Student (2000), Vol. 3, No. 4, 229-244

On Modelling Soccer Data

Dimitris Karlis and Ioannis Ntzoufras

Department of Statistics, Athens University of Economics and Business, Greece

Abstract: This paper examines the plausibility of Poisson regression models for the interpretation and prediction of football (soccer) scores. The paper is divided into two parts; the first part consists of an explanatory analysis on the two basic assumptions of modelling soccer data: Poissonity and independence, while the second presents a practical implementation using data from 1997-98 season league results. In the second part of the paper, a simple Poisson model formulation is presented and an application on data from English Premier Division and 'Italian Serie A' for 1997-98 season is provided. The proposed model setting allows for comparison of different models with different natural interpretation. The model with constant home effect over all the teams is selected by the AIC criterion. Negative binomial models are also considered but the improvement is minor.

Key words: negative binomial distribution, English Premier Division, log-linear models, model selection, overdispersion.

1 Introduction

Statistical models can be helpful tools for prediction and interpretation purposes in soccer; for a wide review on the topic see Bennett (1998, chapter 5) and Emonet and Kuonen (2000). The possible outcomes for a team in soccer, are win, draw or loss. These outcomes have different significance according to the competition (e.g. a league or a knock-out tournament). There are plenty of research papers in this area. Research in soccer statistics can be divided in three main categories. The first category models the outcome of a game. This can be done via paired comparisons models (Bradley and Terry, 1952, Kuk, 1995). Fahrmeir and Tutz (1994) developed models for paired data with time varying abilities. State space models were also developed by Rue and Salvesen (2000) for soccer data. Glickman (1995) described an updating scheme for the abilities of the two competitors after each game for chess. The model can be

used for ranking soccer teams and it may be extended to allow for quantification of home effect. Application of related models and optimal prediction schemes have been presented by Kuonen (1996, 1997) for European national football club tournaments. A recent paper, treating data from France'98 World Cup, is given by Kuonen and Roehrl (2000).

The second approach investigates models for the prediction of the number of goals scored by each team. Reep and Benjamin (1968), Reep *et al.* (1971) and Baxter and Stevenson (1988) used the Poisson and the negative binomial distribution to describe the number of goals. Dixon and Coles (1997) provided an extension allowing for extra probabilities in some scores.

The third category concentrates in modelling other characteristics of the game. Such papers are Reep and Benjamin (1968), Barnett and Hilditch (1993) Ridder *et al.* (1994), Clarke and Norman (1995), Pollard and Reep (1997), Dixon and Robinson (1998).

The aim of this paper is dual. Firstly, some basic assumptions, like the Poisson and the independence assumptions are exploited. Secondly, model based inferences based on intensive simulation are proposed. The paper is organized as follows: The validity of the Poisson assumption is examined in Section 2 while the independence assumption is considered in Section 3. Section 4 proposes a log-linear formulation which allows for model selection. Using this formulation we can fit automatically more than one model and select the one supported by a statistical criterion or test (for example AIC or asymptotic χ^2). These models are applied to English Premier Division and 'Italian Serie A' data of 1997-98 season in Section 5. Finally, concluding remarks are given in Section 6.

2 Poisson or not Poisson?

When modelling the number of goals scored by a team, a fundamental question is whether the Poisson distribution can be used to model the 'true' underlined distribution of goals. The Poisson distribution has a formal theoretical basis and is naturally used for events that occur randomly and at a constant rate over the observed time period. In our case, this is equivalent to assuming that the scoring ability of a team is constant throughout the season. This assumption is restrictive since the ability as well as the composition, the physical conditions and the tactic of each team vary from game to game. The assumption of varying scoring ability leads to mixed Poisson distributions. The negative binomial is the most prominent member of this family, and it has been widely used as an alternative model to the Poisson distribution. The negative binomial distribution can be derived from the simple Poisson distribution assuming that its parameter varies according a to Gamma distribution.

The Poisson distribution has the property that the mean is equal to the variance. The index of dispersion (variance to mean ratio) was calculated for 456

teams participated in 24 championships of different European countries, including Germany, Spain, England, Italy, France and the Netherlands. In Figure 1 one can see the points for all the 456 teams. The line indicates the Poisson distribution, i.e. the variance is equal to the mean. Points above the line correspond to overdipsersed cases relative to the simple Poisson distribution. Of the 456 teams, 58.3% have an index of dispersion greater than one (points above the line). If the Poisson assumption was valid we expect that almost half of the teams would have an index of dispersion greater than one (see, also, Anderson and Siddiqui, 1994). The 95% confidence interval of this proportion is given by (0.538, 0.628), and does not support the previous statement. One can draw similar conclusions by examining the number of goals conceded by each team. Again 58.1% of the teams show overdispersion, which is significantly different from 50%. This strongly implies that the distribution of the number of goals is overdispersed relative to the simple Poisson distribution. The above results, which favor a mixed Poisson distribution, are in agreement with the work of various authors such as Moroney (1956), Reep *et al.* (1971) and Baxter and Stevenson (1988).

However, the overdispersion is relatively small since the 95th percentile of the distribution of dispersion index is equal to 1.55. From Figure 1 we cannot distinguish any pattern of deviation from the Poisson distribution. For example, for the negative binomial distribution the variance is a linear function of the mean, while for the Poisson Inverse Gaussian this function is quadratic.

Differences in winning probabilities between Poisson and negative binomial distributions are minimal for the observed range of overdispersion. Given the complicated nature of the negative binomial distribution and especially the difficulty in estimating the parameters, it is plausible to use the simpler Poisson model. Application to real data (see Section 5.1) argues in favour of the above statement since a negative binomial distribution did not exhibit major differences from the Poisson model.

3 The Independence Assumption

Another question which naturally arises in modelling soccer games is whether the number of goals scored by the two opponents are independent. For each championship studied, a χ^2 test of independence in a contingency table was performed to examine possible dependencies. In 15 out of 24 leagues the independence assumption was not rejected at the 5% significance level. Performing the test in the combined contingency table, containing the results of all the championships, the null hypothesis of independence was rejected at the 5% significance level. However, this rejection of the independence hypothesis was rather an artifact due to the large sample size (8250 games). Spearman's correlation coefficient was highly significant even at a 1% significance level, but its value was consid-

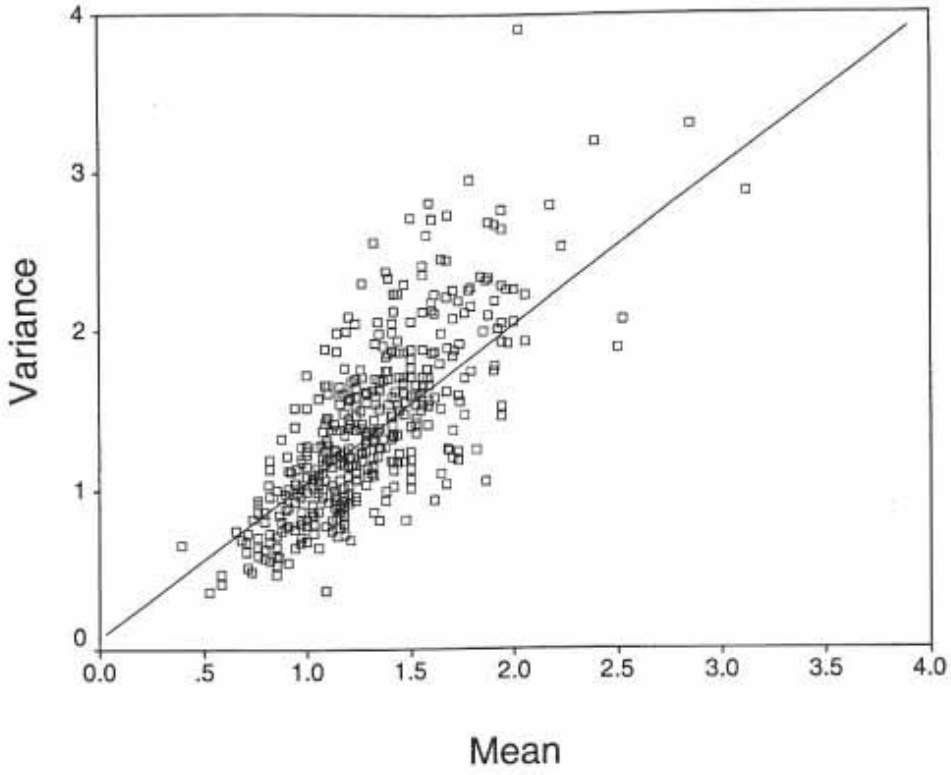

Figure 1. Variance over Mean Ratio for the 456 Teams (Straight line indicates the Poisson distribution)

erably small, namely 0.03, revealing that there is no strong dependence between the two variables.

In order to answer the question more concisely we performed a meta-analysis for the 24 contingency tables, combining the individual p-values as described by Hasselblad (1994). If we have m p-values derived from m independent studies then the quantity $\chi^2 = -2\sum_{i=1}^{m} log(p_i)$ is distributed as a χ^2 variate with $2m$ degrees of freedom. Using this approach we found that the value of the test statistic is 117.013 with 48 degrees of freedom leading to the rejection of the null hypothesis of independence (p-value < 0.001). Thus, there is evidence in favour of the dependence of the two variates, which is, however, rather small.

Karlis and Ntzoufras (2000) prove that under the reasonable assumption that the joint distribution of the number of goals scored by each team is a bivariate Poisson distribution, the outcomes $Z > 0$, $Z < 0$ and $Z = 0$, where Z is the difference on the number of goals, do not depend on the correlation parameter of the bivariate Poisson distribution. Therefore, one can use the independent Poisson formulation to calculate the probabilities of a win, a draw or a loss for each game.

Home	Away Team					
Team	0	1	2	3	4+	Total
0	842	505	272	121	84	1824
1	899	1143	399	195	85	2721
2	677	681	399	116	59	1932
3	393	370	184	83	24	1054
4+	282	265	121	39	12	719
Total	3093	2964	1375	554	264	8250

Table 1: Crosstabulation of Goals Scored by Home and Away Teams Using Data from 24 European Leagues.

4 Poisson and Negative Binomial Model Formulation

In this section some simple candidate Poisson models are presented in detail. Additionally, negative binomial formulation is briefly presented. The use of the full model which includes all main effects (Home, Offense and Defence) and their interactions implies that the offensive and defensive abilities vary in each game depending on the playing ground and the scoring and defending ability of the competing teams. Such a model is not useful for prediction since we need full league data to estimate model parameters. Data of previous years may not reflect performances in present time and estimation of these parameters is problematic. For this reason two simpler candidate models are of great interest. The first model is given by

$$n_{ijk} \sim Poisson(\lambda_{ijk}),\ i = 1, 2 \quad , j, k = 1, \ldots, p, \tag{1}$$

$$log(\lambda_{ijk}) = \mu + h_i + a_j + d_k + h.a_{ij} + h.d_{ik}, \tag{2}$$

where p is the number of teams in the league, n_{ijk} and λ_{ijk} are the observed and the expected number, respectively, of the goals scored by team j, with opponent team k, playing in football ground i (1 for home 2 for away); μ is a constant parameter, h_i is the home/away effect parameter, a_j is the parameter for the offensive performance of j team and d_k encapsulates the defensive performance of k team. The interactions $h.a_{ij}$ and $h.d_{ik}$ determine how much the defensive and offensive abilities differ in home and away games for j and k teams respectively. The motivation for using (2) is the plausible assumption that offensive and defensive abilities of each team change in home and away games.

The above formulation is equivalent to modelling two distinct models for home and away games and therefore can be given by

$$n_{jk}^H \sim Poisson(\lambda_{jk}^H),\ \ log(\lambda_{jk}^H) = \mu^H + a_j^H + d_k^H$$

$$n_{jk}^A \sim Poisson(\lambda_{jk}^A),\ \ log(\lambda_{jk}^A) = \mu^A + a_j^A + d_k^A$$

where

$$n_{jk}^H = n_{2jk}, \quad \mu^H = \mu + h_2, \quad a_j^H = a_j + h.a_{2j}, \quad d_k^H = d_k + h.d_{2k}$$
$$n_{jk}^A = n_{1jk}, \quad \mu^A = \mu + h_1, \quad a_j^A = a_j + h.a_{1j}, \quad d_k^A = d_k + h.d_{1k}$$

In (2) we use sum-to-zero constraints on offensive and defensive parameters and corner constraints on the home/away variable with baseline level the away grounds resulting in

$$\sum_{j=1}^{p} a_j = \sum_{k=1}^{p} d_k = 0, \quad h_1 = h.a_{1j} = h.d_{1k} = 0, \quad \sum_{j=1}^{p} h.a_{2j} = \sum_{k=1}^{p} h.d_{2k} = 0.$$

This parametrization facilitates an easy to use interpretation of model parameters implying that μ is the average of log-mean of the number of goals for away games, h_2 is a measure of average home effect and is given by the difference between the average of log-mean of the number of goals of home and away games, a_j is the away offensive ability of j team expressed in deviations from μ, d_k is the away defensive ability of k team expressed in deviations from μ and $h.a_{2j}$ is the home-away difference in offensive abilities of j team, and $h.d_{2k}$ is home-away difference in defensive abilities of k team.

The second model model is given by

$$log(\lambda_{ijk}) = \mu + h_i + a_j + d_k. \tag{3}$$

This simpler model, also used by Kuonen (1996) and Lee (1997), assumes that offensive and defensive performance are the same for home and away games while home effect is constant over all teams. In this model we use the same parametrization as in (2) resulting in $\sum a_j = \sum d_k = h_1 = 0$. This parametrization implies a straightforward interpretation of the model parameters: μ is the average of log-mean of goals scored in away games, h_2 is the constant home effect while a_j and d_k are the offensive and defensive performances of j and k teams respectively, expressed in deviations from μ. It is obvious that the greater the offensive parameter, the better the offensive performance of the corresponding team, while the lower the defensive parameter the better the defensive performance of the corresponding team.

An alternative modelling approach is to use a negative binomial distribution instead of Poisson. The model formulation slightly changes and according to Venables and Ripley (1994) can be written as

$$n_{ijk} \sim Poisson(\epsilon_{ijk}\lambda_{ijk}), \ \epsilon_{ijk} \sim \Gamma(\theta, \theta)$$

where $\Gamma(a, b)$ denotes the Gamma distribution with mean a/b and variance a/b^2. The parameter θ controls the overdispersion since we now have $E(n_{ijk}) = \lambda_{ijk}$ and $Var(n_{ijk}) = \lambda_{ijk} + \lambda_{ijk}^2/\theta$. Large values of θ imply low over-dispersion. The model formulation is completed by using link functions on λ_{ijk} similar as (2) and (3). More advanced models can be formulated by using regressors also on dispersion parameter θ. Maximum likelihood technique was used for the

estimation of θ via the S-Plus function *glm.nb* provided by Venables and Ripley (1994).

5 Model Based Results

Data of the season 1997-98 of two leagues in Europe, English Premier Division and 'Italian Serie A', are considered. Soccer data form a kind of three-way contingency table with counts the goal scored by team A, against team D, playing in ground H. The factors used for this model are the scoring team A (determining the offensive parameters), the team D against which these goals are scored (determining the defensive parameters) and the home effect H. This contingency table must be handled with great care since it involves zero counts and structural zeros (in the diagonal of scoring and defending teams).

English Premier Division has 20 teams and each team plays 38 games, 19 in home and 19 away football grounds. The total number of games in the league is 380. Due to the high number of games involved in England (league of 20 teams, two cup competitions and European games) league games cannot be separated in full week games. Therefore, some teams may perform more games in one week than other opponents. Every win attributes three points to the winner and every draw one point to each opponent. The team with more points collected is the winner of the league. Positions 2-6 are also of crucial interest since they give the right of playing in European competitions such as 'Champions League' and UEFA cup. Finally, the three teams with the least points collected are relegated in the lower 'first division' and are replaced in the next season by the three best teams of this league.

'Italian Serie A' has 18 teams playing with each opponent twice, once in home and once away football grounds. Each team performs 34 games. The final league consists of 306 football games. Nine games are played each week mainly on Saturday and Sunday, at which each team plays only once. The point system is the same as in the English Premier Division. Positions 2-6 are of crucial interest also since they give the right of playing in European competitions. Finally, the four teams with the least points collected are relegated in the lower division and are replaced at the next season by the four best teams of Serie B league.

Initially some tests were performed to check whether the goals scored by the two opponents are dependent. A cross tabulation of home and away games truncating at 4 goals results in p-values higher than 0.50 for English and over 0.10 for Italian data. Moreover, Spearman correlation is very low, 0.010 for English and -0.032 for Italian data, verifying the findings of Section 3.

Our main interest lies in selecting the model which has adequate fit and good predictive ability. Classical model selection procedures involving χ^2 significance tests should be avoided due to the large number of zeros which makes the χ^2 distribution invalid (see Agresti, 1990) and, furthermore, significance tests will

	Model	Removed Term	AIC English	AIC Italian
1	Full Model		1520.0	1224.0
2	H*A+H*D+A.D	H.A.D	1254.7	1024.4
3	H*A+H*D	A.D	1024.9	746.2
4	H*A+D	H.D	1008.9	726.5
5	H+A+D	H.A	996.0	712.0
[a]	A+D	[5] -H	1020.8	734.2
[b]	H+A	[5] -A	1008.4	746.3
[c]	H+D	[5] -D	1016.4	757.5

Table 2: Table of Backward Model Selection Procedure Details Using AIC for 'English Premier Division' and 'Italian Serie A' Data (Terms A*B indicate that all terms A,B and A.B are included in the model).

support complicated non-parsimonious models due to the large data size (380 games); see Raftery (1995) for a discussion against the use of significance tests when large amount of data are available. For the above reasons, we facilitated a backward method starting from the full model and removing terms that minimize the Akaike information criterion (AIC), introduced by Akaike (1973). All the computations were implemented using S-Plus procedures. Model of independence, $H + A + D$ as given by (3), was selected for both leagues examined. Model of different home effect, as described by (2), is not significantly better than the selected model even if we use approximate χ^2 significance tests.

The selection of $H + A + D$ model in both English and Italian data indicates that a home effect exists but it is constant for all teams. There is a wide discussion in literature on what causes home ground advantage (crowd, dimension of the pitch, psychological effect, travelling distance) which is well discussed by Clark and Norman (1995) and also reviewed by Emonet and Kuonen (2000). We will not pursue this issue further.

5.1 1997-98 English Premier Division Data

5.1.1 Model Based Inference

The selected model of constant home effect, apart from providing us with an insight for the structure of soccer results, can be the basis for predicting future outcomes. Lee (1997) used model based replicated leagues to examine which team was the best. In this section we extend this simulation based approach for calculating final rank probabilities of the league in certain time points during the league.

		Points				
		Observed	Model Based	Observed	Model Parameters	
	Team		Mean ± S.D.	Goals	Offensive	Defensive
1.	Arsenal	78	74.6 ± 7.3	68-33	0.297	-0.387
2.	Manchester Utd.	77	82.2 ± 7.0	73-26	0.361	-0.620
3.	Liverpool	65	68.5 ± 7.5	68-42	0.307	-0.145
4.	Chelsea	63	69.5 ± 7.6	71-43	0.351	-0.118
5.	Leeds	59	59.2 ± 7.7	57-46	0.134	-0.065
6.	Blackburn	58	55.7 ± 7.8	57-52	0.140	0.058
7.	Aston Villa	57	52.9 ± 7.7	49-48	-0.016	-0.031
8.	West Ham	56	52.1 ± 7.6	56-57	0.128	0.149
9.	Derby	55	54.3 ± 7.7	52-49	0.045	-0.007
10.	Leicester	53	58.8 ± 7.7	51-41	0.017	-0.187
11.	Coventry	52	53.2 ± 7.5	46-44	-0.083	-0.121
12.	Southampton	48	49.2 ± 7.6	50-55	0.012	0.107
13.	Newcastle	44	44.9 ± 7.3	35-44	-0.357	-0.132
13.	Tottenham	44	44.7 ± 7.5	44-56	-0.115	0.119
13.	Wimbledon	44	42.9 ± 7.2	34-46	-0.384	-0.089
13.	Sheffield W.	44	44.2 ± 7.5	52-67	0.064	0.307
17.	Everton	40	42.5 ± 7.4	41-56	-0.186	0.116
17.	Bolton	40	39.9 ± 7.3	41-61	-0.181	0.201
19.	Barnsley	35	27.9 ± 6.9	37-82	-0.262	0.494
20.	Crystal Pallace	33	32.2 ± 6.6	37-71	-0.273	0.349

Table 3: English Premier Division: Final League Table Details, Estimated Model Parameters and Average Points of Simulated Leagues (S.D. = Standard Deviation of Points from Generated Leagues).

In order to assess which team was the best, according to the selected model of constant home effect, the estimated parameters were used to generate replications of leagues. The total team points and the ranking of each replicated league were used to assess the distribution of the final league under the assumption that the model used is a sufficient summary of reality and the teams have the same performance as in observed league. This analysis accounts for corrections of games that were surprisingly unexpected or won by luck. For each dataset 10,000 leagues were generated.

Table 3 gives details of the final league table, the estimated offensive and defensive parameters and the average points for the simulated leagues. The constant parameter is equal to 0.061 and the home effect 0.327. According to these parameters the expected number of goals for an average away team is 1.06 while the goals scored by a home team is about 39% higher (expected number of home goals 1.47).

From the average points of the simulated data we clearly see that Manchester United was better, according to the model, than Arsenal which won the league.

		Final Rank							
	Team	1	1.5	2	2.5	3-6	6.5-10.5	10-17.5	18-20
1.	Arsenal	17.7	2.7	32.5	4.2	39.7	2.5	0.1	
2.	Manchester Utd.	64.9	3.6	19.4	1.7	10.2	0.2		
3.	Liverpool	5.1	1.1	13.7	3.0	64.0	11.9	1.2	
4.	Chelsea	6.4	1.2	16.8	3.0	61.8	9.8	1.0	
5.	Leeds	0.4	0.1	2.0	0.7	43.4	39.4	13.8	0.2
6.	Blackburn	0.2	0.0	0.8	0.3	25.1	43.8	26.7	0.8
7.	Aston Villa	0.1	0.0	0.4	0.1	16.8	42.1	38.7	1.8
8.	West Ham			0.2	0.1	14.7	39.9	43.0	2.0
9.	Derby	0.1	0.1	0.5	0.2	22.1	44.0	32.1	1.0
10.	Leicester	0.3	0.1	1.9	0.5	41.8	39.8	15.4	0.2
11.	Coventry			0.4	0.1	17.9	42.4	37.9	1.2
12.	Southampton			0.1	0.1	8.2	31.8	55.0	4.8
13.	Newcastle					2.7	17.7	67.1	12.5
13.	Tottenham					2.8	17.0	66.9	13.3
13.	Wimbledon					1.3	12.2	68.1	18.4
13.	Sheffield W.					2.3	16.4	66.1	15.2
17.	Everton					1.3	11.6	66.5	20.5
17.	Bolton					0.5	6.4	61.3	31.8
19.	Barnsley						0.1	9.8	90.1
20.	Crystal Pallace						0.5	25.5	74.0

Table 4: English Premier Division: Percentages of Final League Ranking (half ranks denote ties).

Manchester United has the best offensive and defensive parameter. The difference of average points is equal to 7.60 in favour of Manchester United. The 95% of the points of the replicated leagues for Arsenal is between 60 and 89 while for Manchester United between 68 and 95. Moreover, if we consider the difference between Arsenal and Manchester United for each replicated league then the 95% of the values is between -27 and 13 (the negative sign favors Manchester United). For the rest of the league positions we can infer that Chelsea was slightly better than Liverpool while Derby was better than both Aston Villa and West Ham. For the last places of the league Crystal Pallace was found better than Barnsley while Bolton was found worse than Everton. Everton and Bolton were tied in 17th place but finally Bolton was relegated due to its worse goal difference.

We can draw similar conclusions by examining the rank percentages of each team. Manchester United is clearly better with 65% probability of winning the title while Arsenal had only 18%. Moreover, Barnsley and Crystal Pallace had high probabilities of relegation but the third relegation place could not clearly identified. Bolton, which was finally relegated, had the highest relegation probability (30%) but Everton, Sheffield, Wimbledon, Tottenham and Southampton had also high probability of relegation (over 10%). Therefore, the performance

of many teams was similar and a lot of uncertainty was involved in the determination of relegated teams.

5.1.2 Prediction From Batch to Batch

In this section the data were divide in 38 batches of ten games. We assume that we have availability of k games ($k/10$ batches) and then we simulate the rest of $380 - k$ games to get probabilities for the final ranking. The purpose was to examine the performances of each team with respect to time. The number of games used are 100, 150, 190, 250, 300, 330, 350, 360 and 370 (or after batches 10, 15, 19, 25, 30, 33, 35, 36, 37). Results, concerning the teams at the first ranks are summarised in Figure 2.

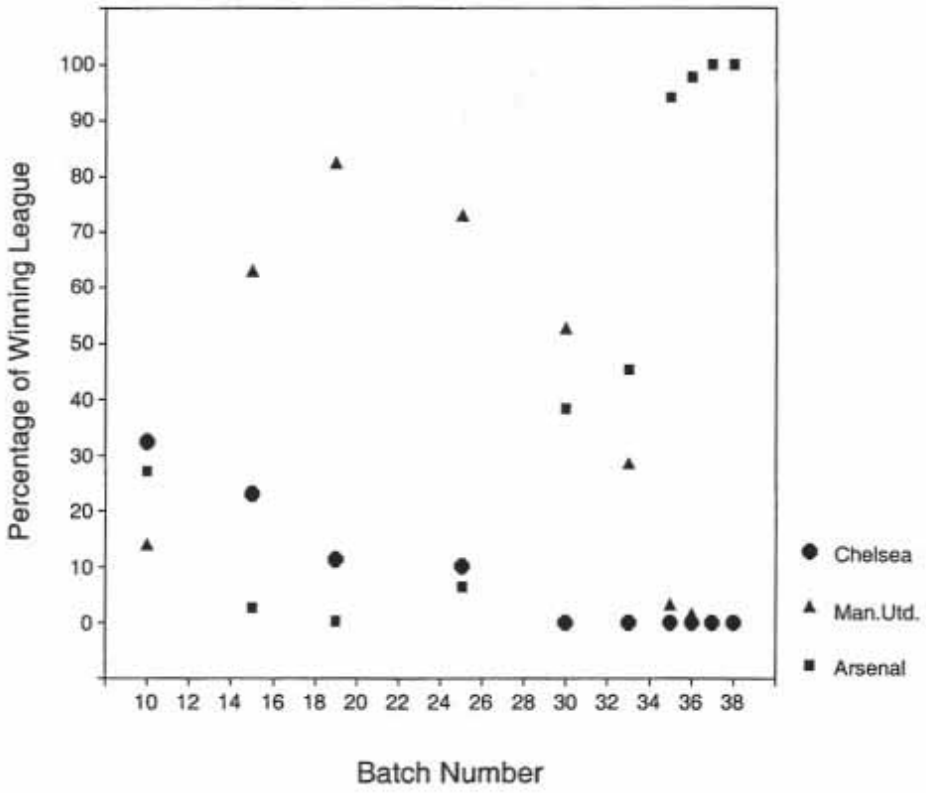

Figure 2. English Premier Division: Percentages of Winning the League per Batch of 10 Games.

From Figure 2 we see that Manchester United had higher probability than Arsenal to win the league until batch 30. Manchester United lost its power gradually after batch 19 but the crucial period was between 33th and 35th game when it lost valuable points in a game against Liverpool. The 4-1 away win of Arsenal against Blackburn was also of crucial significance since the two teams had close fitted values (1.5 for Arsenal and 1.2 for Blackburn). Manchester

United unexpectedly lost points by Newcastle (1-1) diminishing any chance to win the title. Arsenal lost the two last games (both at home ground).

5.2 Results from the 1997-98 'Italian Serie A' Data

'Italian Serie A' first division data analysis leads to similar results. The same model was also selected and generally, despite of its simplicity, this model captures most of data variation.

In Italian data, Juventus and Inter were the main competitors for winning the title. Roma had the best offensive and Inter the best defensive parameter while Juventus had both the second best defence and attack. Home effect was estimated as high as 0.341, that is, any team j with opponent k has expected number of goals about 40.6% higher in home games than in away games.

According to the simulated average points Juventus is first with average points 70.4 while Inter second with 68.7 (getting the second place in Champions League). The standard deviation of the generated league points for both teams is about 6.7 points. If we consider the difference of the points in each simulated league then we have mean difference of 1.7 points in favor of Juventus (standard deviation of 9.8 points).

Analysis assuming availability of games before weeks 10, 17 (end of first round), 22, 26, 30, 31, 32 and 33 shows that Juventus had higher probability of winning the title in all weeks examined. The only week that Inter and Juventus had been really close (in probabilities) was week 26 when Juventus had 37% chance of winning the title while Inter 31% (actual difference of one point). After four games, the actual difference was still one point but Juventus raised its probability to 61%. At week 31 Juventus cleared things out since its probability of winning the league was raised to 93% while the actual difference was 4 points. That week Juventus won over Inter by 1-0 while the expected number of goals of the final model were 1.31 for Juventus and 0.90 for Inter. Using data available until week 30 the corresponding expected number of goals were 1.34 and 0.96 giving winning probability of 44% to Juventus and 27% to Inter.

		Week								
	Team	10	17	22	26	30	31	32	33	34
1.	Juventus	52.6	66.1	67.5	37.1	60.9	93.2	95.9	100.0	winner
2.	Inter	34.6	25.1	11.9	33.4	31.4	4.5	2.5	***	

Table 5: Percentages of Winning the 'Italian Serie A' League per Week for Juventus and Inter (*** = Not possible to win the title).

The final conclusion is that Juventus was the best team for 1997-98 season in Italy. Moreover, Inter was undoubtedly a very good opponent and could have won the league in any small error of Juventus.

5.3 Overdispersion and Negative Binomial Model

Simple statistics of aggregated data from English and Italian leagues indicate that the goals scored by each team are slightly over-dispersed. Dispersion ratios for goals scored by and against each team for each league are given in Table 6.

English Premier Division			'Italian Serie A'		
	DI for Goals			DI for Goals	
Team	Scored	Conceded	Team	Scored	Conceded
Arsenal	1.273	1.380	Atalanta	1.179	1.250
Aston Villa	1.044	1.056	Bari	1.083	1.224
Barnsley	0.860	1.441	Bologna	1.050	1.294
Blackburn	1.829	1.399	Brescia	0.857	0.986
Bolton	1.171	1.264	Empoli	1.287	1.049
Chelsea	1.596	1.155	Fiorentina	1.248	0.912
Coventry	1.302	0.958	Inter	1.179	0.746
Crystal P.	0.915	1.075	Juventus	0.938	0.771
Derby	1.557	1.547	Lazio	1.057	1.152
Everton	1.021	0.907	Lecce	1.027	1.281
Leeds	1.505	0.900	Milan	1.079	1.261
Leicester	0.937	1.422	Napoli	1.015	1.167
Liverpool	0.971	1.066	Parma	0.675	1.011
Man. Utd.	1.333	1.114	Piacenza	1.217	1.018
Newcastle	0.844	0.865	Roma	1.245	1.082
Sheffield W.	0.886	1.607	Sampdoria	1.317	0.975
Southampton	1.155	0.960	Udinese	1.212	1.106
Tottenham	1.518	1.347	Vicenza	0.684	1.040
West Ham	1.274	1.072			
Wimbledon	1.135	1.749			
Total	1.276		Total	1.186	

Table 6: Dispersion Index (DI) for Goals Scored and Conceded by Each Team.

Fitting the negative binomial model to English data with regressors similar to (3) resulted in $\hat{\theta} = 16.91$ and standard error $\sigma_{\hat{\theta}}$ equal to 11.69. The overdispersion parameter θ is much lower than the corresponding parameter for 'Italian Serie A' data ($\hat{\theta} = 849.6 \pm 1251.0$). Note that the larger the value of θ the smaller the overdispersion is. Model parameter estimates are similar to the corresponding Poisson model. Replicated leagues using negative binomial formulation give

similar results to the Poisson model since overdispersion is low (θ in both cases is large compared to λ_{ijk}). These findings verify the results of Section 2.

6 Discussion

Statistical modelling of soccer games is a real challenge since it introduces statistics in everyday activities. In this paper we attempted to answer the basic questions concerning the modelling of soccer games. It was found that the Poisson assumption can be considered as plausible. It is very simple to be applied via standard statistical software and it provides results that do not differ from those obtained by the negative binomial model. The latter, even though it is reasonable demands more intensive computations and special programming for estimation of model parameters.

The number of goals scored by a team is clearly a sufficient indicator for the strength of a team since it must score in order to win. In order to support this statement statistically correlations between the final ranking and the number of goals scored and conceded by each team were calculated from 24 leagues. According to our findings, correlations are as high as 0.85 showing that goal scored can be used to determine the performance of a team.

For practical purposes, when statistical software supporting generalized linear models is not available, one may estimate the offensive ability as the mean number of goals scored and the defensive ability as the mean number of goals scored against. The home effect can be calculated as described in Clarke and Norman (1995) and the probability of a win via simple packages supporting probability function calculation (for example spreadsheets). All these calculations do not need special statistical knowledge and can easily be performed by non statisticians.

The log-linear formulation of our model allows the examination of several models at once. For example one can test whether the home effect is constant over all teams. Results from the 1997-98 data of two distinct European leagues analysed supported the model with constant home effect. No interaction terms were found to improve the model fit.

Finally, it is important that the simple model of constant home effect supported by both English and Italian data of season 1997-98 can satisfactory interpret underlined structures of soccer data.

Acknowledgements: The authors would like to thank Prof. Evdokia Xekalaki and Prof. Petros Dellaportas and an anonymous referee for their comments.

Résumé : Cet article examine la plausibilité d'interpréter et de prédire des scores au football à l'aide des modèles de la régression Poisson. L'article consiste en deux parties. La première partie examine les deux hypothéses principales concernant l'analyse des données au football, à savoir la poissonité et l'indépendance. Dans la deuxième partie de cet article nous trouvons une simple formulation du modèle Poisson et une application de ces données à la Première Division Anglais (English Premier Division) et à la Serie Italienne A (Italian Serie A). Nous comparons ensuite différents modèles et leurs interprétation. La sélection des modèles à l'aide du critère AIC donne un modèle contenant un 'home effect' (jouer à domicile) commun à toutes les équipes. Des modèles de régression binomiale négative ont aussi été considerés, mais sans produire d'amélioration considérable.

References

[1] Agresti, A. (1990). *Categorical Data Analysis.* John Wiley and Sons, New York.

[2] Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *2nd International Symposium on Information Theory*, Akademia Kiado, Budapest 267–281.

[3] Anderson, J. and Siddiqui, M. (1994). The Sampling Distribution of the Index of Dispersion. *Communication in Statistics: Theory and Methods*, **23**, 897–911

[4] Barnett, V. and Hildich, S. (1993). The Effect of an Artificial Pitch Surface on Home Team Performance in Football (Soccer). *Journal of the Royal Statistical Society A*, **156**, 39–50

[5] Baxter, M. and Stevenson, R. (1988). Discriminating Between the Poisson and Negative Binomial Distributions: An Application to Goal Scoring in Association Football. *Journal of Applied Statistics*, **15**, 347-438

[6] Bennet, J (1998). *Statistics in Sport.* First Edition, Edward Arnold, London.

[7] Bradley, R. A. and Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs I The Method of Paired Comparisons. *Biometrika*, **39**, 324–345

[8] Clarke, S. R. and Norman, J. M. (1995). Home Ground Advantage of Individual Clubs in English League. *The Statistician*, **44**, 509-521

[9] Dixon, M. J. and Coles, S. G. (1997). Modelling Association Football Scored and Inefficiencies in Football Betting Market. *Applied Statistics*, **46**, 265-280

[10] Dixon, M. J. and Robinson, M.E. (1998). A Birth Process Model for Association Football Matches. *The Statistician*, **47**, 523–538

[11] Emonet, B. and Kuonen, D. (2000). Revisiting Statistical Applications in Soccer. *Technical Report* **2000.2** Department of Mathematics, Swiss Federal Institute of Technology, Lausanne

[12] Fahrmeir, L. and Tutz, G. (1994). Dynamic Stochastic Models for Time-Dependent Ordered Paired Comparison System. *Journal of the American Statistical Association*, **89**, 1438-1449

[13] Glickman, M. A. (1995). A Comprehensive Guide to Chess Ratings. *American Chess Journal*, **3**, 59–102
[14] Hasselblad, V. (1994). Meta-analysis in Enviromental Statistics. *Handbook of Statistics*, **12**, 691–716
[15] Karlis D. and Ntzoufras, I. (2000). Distributions Based on Poisson Differences with Applications in Sports. *Technical Report* **101**, Dept. of Statistics, Athens University of Economics and Business.
[16] Kuk, A. Y. C. (1995). Modelling Paired Comparison Data with Large Numbers of Draws and Large Variability of Draw Percentages among Players. *The Statistician*, **44**, 523–528.
[17] Kuonen, D. (1996). Modelling the Success of Football Teams in the European Championships. (in French). *Technical Report* **96.1** Department of Mathematics, Swiss Federal Institute of Technology, Lausanne
[18] Kuonen, D. (1997). Statistical Models for Knock-out Soccer Tournaments. *Technical Report* **97.3** Department of Mathematics, Swiss Federal Institute of Technology, Lausanne
[19] Kuonen, D. and Roehrl, A.S.A. (2000). Was France's World Cup Win Pure Chance? *Student*, **3**, 153–166.
[20] Lee, A. J. (1997). Modeling Scores in the Premier league: Is Manchester United Really the Best? *Chance*, **10(1)**, 15-19
[21] Moroney, M.J. (1956). *Facts from Figures*. 3rd Edition, Penguin, London.
[22] Pollard, R. and Reep, C. (1997). Measuring the Effectiveness of Playing Strategies at Soccer. *The Statistician*, **46**, 541-550
[23] Raftery, A.E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology 1995* (P.V. Marsden ed.). Blackwell, Oxford.
[24] Ridder, G. , Cramer, J. S. and Hopstaken, P. (1994). Down to Ten: Estimating the Effect of a Red Card. *Journal of the American Statistical Association*, **89**, 1124-1127
[25] Reep, C. and Benjamin, B. (1968). Skill and Chance in Association Football. *Journal of the Royal Statistical Society A*, **131**, 581-585
[26] Reep, C. , Pollard, R. and Benjamin, B. (1971). Skill and Chance in Ball Games. *Journal of the Royal Statistical Society A*, **133**, 623-629
[27] Rue, H. and Salvesen, O. (2000). Prediction and Retrospective Analayis of Soccer Matches in a League. *The Statistician*, **49**, 399-418
[28] Venables, W. N. and Ripley, B. D. (1994). *Modern Applied Statistics with S-Plus*. Springer-Verlag, New York.