On Modelling Association Football Data

Dimitris Karlis and Ioannis Ntzoufras
Athens University of Economics and Business, Greece

SUMMARY

This paper examines the plausibility of Poisson regression models for the interpretation and prediction of association football (soccer) scores. The paper is divided into two parts; the first part consists of a thorough 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. According to our findings departures from Poisson distribution are minor and therefore can be ignored. The dependence of the goals scored between the two opponents is low, while in a bivariate Poisson formulation, the correlation coefficient does not affect the outcome of the game. In the second part of the paper we present a simple Poisson model formulation and an application on data from English Premier division and Italian serie A of 1997-98 season. These models are used to extract information for the home advantage, offensive and defensive performance of each team and probabilistically quantify their final ranking. 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 the improvement is minor.

1 Introduction

Sports is a blooming field for applying statistical methodologies as well as a platform for developing new ones. The amount of money invested by well organized sports related industries (including proffesional sports clubs) has extremely increased over the last decades (see Szymanski and Smith, 1997). Statistical modelling can help the managers of these companies to make crucial decisions in certain circumstances. Furthermore, betting on the outcome of soccer matches has a long tradition in United Kingdom and other European countries. Football pools typically involve the selection of matches whose the outcome is hard to predict. Such bets involve either the final score or other specific chatacteristics of the match, such as halftime result or the scorer of the first goal. In making bets, the challenge is to trace matches of which probabilities have been determined inaccurately and thus the expected gain is high. Statistical models can be helpful tools for such purposes; see, for example, Jackson (1994).

There are many articles analysing soccer data or data from other sports using methodologies that can be easily applied in soccer data as well. The dispersion of sports related papers in many journals, usually from different scientific fields, makes it difficult to gather all the results. For a wide review on the topic see Bennet (1998, chapter 5).

Regarding soccer, the possible outcomes for a team are win, draw or loss. These outcomes have different significance according to the competition. For example in long period leagues, loosing a game may not affect very much the final results while in knock-out competitions it has direct effect since the team is disqualified from the next round.

Research in soccer statistics can be divided in three main categories. The first category models the outcome of a game. The Bradley-Terry model (Bradley and Terry, 1952) for paired comparisons is the most important formulation used for modelling the outcome of soccer games. An alternative model adjusting also for draws is described in Kuk (1995). The drawback of the above models is that team abilities are assumed to be constant throughout the period under consideration. Fahrmeir and Tutz (1994) developed models for paired data with time varying abilities using data from the German soccer league. State space models were also developed by Rue and Salvesen (1997) for soccer data and Glickman and Stern (1998) for American football data. Moreover, Glickman (1995), in a paper devoted to chess, described an updating scheme for the abilities of the two competitors after each game which can easily be applied to soccer data. The model can be used for ranking soccer teams and it may be extended to allow for quantification of home effect. For extensive application of this model to data from European knock-out competitions and some extensions see Kuonen (1997a, b).

The second approach investigates models for the prediction of the number of goals scored by each team. Reep and Benjamin (1968) and Reep *et al.* (1971) used the Poisson and the negative binomial distribution to describe the number of goals. If pure chance governs the game, the Poisson distribution is an appropriate choice. Departures from the Poisson distribution lead to the negative binomial formualtion which arises when the skill of each team varies according to a Gamma distribution. Baxter and Stevenson (1988) provided a detailed discussion on the use of negative binomial on soccer data. On the other hand, several sport statisticians adopted the Poisson distribution. For example, Maher (1982) and Lee (1997) used Poisson regression models with covariates regarded as the offensive and defensive abilities of the teams. Rue and Salvesen (1997) described a dynamic model while Dixon and Coles (1997) have recently extended the model by adding assumptions about specific scores observed to be underestimated from the simple Poisson model.

The third category concentrates in modelling other characteristics of the game (for example tactics or player positioning). Clarke and Norman (1995) report significant home effect using simple calculations. Reep and Benjamin (1968) modelled the number and the patterns of passing moves within a game. Barnett and Hilditch (1993) applied standard nonparametric tests to check if home teams with artificial pitches have a significant advantage. Ridder *et al.* (1994) examined the effect of sending-off a player while Pollard and Reep (1997) analysed the effect of different playing strategies. Stefani (1983) described betting strategies for football pools. Dixon and Robinson (1998) modelled the scoring times as a birth proccess using survival analysis models. Finally, soccer examples have been used in introducing new statistical methodologies; see for example Mosteller (1997) and Wright (1997).

Discussion on the distribution of the number of goals and the validity of the Poisson assumption is examined in Section 2. Section 3 considers whether the independence assumption for the goals scored by the two opponents of each game is plausible. Section 4 explains the structure of soccer data used for Poisson regression while Section 5 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 in English Premier Division and 'Italian Serie A' data of 1997-98 season in Section 6. Finally, concluding remarks are given in Section 7.

2 Poisson or not Poisson?

When modelling the number of goals scored by a team, a fundamental question is whether Poisson can be used to approximate the 'true' underlined distribution of goals. The Poisson distribution has a formal theoretical basis and is naturally used for events that occur randomly 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 Gamma distribution.

The Poisson distribution has the unique property that the mean is equal to the variance.

We calculated the index of dispersion (variance to mean ratio) for 456 teams participated in 24 championships of different European countries, including Germany, Spain, England, Italy, France and the Netherlands for the last 5 years. Of the 456 teams, 58.3% have an index of dispersion greater than one. 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, given by (0.538, 0.628), does not support the previous statement. We 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.

However, the overdispersion is relatively small since the 95th percentile of the distribution of dispersion index is equal to 1.55. Figure 1 presents the histogram of the dispersion index for 456 European teams. Moreover, from Figure 2 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 (see Johnson *et al.* , 1992) this function is quadratic.

In order to examine the effect of the overdispersion in the probability of each soccer game outcome, we calculated the probability for team A to win over team B assuming a negative binomial distribution with overdispersion varying from 1 (Poisson distribution) to 1.5, which is the interval observed in the 24 leagues examined. Even when the overdispersion of both teams is near 1.5, the probability of winning the game differs less than 5% of the probability when assuming Poisson distribution. This shows that the deviation from the Poisson distribution is not of much concern. 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 6.1) argues in favour of the above statement since a negative binomial distribution did not exhibit major differences from the Poisson model. Figure 1

To conclude, the distribution of goals is slightly overdispersed relative to the Poisson assumption. Since the overdispersion is small, Poisson model formulation will give acceptable results while the computational gain is high. Moreover, there is not any overdispersion pattern supporting a particular mixed Poisson distribution. Figure 2

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 was performed to examine possible dependencies. In 15 out of 24 leagues the independence assumption was not rejected. If we combine the results of all the championships, the null hypothesis is rejected. However, the 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 0.001 level of significance, but its value was considerably 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 chi-square 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. Thus, there is evidence in favour of the dependence of the two variates, which is, however, rather small. Is this dependence important for the determination of the outcome? We attempt to answer this question by imposing some simple and relatively reasonable assumptions. Table 1

Suppose that random variables X, Y represent the number of goals scored by each team respectively. Furthermore, suppose that their joint distribution is a bivariate Poison (see Kocherlakota and Kocherlakota, 1992). Thus the probability function is given by

$$P(X=x, Y=y) = P(x,y) = exp\left\{-(\theta_1+\theta_2+\theta_3)\right\} \frac{\theta_1^x}{x!}\frac{\theta_2^y}{y!} \sum_{i=0}^{min(x,y)} \left(\begin{array}{c} x \\ i \end{array} \right) \left(\begin{array}{c} y \\ i \end{array} \right) i! \left(\frac{\theta_3}{\theta_1\theta_2} \right)^i . \quad (1)$$

The bivariate Poisson distribution defined in (1) allows for dependence between the two random variables. Marginally each random variable follows a Poisson distribution with parameters $\theta_1 + \theta_3$ and $\theta_2 + \theta_3$ respectively. A natural interpretation of the parameters is that θ_1 and θ_2 are the parameters reflecting the scoring ability of the two teams while parameter θ_3 reflects game conditions. Parameter θ_3 is the covariance between X and Y.

Suppose further that the random variable Z is the difference between X and Y ($Z = X - Y$). The sign of Z represents the winner of the game while zero value corresponds to a draw. The distribution of Z is given by

$$P(Z=z) = \sum_{k=0}^{\infty} P(X=k)P(Y=k-z|X=k) \quad (2)$$

Although the marginal distribution of X is also a Poisson, the conditional distribution of

Y given X is the convolution of a Poisson with a binomial distribution (see Kocherlakota and Kocherlakota, 1992). Description of the distribution of Z is provided by Irwin (1937) and Skellam (1946) when X and Y are independent and dependent Poisson variates, respectively.

The following theorem provides evidence that the probability of a soccer game outcome (win, loss, draw), when bivariate Poisson distribution is assumed, does not depend on the correlation coeficient.

> **Theorem**: If the joint distribution of (X,Y) is the bivariate Poisson distribution as given by (1) then the probabilities $P(Z=0)$, $P(Z>0)$, and $P(Z<0)$ do not depend on θ_3 .

Proof: Suppose that r, s are integers and that $r < s$. We can see that $P(Z<0) = \sum_{r=0}^{\infty} \sum_{s=r+1}^{\infty} P(X=r, Y=s)$ and using (1) it follows that

$$P(Z<0) = exp\{-(\theta_1+\theta_2+\theta_3)\} \sum_{r=0}^{\infty} \sum_{s=r+1}^{\infty} \frac{\theta_1^r}{r!} \frac{\theta_2^s}{s!} \sum_{i=0}^{r} \binom{r}{i} \binom{s}{i} i! \left(\frac{\theta_3}{\theta_1\theta_2}\right)^i =$$

$$= exp\{-(\theta_1+\theta_2+\theta_3)\} \sum_{i=0}^{\infty} \sum_{r=i}^{\infty} \sum_{s=r+1}^{\infty} \frac{\theta_1^{r-i}\theta_2^{s-i}}{(r-i)!i!(s-i)!}\theta_3^i$$

By setting $t = r - i$ and $\rho = s - i$ we obtain that

$$P(Z<0) = exp\{-(\theta_1+\theta_2+\theta_3)\} \sum_{i=0}^{\infty} \sum_{t=0}^{\infty} \sum_{\rho=t+1}^{\infty} \frac{\theta_1^t\theta_2^\rho}{t!i!\rho!}\theta_3^i =$$

$$= exp\{-(\theta_1+\theta_2+\theta_3)\} \sum_{i=0}^{\infty} \frac{\theta_3^i}{i!} \sum_{t=0}^{\infty} \sum_{\rho=t+1}^{\infty} \frac{\theta_1^t\theta_2^\rho}{t!\rho!} =$$

$$= exp\{-(\theta_1+\theta_2+\theta_3)\}\, exp(\theta_3) \sum_{t=0}^{\infty} \sum_{\rho=t+1}^{\infty} \frac{\theta_1^t\theta_2^\rho}{t!\rho!} =$$

$$= exp\{-(\theta_1+\theta_2)\} \sum_{t=0}^{\infty} \sum_{\rho=t+1}^{\infty} \frac{\theta_1^t\theta_2^\rho}{t!\rho!}$$

which does not depend on θ_3. With similar arguements we can show that

$$P(Z>0) = exp\{-(\theta_1+\theta_2)\} \sum_{t=0}^{\infty} \sum_{\rho=t+1}^{\infty} \frac{\theta_1^\rho\theta_2^t}{t!\rho!}.$$

The above probabilities are equal to the corresponding probabilities from two independent Poisson distributions with means θ_1 and θ_2. We can also show that

$$P(Z=0) = exp\{-(\theta_1+\theta_2)\} \sum_{t=0}^{\infty} \frac{\theta_1^t\theta_2^t}{t!\rho!}.$$

The importance of the above theorem is that the dependence between the X and Y does not affect the probabilities of soccer outcomes (win, loss, draw) while this is not true for the probabilities of the type $P(Z = z)$, where z takes the values $\ldots, -2, -1, 1, 2, \ldots$. Therefore, we can use the indepedent Poisson formulation to accurately calculate the probabilities of a win, a draw or a loss for each game. The result of the theorem is not useful when the practitioner needs to predict the number of goals scored by each team or the total number of goals scored in the game.

In conclusion, violation of the independence assumption in Poisson distribution does not affect the probabilities of soccer outcomes and therefore simple Poisson models can be utilized to accurately calculate these probabilities.

4 English Premier Division and 'Italian Serie A' Data

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 play 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 seperated 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 league. 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.

5 Poisson and Negative Binomial Model Formulation.

In this section we present possible simple model formulations that can be used for soccer data. In the previous sections it is evident that the single data can be thought as Poisson distrubuted. In this section some simple canditate Poisson models are presented in detailed. Additionally, negative binomial formulation is briefly presented.

The full Poisson model takes the following form

$$n_{ijk} \sim Poisson(\lambda_{ijk}), \tag{3}$$

$$log(\lambda_{ijk}) = \mu + h_i + a_j + d_k + h.a_{ij} + h.d_{ik} + a.d_{jk} + h.a.d_{ijk} \tag{4}$$

where 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 (away or home); μ is a constant parameter, h_i is the home 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 rest of the parameters are interactions between the three main factors that are interpreted accordingly. The full model implies that the offensive and defensive abilities vary in each game depending on the playing ground, the scoring and defending ability of the competing teams. Such a model is not useful for prediction since we need full league data (which are not available) to estimate model parameters. Data of previous years may not reflect performances in present time and estimation of these parameters is problematic.

Two simpler canditate models are of great interest. The first model is given by

$$log(\lambda_{ijk}) = \mu + h_i + a_j + d_k + h.a_{ij} + h.d_{ik}. \tag{5}$$

The motivation for using (5) 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 dinstinct models for home and away games and therefore can be given by

$$n^H_{jk} \sim Poisson(\lambda^H_{jk}), \quad log(\lambda^H_{jk}) = \mu^H + a^H_j + d^H_k$$

$$n^A_{jk} \sim Poisson(\lambda^A_{jk}), \quad log(\lambda^A_{jk}) = \mu^A + a^A_j + d^A_k$$

where

$$n^H_{jk} = n_{2jk}, \quad \mu^H = \mu + h_2, \quad a^H_j = a_j + h.a_{2j}, \quad d^H_k = d_k + h.d_{2k}$$

$$n^A_{jk} = n_{1jk}, \quad \mu^A = \mu + h_1, \quad a^A_j = a_j + h.a_{1j}, \quad d^A_k = d_k + h.d_{1k}$$

In (5) 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 the difference of the average of log-mean of the number of goals of home games from 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.

The second model model is given by

$$log(\lambda_{ijk}) = \mu + h_i + a_j + d_k. \qquad (6)$$

This simple model, also used by 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 (5) resulting in $\sum a_j = \sum d_k = h_1 = 0$; where h_1 indicates goals scored by away teams. This parametrization implies a straigtforward 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 alterntive 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 $Var(n_{ijk}) = \lambda_{ijk} + \lambda^2_{ijk}/\theta$. Large values of θ imply low over-dispersion. The model formulation is completed by using link functions on λ_{ijk} similar as (6) and (5). 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).

6 Model Based Results

In this section we present results from the English and Italian league for the season 1997-98. Initially some tests were performed to check whether the goals scored by the two opponents are dependent. A crosstabulation 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, veryfying the findings of Section 3. Table 2

Our main interest lies in selecting the model which has adequate fit and good predictive ability. Classical procedures using approximate significance tests should be avoided due to the large number of zeros. Furthermore, significance tests will support compicated non-parsimonious models due to the large data size (380 games). For the above reasons, Akaike information criterion (AIC), introduced by Akaike (1973), was used. Model of independence, $H + A + D$ as given by (6), was selected for both leagues examined. Model of different home effect, as described by (5), is not significantly better than the selected model even if we use approximate χ^2 significance tests.

6.1 1997-98 English Premier Division Data

6.1.1 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. Table 3

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 unfair or won by luck; also see Lee (1997). For each dataset 10,000 leagues were generated. Table 4

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 away team is 1.06 while the goal 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 than Arsenal which won the league. Manchester has the best offensive and defensive parameter. The difference of average points is equal to 7.60 in favour of United. The 95% of the points of the replicated leagues for Arsenal is between 60 and 89 while for Manchester between 68 and 95. Morever if we consider the difference between Arsenal and Manchester for each replicated league then the 95% of the values is between -27 and 13 (the negative sign favors Manchester). For the rest of the places we see that Chelsea was slightly better than Liverpool while Derby was better than both Aston Villa and West Ham. This is plausible since Derby had better goal difference than the other two teams. The term 'goal difference' is used to refer to the total number goals scored by a team minus total number of goals conceded by the same team. 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. 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 is not clearly defined. 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.

6.1.2 Prediction From Batch to Batch

In this section we divide the data 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 are summarised Figures 3, 5 and 7. Figures of actual points per game for these batches are also given for comparison purposes (see Figures 4, 6 and 8).

Figures 4 - 8

From Figure 3 we see that United had higher probability than Arsenal to win the league until batch 30. 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). United unexpectably lost points by Newcastle (1-1) diminishing any chance to win the title. Arsenal lost the two last games (both at home ground). The first was really a tough game since Arsenal's opponent, Liverpool, had slightly better expected number of goals and probability of winning the game. The four goal difference was a surprise for a game that a-priori appeared to be between two opponents of equal power.

Figure 5 depicts the probabilities of direct exit to European competitions (positions 1-6). Chelsea had high probability of a European exit but finally won the European Cups Winner Cup getting a place in this competition instead of the UEFA cup and therefore the 7th team also won the right to play in Europe. Aston Villa, which finally qualified as 7th for the UEFA cup, had a very low probability of getting a direct European exit. Liverpool, Chelsea and Leeds earned a place between positions 1 and 6 easily. Chelsea had probability over 98% even after the 100th game (18th of October), while Liverpool ensured its European exit after the 190th game (26th of December) and Leeds after the 300th game (28th of March). Blackburn started very well but showed a decline between games 300 and 360 (28th of March and 2nd of May). The main competitor of Blackburn, West Ham, had 61% of getting a place between 1-6 on 25th of April (after 350th game).

For the relegation places, only Barnsley was clearly worse than the other teams. Crystal Pallace started really well and only after the 25th batch showed high probability of relegation. Bolton, Everton and Tottenham were the main competitors for the third relegation place.

6.2 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.

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). Fiorentina, Lazio, Roma and Udineze are found in places 3-6 getting the UEFA cup places. Note that Lazio won the cup and therefore it had no interest in earning one of these places. Parma, which got one of the UEFA places, is ranked at the seventh place according to the average points. The four worst teams according to the average points of the simulated leagues are Vicenza, Atalanta, Lecce and Napoli. The three of them were actually relegated while Vicenza finally avoided relegation. Brescia, which was finally relegated instead of Vicenza, has the 13th place on the average points ranking with 5 points more than the latter. This may indicate that Brescia did not deserve to be relegated and was better than other competitors. Three teams had considerably higher probabilities of relegation: Napoli, Lecce and Atalanta had 98%, 87% and 64% respectively. Probabilities for the forth relegation place are mainly divided between Vicenza (51%), Brecia (24%) and Piacenza (15%).

Analysis assuming availiability 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 raized 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 proability of 44% to Juventus and 27% to Inter. Table 5

The final conclusion is that Juventus was the best team for 1997-98 season in Italy. Moreover, Inter was undoubtelly a very good opponent and could have won the league in any small error of Juventus. We may also add that Lazio was a really good team deserving a higher position in the league. Winning the cup has reduced its motivation for chasing a higher position. Finally, Brescia did not deserve to be relegated and seems that it was a much better team than Vicenza, which avoided relegation.

6.3 Overdispesion and Negative Binomial Model

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

Fitting the negative binomial model to English data with regressors similar to (6) 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 ($\hat{\theta} = 849.6 \pm 1251.0$). Model paramater 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.

7 Discussion

Statistical modelling of soccer games is a real chalenge since it introduces statistics in every-day 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 we calculated correlations between the final ranking and the number of goals scored and conceded by each team 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, we 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 we 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 English, Italian of season 1997-98 can satisfactory interpret underlined structures of soccer data. Adding other possible regressors such as red cards, injuries or psychological factors can improve the model fit and prediction of future results.

Acknowledgements

The authors would like to thank Professor Eudokia Xekalaki and Professor Petros Dellaportas for their comments on an earlier version of this paper.

Address for Correspondence

Dr. Dimitris Karlis, Department of Statistics, Athens University of Economics and Business, 76 Patission Street, 10434, Athens, Greece. Electronic mail: karlis@stat-athens.aueb.gr.

References

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

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

[3] 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.

[4] 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.

[5] BENNET, J. (1998). *Statistics in Sport.* First Edition, London: Edward Arnold.

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

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

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

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

[10] 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.

[11] GLICKMAN, M.E. (1995). A Comprehensive Guide to Chess Ratings. *American Chess Journal*, **3**, 59–102.

[12] GLICKMAN, M.E. and STERN, H.S. (1998). A State-space Model for National Football League Scores. *Journal of the American Statistical Association*, **93**, 25-35.

[13] HASSELBLAD, V. (1994). Meta-analysis in Enviromental Statistics. *Handbook of Statistics*, **12**, 691–716.

[14] IRWIN, W. (1937). The Frequency Distribution of the Difference Between Two Poisson Variates Following the Same Poisson Distribution. *Journal of the Royal Statistical Society A*, **100**, 415–416.

[15] JACKSON, D.A. (1994). Index Betting on Sports. *The Statistician*, **43**, 309-315.

[16] JOHNSON, N.L., KOTZ, S. and KEMP, A.W. (1992). *Univariate Discrete Distributions*. Second Edition, New York: John Wiley & Sons.

[17] KOCHERLAKOTA, S. and KOCHERLAKOTA, K. (1992). *Bivariate Discrete Distributions*. New York: Marcel Dekker.

[18] 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.

[19] KUONEN, D. (1997a). Statistical Models for Knock-out Soccer Tournaments. *Technical Report*, Department of Mathematics, Chair of Applied Statistics, Ecole Polytechnique Federale De Lausanne.

[20] KUONEN, D. (1997b). Modelling the Success of Football Teams in the European Championships (in French). *Technical Report*, Department of Mathematics, Chair of Applied Statistics, Ecole Polytechnique Federale De Lausanne.

[21] LEE, A.J. (1997). Modeling Scores in the Premier league: Is Manchester United Really the Best? *Chance*, **10(1)**, 15-19.

[22] MAHER, M.J. (1982). Modelling Association Football Scores. *Statistica Neerlandica*, **36**, 109–118.

[23] MOSTELLER, F. (1997). Lessons from Sports Statistics. *The American Statistician*, **51**, 305-310.

[24] POLLARD, R. and REEP, C. (1997). Measuring the Effectiveness of Playing Strategies at Soccer. *The Statistician*, **46**, 541-550.

[25] 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.

[26] REEP, C. and BENJAMIN, B. (1968). Skill and Chance in Association Football. *Journal of the Royal Statistical Society A*, **131**, 581-585.

[27] REEP, C., POLLARD, R. and BENJAMIN, B. (1971). Skill and Chance in Ball Games. *Journal of the Royal Statistical Society A*, **134**, 623-629.

[28] RUE, H. and SALVENSEN, O. (1997). Predicting Soccer Matches in a League. *Technical Report*, Department of Mathematical Sciences, Norwegian University of Science and Technology, Norway.

[29] SZYMANSKI, S. and SMITH, R. (1997). The English Football Industry: Profit, Performance and Industrial Structure. *Internenational Review of Applied Economics*, **11**, 135–153.

[30] SKELLAM, J.G. (1946). The Frequency Distribution of the Difference Between Two Poisson Variates Belonging to Different Populations. *Journal of the Royal Statistical Society A*, **104**, 296.

[31] STEFANI, R.T. (1983). Observed Betting Tendencies and Suggested Betting Strategies for European Football Pools. *The Statistician*, **32**, 319–329.

[32] VENABLES, W.N. and RIPLEY, B.D. (1994). *Modern Applied Statistics with S-Plus.* New York: Springer Verlag.

[33] WRIGHT, D. (1997). Football Standings and Measurement levels. *The Statistician*, **46**, 105-110.

List of Tables

1 Crosstabulation of Home and Away Goals Using Data from 24 European Leagues. 19

2 Model Selection Details for English and Italian League 20

3 English Premier Division: Final League Table Details, Estimated Model Parameters and Average Points of Simulated Leagues. 21

4 English Premier Division: Percentages of Final League Ranking 22

5 Percentages of Winning the 'Italian Serie A' League per Week for Juventus and Inter 23

6 Dispersion Index (DI) for Goals Scored and Conceded by Each Team. 24

List of Figures

1 Histogram of Variance over Mean Ratio Observed Values. 25

2 Variance over Mean Ratio for the 456 Teams 26

3 English Premier Division: Percentages of Winning the League per Batch of 10 Games 27

4 English Premier Division: Average Points per Game for Each Batch of 10 Games 28

5 English Premier Division: Predicted Probabilities of European Exit per Batch of 10 Games 29

6 English Premier Division: Average Points per Game for Each Batch of 10 Games 30

7 English Premier Division: Predicted Probabilities of Relegation per Batch of 10 Games 31

8 English Premier Division: Average Points per Game for Each Batch of 10 Games 32

		number of goals scored by the quest team					
		0	1	2	3	4+	Total
goals scored by the home team	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 Home and Away Goals Using Data from 24 European Leagues.

	Model	Removed Term	AIC English	AIC Italian
1	`Saturated`		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: Model Selection Details for English and Italian 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 Und.	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).

	Team	Final Rank							
		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 Und.	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).

	Team	Week 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).

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. Und.	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.

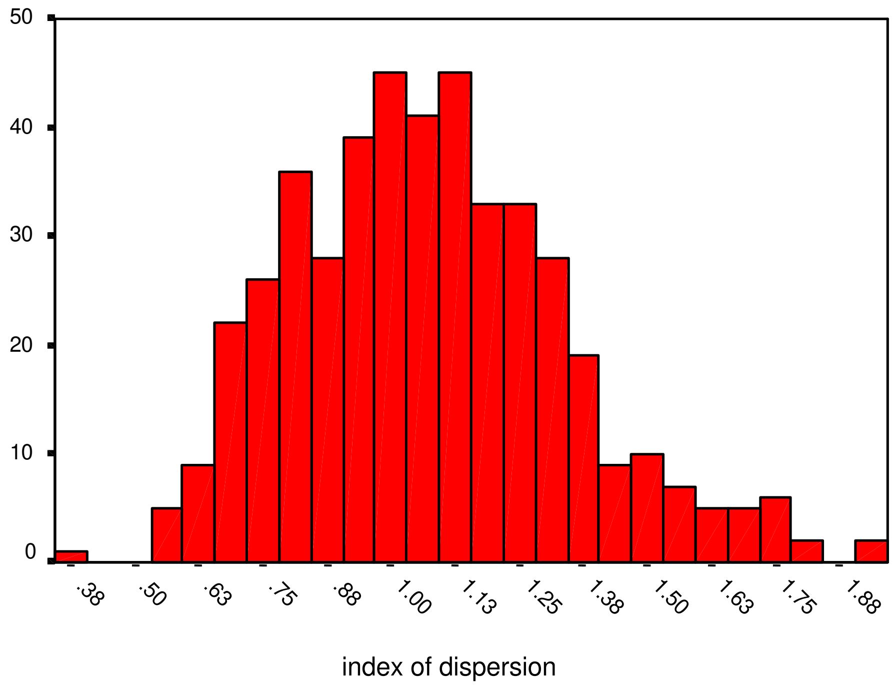

Figure 1: Histogram of Variance over Mean Ratio Observed Values.

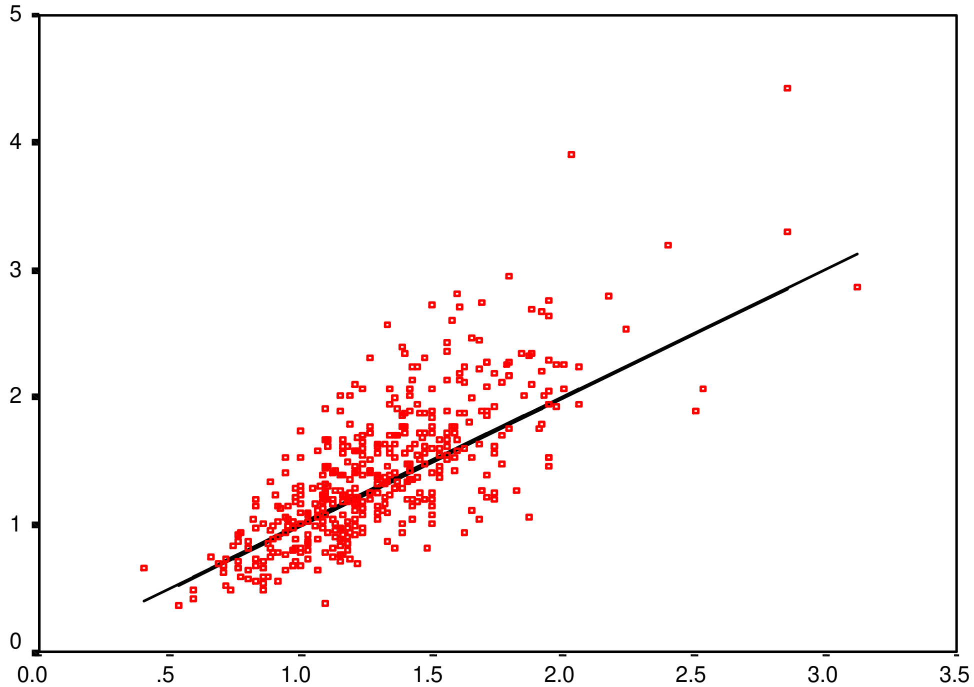

Figure 2: Variance over Mean Ratio for the 456 Teams (Straight Line Indicates the Poisson Distribution).

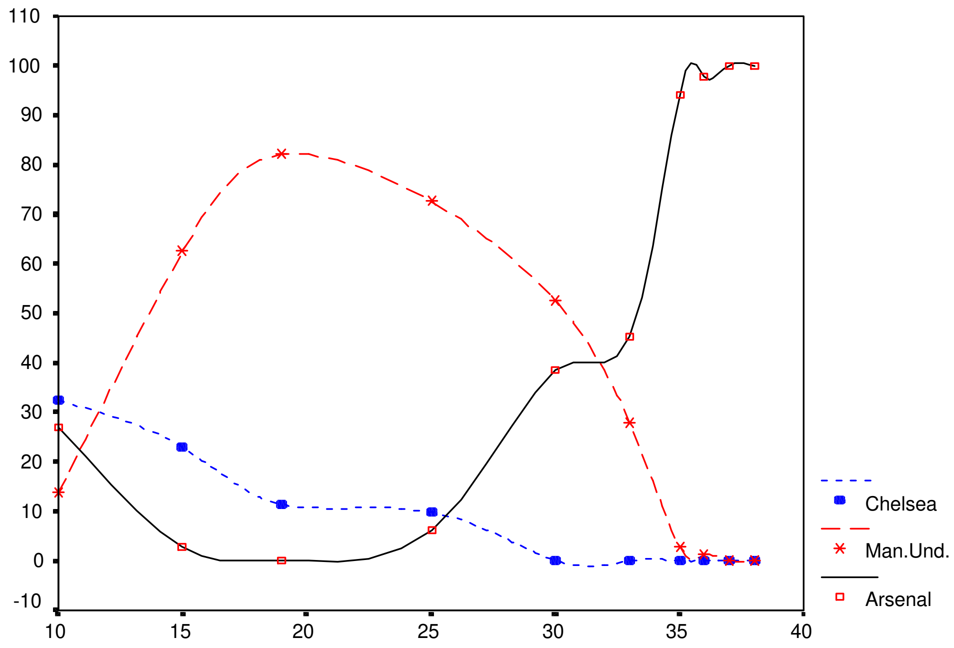

Figure 3: English Premier Division: Percentages of Winning the League per Batch of 10 Games

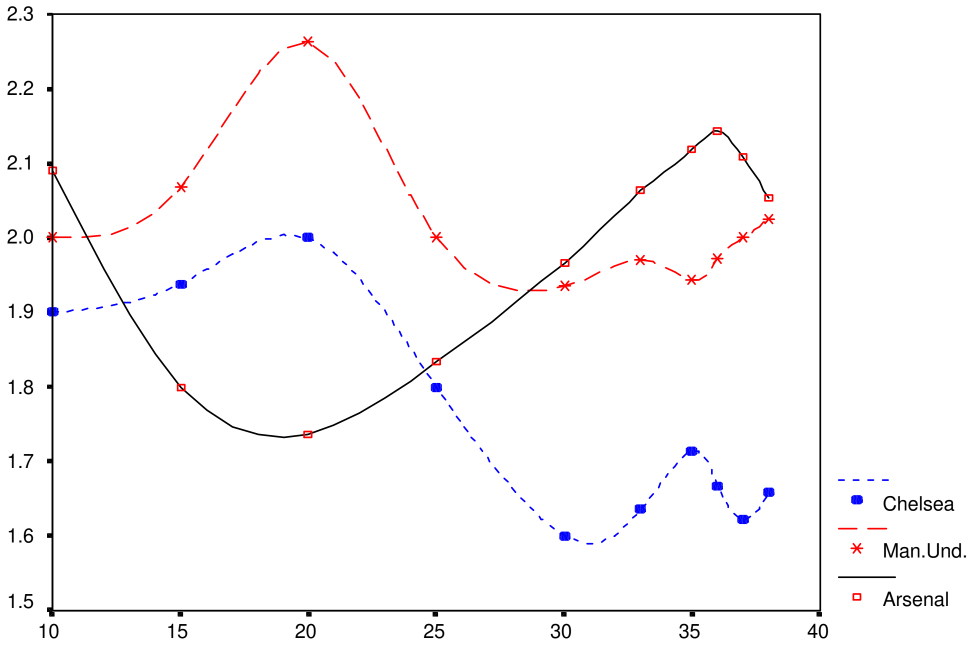

Figure 4: English Premier Division: Average Points per Game for Each Batch of 10 Games

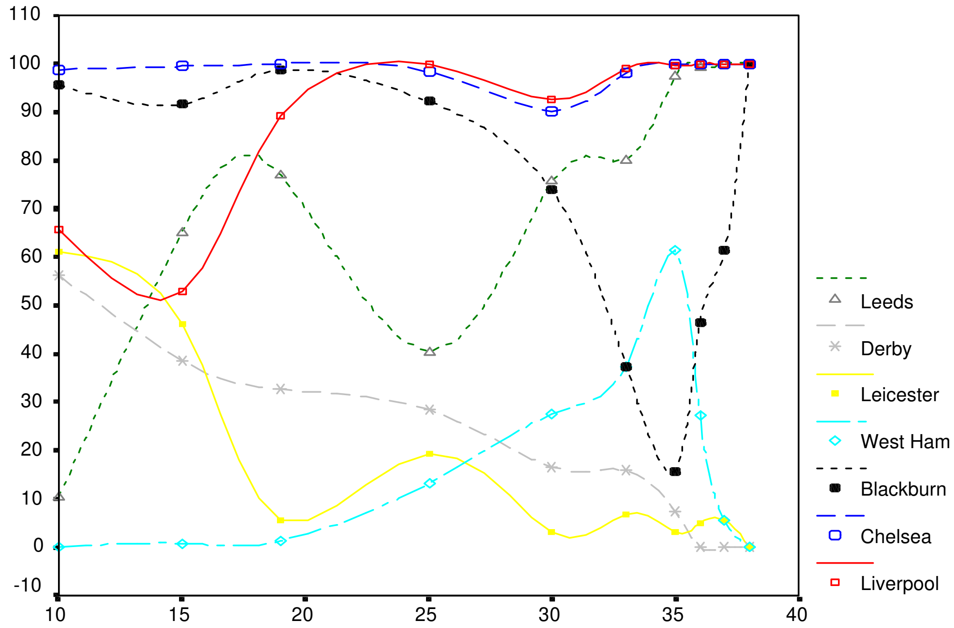

Figure 5: English Premier Division: Predicted Probabilities of European Exit per Batch of 10 Games

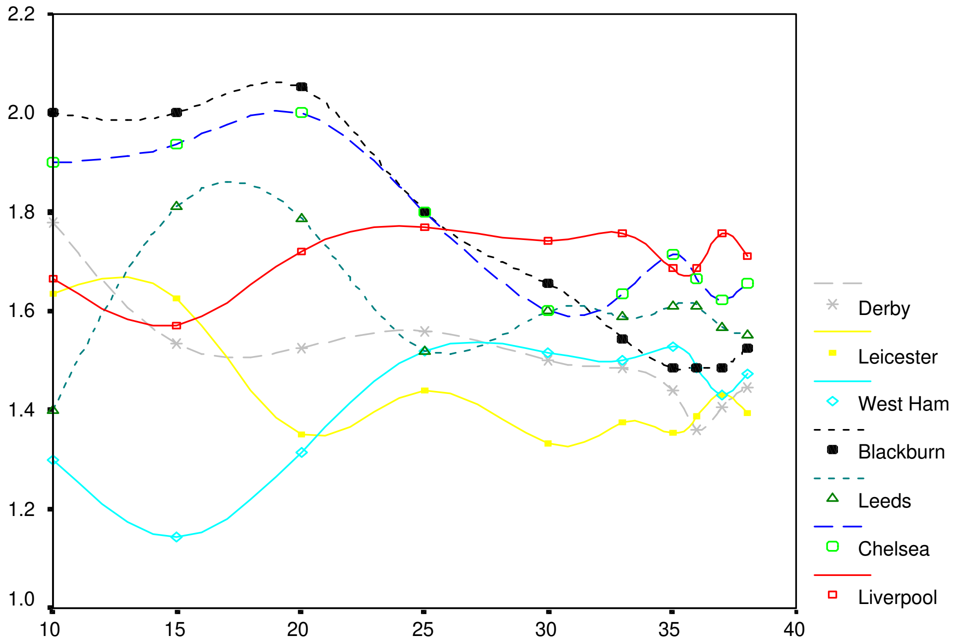

Figure 6: English Premier Division: Average Points per Game for Each Batch of 10 Games

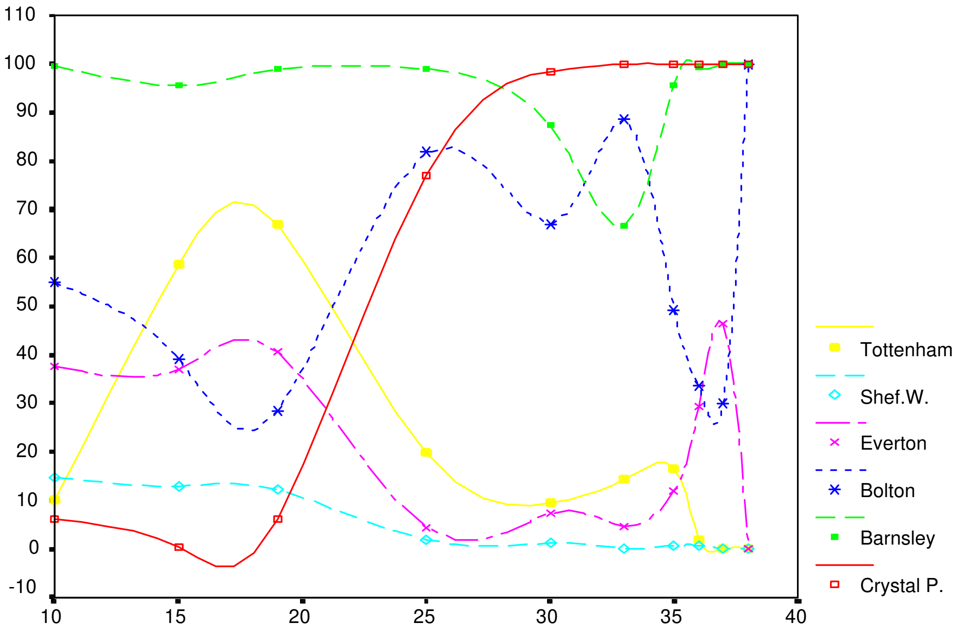

Figure 7: English Premier Division: Predicted Probabilities of Relegation per Batch of 10 Games

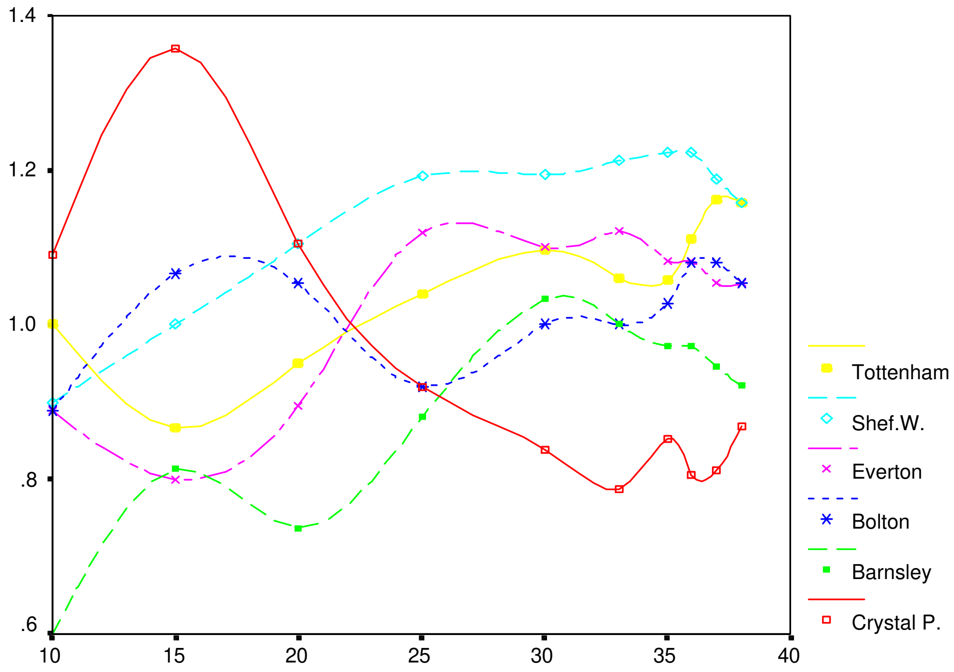

Figure 8: English Premier Division: Average Points per Game for Each Batch of 10 Games