tables A1 – A4 present our predictions for the 2010 FIFA World Cup games separated by round using the static and dynamic models. In all tables, the probability of a draw can be computed by complementary probability.

Table A1. Probabilistic predictions for first round of games on the 2010 World Cup games using the static (S) and the dynamic (D) models.

	A wins (S)	*B* wins	*A* wins (D)	*B* wins	True result
SouthAfrica × Mexico	0.2733	0.4552	0.2733	0.4552	1 × 1
Uruguay × France	0.2977	0.4830	0.2977	0.4830	0 × 0
Argentina × Nigeria	0.5094	0.2702	0.5094	0.2702	1 × 0
KoreaRepublic × Greece	0.3704	0.3693	0.3704	0.3693	2 × 0
England × USA	0.5006	0.2779	0.5006	0.2779	1 × 1
Algeria × Slovenia	0.3525	0.3352	0.3525	0.3352	0 × 1
Germany × Australia	0.5149	0.2631	0.5149	0.2631	4 × 0
Serbia × Ghana	0.3882	0.3543	0.3882	0.3543	0 × 1
Holland × Denmark	0.5022	0.2761	0.5022	0.2761	2 × 0
Japan × Cameron	0.3684	0.3712	0.3684	0.3712	1 × 0
Italy × Paraguay	0.5712	0.2074	0.5712	0.2074	1 × 1
NewZealand × Slovakia	0.3477	0.3390	0.3477	0.3390	1 × 1
IvoryCoast × Portugal	0.2769	0.5018	0.2769	0.5018	0 × 0
Brazil × Korea DPR	0.5763	0.2037	0.5763	0.2037	2 × 1
Honduras × Chile	0.3139	0.4014	0.3139	0.4014	0 × 1
Spain × Switzerland	0.5097	0.2682	0.5097	0.2682	0 × 1

Table A2. Probabilistic predictions for second round of games on the 2010 World Cup games using the static (S) and the dynamic (D) models.

	A wins (S)	*B* wins	*A* wins (D)	*B* wins	True result
SouthAfrica × Uruguay	0.2665	0.4666	0.3033	0.4068	0 × 3
France × Mexico	0.4918	0.2883	0.4529	0.3101	0 × 2
Greece × Nigeria	0.3712	0.3679	0.3545	0.3511	2 × 1
Argentina × KoreaRepublic	0.5074	0.2724	0.4603	0.3138	4 × 1
Germany × Serbia	0.4916	0.2875	0.5913	0.2063	0 × 1
Slovenia × USA	0.2811	0.4408	0.2913	0.4257	2 × 2
England × Algeria	0.5673	0.2112	0.5920	0.1764	0 × 0
Ghana × Australia	0.3755	0.3595	0.4058	0.3109	1 × 1
Holland × Japan	0.5040	0.2742	0.5178	0.2606	1 × 0
Cameron × Denmark	0.3696	0.3706	0.3529	0.3540	1 × 2
Slovakia × Paraguay	0.3399	0.3462	0.3416	0.3466	0 × 2
Italy × NewZealand	0.5701	0.2085	0.5465	0.2213	1 × 1
Brazil × IvoryCoast	0.5042	0.2752	0.5536	0.2252	3 × 1
Portugal × Korea DPR	0.5739	0.2050	0.5215	0.2345	7 × 0
Chile × Switzerland	0.3548	0.3773	0.3532	0.3728	1 × 0
Spain × Honduras	0.5609	0.2170	0.5572	0.1965	2 × 0

Table A3. Probabilistic predictions for third round of games on the 2010 World Cup games using the static (S) and the dynamic (D) models.

	A wins (S)	*B* wins	*A* wins (D)	*B* wins	True result
Mexico × Uruguay	0.3662	0.3850	0.3701	0.3835	0 × 1
France × South Africa	0.5771	0.2031	0.5209	0.2137	1 × 2
Nigeria × Korea Republic	0.3674	0.3718	0.2981	0.4323	2 × 2
Greece × Argentina	0.2718	0.5079	0.2342	0.5540	0 × 2
Slovenia × England	0.2037	0.5752	0.2864	0.4664	0 × 1
USA × Algeria	0.4332	0.2905	0.5278	0.1884	1 × 0
Ghana × Germany	0.2709	0.5074	0.2479	0.5310	0 × 1
Australia × Serbia	0.3454	0.3951	0.3299	0.3755	2 × 1
Slovakia × Italy	0.2047	0.5741	0.1878	0.5709	3 × 2
Paraguay × New Zealand	0.3428	0.3452	0.3943	0.3140	0 × 0
Denmark × Japan	0.3717	0.3680	0.3939	0.3237	1 × 3
Cameron × Holland	0.2756	0.5027	0.2345	0.5345	1 × 2
Portugal × Brazil	0.3932	0.3975	0.4402	0.3648	0 × 0
Korea DPR × Ivory Coast	0.2837	0.4382	0.2740	0.4094	0 × 3
Chile × Spain	0.2574	0.5202	0.2689	0.4918	1 × 2
Switzerland × Honduras	0.4151	0.3052	0.4271	0.2399	0 × 0

Table A4. Probabilistic predictions for knockout stage games on the 2010 World Cup games using the static (S) and the dynamic (D) models.

	A wins (S)	*B* wins	*A* wins (D)	*B* wins	Phase[a]	True result
Uruguay × Korea Republic	0.3989	0.3485	0.3651	0.3872	4	2 × 1
USA × Ghana	0.3770	0.3625	0.4263	0.2999	4	1 × 2
Germany × England	0.3951	0.3952	0.4475	0.3287	4	4 × 1
Argentina × Mexico	0.4925	0.2878	0.5487	0.2388	4	3 × 1
Holland × Slovakia	0.5737	0.2049	0.5258	0.2459	4	2 × 1
Brazil × Chile	0.5238	0.2551	0.5287	0.2433	4	3 × 0
Paraguay × Japan	0.2887	0.4320	0.2874	0.4354	4	0(5) × 0(3)
Spain × Portugal	0.3941	0.3961	0.3380	0.4523	4	1 × 0
Holland × Brazil	0.2031	0.5769	0.2259	0.5552	5	2 × 1
Uruguay × Ghana	0.4027	0.3438	0.4618	0.2762	5	1(4) × 1(2)
Argentina × Germany	0.3989	0.3920	0.4184	0.3818	5	0 × 4
Paraguay × Spain	0.2072	0.5717	0.1821	0.5922	5	0 × 1
Uruguay × Holland	0.4670	0.2661	0.4135	0.3286	6	2 × 3
Germany × Spain	0.3949	0.3954	0.4757	0.3222	6	0 × 1
Uruguay × Germany	0.3000	0.4797	0.2864	0.5002	7	2 × 3
Holland × Spain	0.2044	0.5745	0.3220	0.4456	7	0 × 1

Note: [a]Phase 4 represents the round of 16, phase 5 represents the quarter finals, phase 6 represents the semifinal and phase 7 represents the final and third place game.