International Journal of Scientific Knowledge
(Computing and Information Technology)
Volume 1 Issue 2, Oct 2012

ISSN 2305-1493 www.ijsk.org

A Bayesian Computation for the Prediction of Football Match Results Using Artificial Neural Network

Mehmet Ali Cengiz[1]
Naci Murat[2]
Haydar Koç[3]
Talat Şenel[4]

19 Mayıs University[1, 2, 3, 4]
Turkey

,

Email: macengiz@omu.edu.tr, nacimurat@omu.edu.tr, haydar.koc@omu.edu.tr, tlsenel@omu.edu.tr

Abstract. The problem of modelling football data has become increasingly popular in the last few years and many different models have been proposed with the aim of estimating the characteristics that bring a team to lose or win a game, or to predict the score of a particular match. We propose a Bayesian model to test its predictive strength on data for the Turkish Super League 2008-2009. We compare the predictive performance of Bayesian approach with a Poisson Log linear approach and Artificial Neural Network approaches.

Keywords: Artificial Neural Network, Bayesian Model, Poisson Log linear

1 **Introduction**

Statistical soccer prediction is a method used in sports betting, to predict the outcome of soccer matches by means of statistical tools. The problem of modelling football [soccer] data has become increasingly popular in the last few years and many different models have been proposed with the aim of estimating the characteristics that bring a team to lose or win a game, or to predict the score of a particular match.

First statistical model for soccer predictions was proposed by Moroney M J [8]. He showed that both Poisson distribution and negative binomial distribution provided an adequate fit to results of soccer games. Reep C and Benjamin B analysed the numbers of ball passing between players during matches using negative binomial distribution. [9] Maher M J proposed the first model predicting outcomes of soccer matches between teams with different skills using the Poisson distribution [7]. He developed a model in which the home and away team scores follow independent Poisson distributions, with means reflecting the attacking and defensive capabilities of the two teams. Dixon M J and Coles S G [3] employ a similar approach as Maher M J [7] to develop a forecasting model capable of generating previous match outcome probabilities. All parameters are estimated using data from previous matches only. In the same way, Rue H and Salvesen O [10] assume that the time-varying attacking and defensive parameters of all teams vary randomly over time.

Caurneya K S and Carron A V [2] proposed the methods for modelling the home field advantage

International Journal of Scientific Knowledge
(Computing and Information Technology)
Volume 1 Issue 2, Oct 2012

ISSN 2305-1493 www.ijsk.org

factor. Karlis D and Ntzoufras I [6] showed empirical low levels of correlation between the numbers of goals scored in a single game by the two opponents. Karlis D and Ntzoufras I [5] advocated the use of a bivariate Poisson distribution that has a more complicated formulation for the likelihood function, and includes an additional parameter explicitly accounting for the covariance between the goals scored by the two competing teams. They specify the model in a frequent framework, and their main purpose is the estimation of the effects used to explain the number of goals scores [1]. Bayesian methods can be used to update the prior estimates of these parameters as new match results information is received, and Markov chain Monte Carlo iterative simulation techniques are used for inference.

In this paper we use the same model in [5], which is a Bayesian approach in WinBUGS [11], which has become the standard software for Bayesian analysis for predicting match outcomes for Turkish super League for the season 2008-2009 and compare the results with a Poisson Loglinear approach and Artificial Neural Network approaches.

2 **Models**

We use the same model in [5]. We first consider Turkish super League for the season 2008-2009. The league is made by a total of T = 18 teams, playing each other twice in a season (one at home and one away). We indicate the number of goals scored by the home and by the away team in the g_{th} game of the season $(g = 1,...,G = 306)$ as y_{g1} and y_{g2} respectively. The vector of observed counts $y = (y_{g1}, y_{g2})$ is modelled as independent Poisson:

$$(y_{gj} \mid \theta_{gj} \sim Poisson(\theta_{gj})) \quad (1)$$

where the parameters $\theta = (\theta_{g1}, \theta_{g2})$ represent the scoring intensity in the g-th game for the team playing at home $(j = 1)$ and away $(j = 2)$, respectively. We model these parameters according to a formulation that has been used widely in the statistical literature [5], [6] assuming a log-linear random effect model:

$$\log\theta_{g1} = \text{home} + \text{att}_{h(g)} + \text{def}_{a(g)}$$
$$\log\theta_{g2} = \text{att}_{a(g)} + \text{def}_{h(g)} \quad (2)$$

The parameter home represents the advantage for the team hosting the game and we assume that this effect is constant for all the teams and throughout the season. In addition, the scoring intensity is determined jointly by the attack and defence ability of the two teams involved, represented by the parameters att and def, respectively. The nested indexes $h(g), a(g) = 1,...,T$ identify the team that is playing at home (away) in the g_{th} game of the season.

The data structure for the model is presented in TABLE I and consist of the name and code of the teams, and the number of goals scored for each game of the season. As is possible to see, the indexes $h(g)$ and $a(g)$ are uniquely associated with one of the 18 teams.

In line with the Bayesian approach, we have to specify some suitable prior distributions for all the random parameters in our model. The variable home is modelled as a fixed effect, assuming a standard flat prior distribution (notice that we use here the typical notation to describe the Normal distribution in terms of the mean and the precision):

$$home \sim Normal(0,0.0001) \quad (3)$$

Conversely, for each $t = 1,...,T$, the team-specific effects are modelled as exchangeable from a common distribution:

$$att_t \sim Normal(\mu_{att}, \tau_{att}) \ ,$$
$$def_t \sim Normal(\mu_{def}, \tau_{def})$$

As suggested by various works, we need to impose some identifiably constraints on the team-specific parameters. In line with [5], we use a sum-to-zero constraint, that is

International Journal of Scientific Knowledge
(Computing and Information Technology)
Volume 1 Issue 2, Oct 2012

ISSN 2305-1493 www.ijsk.org

$\Sigma_{t=1}^{T} att_t = 0$, $\Sigma_{t=1}^{T} def_t = 0.$ (4)

However, we also assessed the performance of the model using a corner constraint instead, in which the team-specific effect for only one team are set to 0, for instance $att_1 = 0$ and $def_1 = 0$. Even if this latter method is slightly faster to run, the interpretation of these coefficients is incremental with respect to the baseline identified by the team associated with an attacking and defending strength of 0 and therefore is less intuitive. Finally, the hyper-priors of the attack and defence effects are modelled independently using again flat prior distributions:

$$\begin{aligned} \mu_{att} &\sim Normal(0,0.000.1) \ , \\ \mu_{def} &\sim Normal(0,0.000.1) \\ \tau_{att} &\sim Gamma(0.1,0.1) \ , \\ \tau_{def} &\sim Gamma(0.1,0.1) \end{aligned} \quad (5)$$

3 **Application**

Posterior summaries of the Poisson log-linear model parameters are provided in TABLE I. As we can see from the estimated model Fenerbahce had the highest attacking parameter while Kayserispor had the lowest (i.e. best) defensive parameter.

Posterior credible intervals of attacking and defensive parameters for each team are depicted in Fig. 1 and 2

Models in sports are used mainly for prediction. Here we briefly illustrate how we can obtain predictions for six future games. The approach can be easily generalized to additional games using the same approach. Posterior probabilities of each predicted outcome are summarized in Table 3. Outcome probabilities indicate Antalyaspor's probability of winning the game against MKE Ankaragücü is about 43%. Concerning the second game (Trabzonspor vs. Fenerbahçe), the posterior model probabilities confirm that the posterior probability of away team is greater than the posterior probability of home team.

In order to compare the predictive accuracy, Poisson loglinear approach, two Artificial Neural Network approaches were also considered. The main concept is to use the number of goals scored by the home and by the away team and being home team effect as input parameters, to predict the number of goals that will be scored by the home and by the away team for future game (output parameter). In our study a feed forward multi layer network architecture (ANN1) and a radial basis function approach (ANN2) were employed which are widely used in ANN applications. For training and testing the proposed ANN model the overall data set was randomly divided into two separate sets. The first set, namely the training set, consisted of 75% data records while the remaining 25% of the data records formed the test set. For comparison, the overall data set was divided as the same way with ANN approach for Bayesian approaches. Table 4 provides RMSE and MAPE values for all models with three different approaches.

4 **Conclusions**

In this study, we give some details for a Bayesian Approach given by Karlis and Ntzoufras (2003) to predict future match scores for Turkish super League [5]. We obtained fairly good posterior model probabilities for unplayed six matches. We also used a Poisson Loglinear approach and two Artificial Neural Network approaches to compare with the Bayesian Poisson Loglinear approach for prediction performance.
Two Artificial Neural Network approaches give the similar the predictive performance. Although Artificial Neural Network approaches give the better the predictive performance than the Poisson loglinear approach, Bayesian approach gives much smaller RMSE and MAPE values than the others.

5 **Refrences**

ISSN 2305-1493 www.ijsk.org

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

2. Caurneya K S and Carron A V (1992). The home advantage in sports competitions: A literature review. *Journal of Sport and Exercise Physiology*, 14, 13-27.

3. Dixon M J and Coles S G (1997). Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics*, 46, 2, 265-280.

4. Karlis D and Ntzoufras I (2000). On modelling soccer data. *Student*, 3, 229–244.

5. Karlis D and Ntzoufras I (2003). Analysis of sports data by using bivariate Poisson models. *J. R. Statist. Soc. D*, 52, 381–393.

6. Karlis D and Ntzoufras I (2009). Bayesian modelling of football outcomes: Using the Skellam's distribution for the goal difference. *IMA J. Numerical Analysis*, 20, 133-145.

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

8. Moroney M J (1956). *Facts from figures*. 3rd edition, Penguin, London.

9. Reep C and Benjamin B (1968). Skill and chance in association football. *Journal of the Royal Statistical Society A*, 131, 581-585.

10. Rue H and Salvesen O (1999). Predicting and retrospective analysis of soccer matches in a league. Technical Report. Norwegian University of Science and Technology, Trondheim.

11. Spiegelhalter D, Thomas A, Best N and Lunn D (2003). WinBUGS Version 1.4 User Manual. MRC Biostatistics Unit..

International Journal of Scientific Knowledge
(Computing and Information Technology)
Volume 1 Issue 2, Oct 2012

ISSN 2305-1493

IJSK

www.ijsk.org

Table 1. Posterior Summaries Of Team Attacking Parameters

Team		Posterior				
					percentiles	
		Node	Mean	Sd	2.5%	97.5%
1.	Ankaraspor a.ş	a[1]	-0.1705	0.162	-0.5065	0.1317
2.	Antalyaspor a.ş	a[2]	-0.2284	0.1664	-0.5652	0.08546
3.	Beşiktaş a.ş	a[3]	0.3302	0.1285	0.07631	0.5769
4.	Bursaspor	a[4]	0.09943	0.1442	-0.1942	0.3695
5.	Büyükşehir Bld.spor	a[5]	-0.1101	0.1593	-0.436	0.1889
6.	Denizlispor	a[6]	-0.07572	0.1552	-0.3973	0.2126
7.	Eskişehirspor	a[7]	0.06513	0.1465	-0.2331	0.3433
8.	Fenerbahçe	a[8]	0.3376	0.1283	0.07977	0.5809
9.	Galatasaray a.ş	a[9]	0.2934	0.1318	0.0289	0.5473
10.	Gaziantepspor	a[10]	0.08419	0.1452	-0.2114	0.3602
11.	Gençlerbiliği	a[11]	-0.09018	0.1575	-0.4103	0.2065
12.	Hacettepespor	a[12]	-0.55	0.2014	-0.9569	-0.1724
13.	Kayserispor	a[13]	-0.2269	0.1686	-0.5744	0.08778
14.	Kocaelispor	a[14]	0.1467	0.1457	-0.1463	0.4279
15.	Konyaspor	a[15]	-0.1934	0.1634	-0.5237	0.1175
16.	Mke Ankaragücü	a[16]	-0.163	0.1611	-0.4907	0.1404
17.	Sivasspr	a[17]	0.2236	0.1338	-0.0488	0.4773
18.	Trabzonspor a.ş	a[18]	0.2309	0.1345	-0.0424	0.4878
		home	0.3073	0.07413	0.1625	0.4523

Table 2. Posterior Summaries Of Team Defensive Parameters

Team		Posterior				
					percentiles	
		Node	Mean	S.d	2.5%	97.5%
1.	Ankaraspor a.ş	d[1]	-0.01591	0.1515	-0.3254	0.27
2.	Antalyaspor a.ş	d[2]	-0.01906	0.1504	-0.3229	0.2675
3.	Beşiktaş a.ş	d[3]	-0.3297	0.1788	-0.6973	0.00548
4.	Bursaspor	d[4]	-0.1291	0.1632	-0.4584	0.1804
5.	Büyükşehir Bld.spor	d[5]	0.08286	0.148	-0.2117	0.3648
6.	Denizlispor	d[6]	0.2056	0.1378	-0.07479	0.4671
7.	Eskişehirspor	d[7]	0.1535	0.1403	-0.133	0.4188
8.	Fenerbahşe	d[8]	-0.14	0.1624	-0.4659	0.165
9.	Galatasaray a.ş	d[9]	-0.06114	0.1543	-0.3751	0.2323
10.	Gaziantepspor	d[10]	0.1286	0.1432	-0.1601	0.3996
11.	Gençlerbiliği	d[11]	0.1008	0.1459	-0.1941	0.3802
12.	Hacettepespor	d[12]	0.3803	0.1251	0.1287	0.6181
13.	Kayserispor	d[13]	***-0.4748***	0.1903	-0.8571	-0.1193
14.	Kocaelispor	d[14]	0.5573	0.1187	0.3172	0.78

International Journal of Scientific Knowledge
(Computing and Information Technology)
Volume 1 Issue 2, Oct 2012

ISSN 2305-1493

IJSK

www.ijsk.org

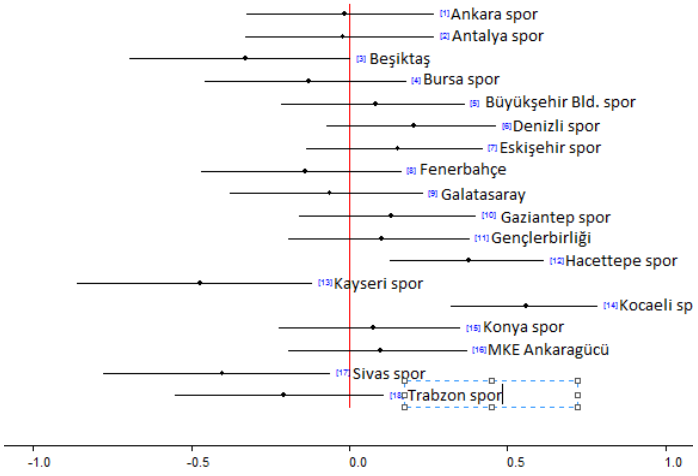
15.	Konyaspor	d[15]	0.07554	0.1457	-0.223	0.3481
16.	Mke Ankaragücü	d[16]	0.0967	0.144	-0.1971	0.3722
17.	Sivasspor	d[17]	***-0.4052***	0.1846	-0.7839	-0.0636
18.	Trabzonspor a.ş	d[18]	-0.2065	0.1699	-0.5601	0.1098
		μ	0.003371	0.05829	-0.1115	0.1159

Table 3. Posterior Probabilities Of Each Game Outcome For Last Six Games

Number	Home team	Away team	Actual score	Posterior probabilities		
				Home wins	Draw	Away wins
301	Antalyaspor	MKE Ankaragücü	1-0	0.4331	0.2916	0.2753
302	Trabzonspor	Fenerbahçe	1-2	0.3707	0.246	0.3833
303	Konyaspor	Ankaraspor	3-0	0.3623	0.3074	0.3303
304	Galatasaray	Sivasspor	2-1	0.3930	0.2676	0.3394
305	Büyükşehir Bld.	Bursaspor	0-1	0.3277	0.2851	0.3871
306	Gençlerbirliği	Kayserispor	0-4	0.2949	0.3371	0.368

Table 4. Results For The Predictive Performance For All Models

	RMSE	MAPE
Loglinear model	3.128	5.876
ANN1	2.983	4.976
ANN2	2.976	4.983
Bayesian approach	2,441	3,805



International Journal of Scientific Knowledge
(Computing and Information Technology)
Volume 1 Issue 2, Oct 2012

ISSN 2305-1493

IJSK

www.ijsk.org

Fig. 1 95% Posterior intervals for team attacking parameters

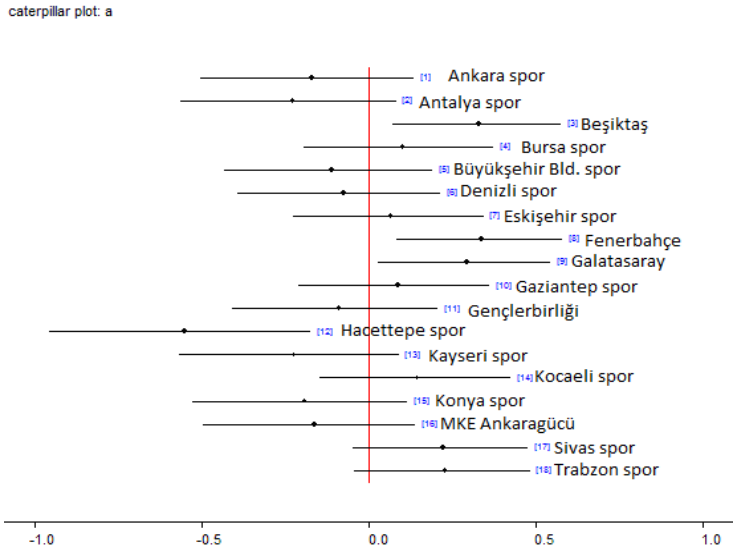


Fig. 2 95% Posterior intervals for team defensive parameters