DE GRUYTER
OPEN

Croatian First Football League: Teams' performance in the championship

Dušan Mundar
Faculty of Organization and Informatics, University of Zagreb, Varaždin, Croatia
dusan.mundjar@foi.hr

Diana Šimić
Faculty of Organization and Informatics, University of Zagreb, Varaždin, Croatia
diana.simic@foi.hr

Abstract
The goal of our research was to use simulation modelling for prediction of the Croatian First Football League seasonal ranking and analyse variation in teams' performance during a season. We have developed a model of the number of goals scored by a team in a match based on the Poisson distribution. Parameters of the model were estimated from the data on consecutive matches in a season. Variation in a team's performance was modelled as a moving parameter estimate. The final rankings were predicted from 1000 simulation runs of the second part of the season based on parameter estimates from the first part of the season. For each team the most frequent outcome of the simulation defined the team's rank. The method was tested on seasons 2014/15 and 2015/16. Prediction was correct for six teams in the season 2014/15 and five teams in the season 2015/16. Proposed methods enable dynamic monitoring of a team's performance and prediction of final rankings during the season. An advantage of the prediction method is that in addition to predicting the final ranking it also estimates probabilities of alternative positions.

Keywords: football, performance monitoring, Poisson distribution, predictive analytics, simulation, sports analytics.

JEL classification: C15, C51, L83, Z20.
DOI: 10.1515/crebss-2016-0006

Received: May 25, 2016
Accepted: August 24, 2016

Introduction
Our aim was to develop methodology to dynamically monitor teams' performance and predict final rankings of teams in a league after the first half of the season is played and apply it to the Croatian football league. There is a wide range of benefits that such an analysis can provide. Teams' management can gain more objective assessment on (relative) strength of their team. Applied to a sliding window of matches during a season results can also signal decrease in team's quality that cannot be attributed to natural variation.

Models in current literature are becoming complex and require large amount of data. On the other hand, Croatian league is relatively small (10 teams, 180 matches per season). Teams' performance monitoring and prediction of final rankings had to

be based on final scores of ca. 90 matches. This imposed a strong restriction on the feasible number of parameters (i.e. degrees of freedom of the model). Therefore, we had to make some strong assumptions. We assume that number of goals scored by a team is a Poisson random variable with a constant parameter during the season, modified only by home advantage. Numbers of goals scored by two teams in a match are assumed to be independent. It is interesting to determine whether such a simple model is rich enough to enable prediction of the final rankings in Croatian Football League from data on only a part of a season.

The rest of the paper is structured as follows. The next section presents literature review and some directions of further development in the area of football match score prediction. Research Methodology section introduces basic information on the Croatian Football League and the structure of data as well as methodology. The fourth section provides results for seasons 2014/15 and 2015/16 with discussion. Main findings are summarised in Conclusions.

Literature review

Sports analytics and predictive analytics captured the attention of both scientific and general audience after the appearance of the book Moneyball (Lewis, 2004) and the subsequent movie (Moneyball, 2011). They display now famous results of baseball team Oakland Athletics' attributed to use of statistical analysis of the game and of players' performance. Football analytics still do not get such attention of general audience; however, they have been around for more than 30 years.

The first models of football results were published in the 80's (e.g. Maher, 1982). Modelling number of goals scored by a team as a Poisson random variable was one of the first approaches. Further developments include prediction of league results (e.g. Lee, 1997) and use of bivariate distribution (e.g. Karlis and Ntzoufras, 2003). More recently, Constantinou et al. (2012) demonstrate use of a Bayesian network model for forecasting of match outcomes, Arabzad et al. (2014) use artificial neural networks to predict outcomes, and Dobravec (2015) demonstrates an approach for forecasting world cup results using a matrix-factorization model. Models in current literature are becoming more complex and require large number of parameters (e.g. team strength in attack and defence, separate estimate of home effect per team, etc.) which also requires structure

Research methodology

Data

Data on Croatian Football League used in the paper and available at www.rezultati.com. They consist of games schedules and number of goals scored per team in each of the games played in seasons 2014/15 and 2015/16. The first Croatian Football League comprises 10 teams. During a season, every team meets four times with each other team, twice on home and twice on the opponent's stadium. Schedule of games is known before the season starts. Teams score 3 points for every win and 1 point for every draw. Final ranking is based on the total number of points scored during the season.

Tables 1 and 2 present number of games played, total number of goals scored by a team and rankings at the end of the winter part of the Season 2014/15 (Table 1) and Season 2015/16 (Table 2).

Number of games played in the winter part of a season was 95 for the Season 2014/15, and 105 for the Season 2015/16. Data on scores of these games were used for prediction of final ranking for the seasons.

Methods

We have modelled number of goals scored by each team as a Poisson random variable:

$$P(X = k) = \frac{\lambda_t^k}{k!} e^{-\lambda_t}, \; k = 0,1,\ldots, \tag{1}$$

Parameter λ_t depends on number of goals achieved by a team in previous matches. Since number of goals is larger when team plays home, and smaller when team plays away, we have added a home-bonus factor *h*. If A is the team playing home and B is the team playing away then:

$$\begin{aligned} \lambda_A &= \lambda(1+h)(1+t_A), \\ \lambda_B &= \lambda(1-h)(1+t_B), \end{aligned} \tag{2}$$

where λ is an average number of goals scored per team in all matches, t_A and t_B are team specific factors.

Goals in a match where home team A is playing against away team B were modelled independently. If X was number of goals scored by team A and Y was number of goals scored by team B probability of score *n:m* was:

$$P(X = n, Y = m | \lambda_A, \lambda_B) = P(X = n | \lambda_A) P(Y = m | \lambda_B), \tag{3}$$

where X and Y were Poisson random variables with parameters λ_A and λ_B.

Table 1 Rankings at the end of the winter part of the Season 2014/15

Team	Games Played Home	Games Played Away	Goals Scored Home	Goals Scored Away	Points Scored
Din. Zagreb	10	9	30	18	42
Rijeka	9	10	29	14	33
Hajduk Split	10	9	23	16	32
Lok. Zagreb	9	10	17	19	31
RNK Split	9	10	14	8	21
Slaven Belupo	9	10	12	6	19
Osijek	10	9	13	5	18
NK Zagreb	10	9	14	11	18
Istra 1961	9	10	8	11	16
Zadar	10	9	14	4	9

Table 2 Rankings at the end of the winter part of the Season 2015/16

Team	Games Played Home	Games Played Away	Goals Scored Home	Goals Scored Away	Points Scored
Din. Zagreb	11	10	27	14	45
Rijeka	10	11	21	14	43
Hajduk Split	10	11	15	14	39
RNK Split	11	10	16	6	32
Lok. Zagreb	10	11	18	16	30
Int. Zaprešić	11	10	11	8	22
Slaven Belupo	10	11	13	11	21
Istra 1961	11	10	13	6	19
Osijek	10	11	8	7	19
NK Zagreb	11	10	10	6	9

Home-bonus *h* was estimated as $h = \lambda_h/\lambda - 1$, where λ_h was the average number of goals scored per home team in all matches. Team specific factor *t* for a team that played *n* matches at home and *m* matches away was estimated as $t = g/e - 1$, where *g* was average number of goals scored by the team and e was expected number of goals scored by an average team that played *n* matches at home and *m* matches away, given by:

$$e = \frac{\lambda_h n + \lambda_a m}{n + m}. \quad (4)$$

We have estimated Poisson parameters λ_h, λ_a, home-bonus h, and team specific factors from available data. For the purpose of monitoring teams' performance, we have estimated the parameters on a sliding window of consecutive matches. For prediction of the final ranking, we have used parameter estimates from the first part of the season for simulation of match results for the rest of the season. Based on results of the simulation we have allocated points to the teams and generated a ranking. Simulations were run 1000 times. For each team probability of each position in final ranking was estimated as a proportion of simulation runs resulting with the team in that position.

Prediction accuracy for Seasons 2014/15 and 2015/16 were compared with accuracy of naïve predictions based on ranking after the winter part of the seasons using Euclidean distance between predicted and actual rankings.

Variation in a team's performance was modelled by moving parameter estimates. For each team specific factors and home-bonus were calculated in moving time-frames of five consecutive games. Expression (2) was used to estimate teams' expected average number of goals per game, without correction for home advantage, over moving time frames. These estimates were used to monitor teams' performance dynamics.

Results

Season 2014/15

During the Season 2014/15, total number of games played was 180. In the winter part of the Season, there were 19 rounds with the total of 95 matches. Average number of goals scored per team (λ) was 1.505, average number of goals scored per home team (λ_H) was 1.832, and average number of goals scored per away team (λ_A) was 1.179. Home-bonus *h* was 0.216, i.e. teams achieved 21.6% more goals per game when playing home then in general.

Estimates of teams' parameters are presented in Table 3. It is interesting to compare data in Table 1 and parameter estimates in Table 3. For the upper part of the table (the best four teams), their order in respect to the total number of points after the winter part of the Season corresponded to the order in respect to Poisson lambdas. For the middle and lower part of the table, these orders were not equal. Teams with higher number of goals scored tended to have higher lambdas then we would have expected based only on point rankings. Therefore, we expected to find a difference between a naïve prediction based on interim ranking after the winter part of the Season and our prediction based on simulation.

Table 4 presents results of the simulation, along with rankings after the winter part of the Season and the final ranking of the Season. It shows that final ranking of teams is highly predictable. Six of the teams had final ranking position that was predicted by the model as the most likely. Slaven Belupo, Osijek and Istra 1961 had probability of the most likely position lower than 0.5. For these teams predicted final position missed achieved final position by one. The only surprise was RNK Split for which the

most likely position had probability above 0.5. Their achieved final position was equal to the position predicted as the second most likely. Euclidean distance between predicted and achieved final ranking was 2.00. In comparison, Euclidean distance between naïve prediction (using interim ranking after the winter part of the Season as prediction of the final ranking) and the final ranking was 2.83. Maximal deviation of naïve method was two, and for our model, it was one.

Table 3 Poisson parameters λ for the winter part of the Season 2014/15

Team	λ	λ_h	λ_a
Din. Zagreb	2.498	3.039	1.956
Rijeka	2.289	2.786	1.793
Hajduk Split	2.029	2.469	1.590
Lok. Zagreb	1.917	2.332	1.501
NK Zagreb	1.301	1.583	1.019
RNK Split	1.171	1.425	0.917
Istra 1961	1.012	1.231	0.792
Slaven Belupo	0.958	1.166	0.751
Osijek	0.937	1.140	0.734
Zadar	0.937	1.140	0.733

Legend:
λ – Poisson parameter for play at neutral stadium
λ_h – Poisson parameter for home game
λ_a – Poisson parameter for away game

Table 4 Results of the simulation for Season 2014/15 with interim and final rankings

Team	Most likely position (probability)	2nd most likely position (probability)	3rd most likely position (probability)	Naïve ranking	Final ranking
Din. Zagreb	1 (0.973)	2 (0.027)	-	1	1
Rijeka	2 (0.980)	1 (0.016)	3 (0.004)	2	2
Hajduk Split	3 (0.833)	4 (0.164)	2 (0.003)	3	3
Lok. Zagreb	4 (0.867)	3 (0.128)	5 (0.005)	4	4
NK Zagreb	5 (0.833)	6 (0.147)	7 (0.017)	7	5
Slaven Belupo	7 (0.407)	8 (0.286)	6 (0.151)	6	6
RNK Split	6 (0.597)	7 (0.214)	5 (0.109)	5	7
Osijek	9 (0.404)	10 (0.277)	8 (0.220)	8	8
Istra 1961	8 (0.356)	9 (0.238)	7 (0.224)	9	9
Zadar	10 (0.658)	9 (0.240)	8 (0.080)	10	10

Legend:
Naïve ranking – ranking after the winter part of the Season
Final ranking – ranking at the end of the Season

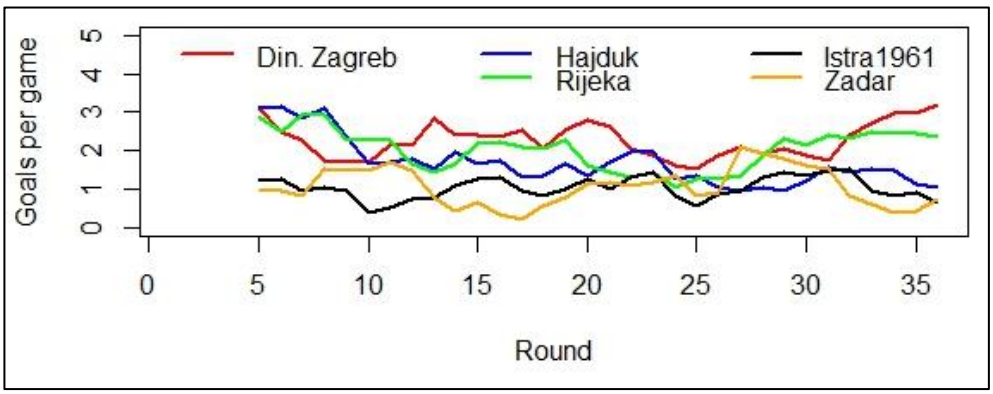

Figure 1. Expected number of goals per game by teams in Season 2014/2015 based on the time frame of last five games

Figure 1 presents variation in teams' performance for some of the teams in the Season 2014/15. The analysis indicates that teams that ended in the last two positions had decreased their performance in the final rounds. Three top teams had some decline in performance in the middle of the Season, but two of them (Din. Zagreb and Rijeka) improved it in the finals.

Season 2015/16

During the winter part of the Season 2015/16 there were 105 matches played in 21 rounds. Average number of goals scored per team (λ) was 1.209, average number of goals scored per home team (λ_h) was 1.448, and average number of goals scored per away team (λ_a) was 0.971. Home-bonus h was 0.197 i.e. teams achieved 19.7% more goals per game when playing home and 19.7% less goals per game when playing away then when playing a neutral game.

Expected number of goals scored by an average team (e) for teams that played 11 games at home and 10 games away was $25.638/21 = 1.221$. Expected number of goals scored by a team that played 10 games home and 11 games away was $25.162/21 = 1.198$. These numbers were used to calculate team specific factors. For example, RNK Split scored 22 goals in winter part of the Season, playing 11 games home and 10 games away, so it's team specific factor $t = ((22/(11+10)/1.221) - 1 = -0.142$.

Table 5 presents Poisson parameters for all teams and different match scenarios (neutral ground, home, away). It is interesting to notice differences in team rankings based on total number of points after the winter part of the Season (Table 2) and those based on estimates of Poisson parameters (Table 5). For instance, Lokomotiva Zagreb and Slaven Belupo ranked higher in Table 5, and RNK Split ranked lower.

Table 5 Poisson parameters λ for the winter part of the Season 2015/16

Team	λ	λ_h	λ_a
Din. Zagreb	1.934	2.315	1.553
Rijeka	1.682	2.014	1.351
Lok. Zagreb	1.634	1.956	1.313
Hajduk Split	1.394	1.668	1.120
Slaven Belupo	1.154	1.381	0.927
RNK Split	1.038	1.242	0.834
Int. Zaprešić	0.896	1.073	0.720
Istra 1961	0.896	1.073	0.720
NK Zagreb	0.755	0.903	0.606
Osijek	0.721	0.863	0.579

Legend:
λ – Poisson parameter for play at neutral stadium
λ_h – Poisson parameter for home game
λ_a – Poisson parameter for away game

The second part of the Season 2015/16 comprised 15 rounds with a total of 75 games. Maximum number of points that a team could gain during the second part of the Season was 45. Table 6 shows three most likely positions for each team with estimated probabilities based on simulation. We predicted RNK Split to drop by one position, but they ended two positions lower than in interim ranking. Lokomotiva Zagreb improved its position as our model suggested. Inter Zaprešić significantly improved their performance in the second part of the season and was better than predicted. Osijek also significantly improved performance in the final 10 games (losing only 3 games, unexpected from a team near the bottom of the table).

Overall achieved final position was equal to the predicted most likely position for five teams. Euclidean distance between our prediction and the final ranking was 2.83, same as Euclidean distance between naïve prediction (using interim ranking after the winter part of the Season as prediction of the final ranking) and the final ranking. Maximal deviation of both naïve method and our model was two.

Table 6 Results of the simulation for Season 2015/16 with interim and final rankings

Team	Most likely position (probability)	2nd most likely position (probability)	3rd most likely position (probability)	Naïve ranking	Final ranking
Din. Zagreb	1 (0.856)	2 (0.139)	3 (0.005)	1	1
Rijeka	2 (0.757)	1 (0.134)	3 (0.099)	2	2
Hajduk Split	3 (0.635)	4 (0.252)	2 (0.094)	3	3
RNK Split	5 (0.785)	6 (0.165)	4 (0.033)	4	6
Lok. Zagreb	4 (0.704)	3 (0.257)	5 (0.028)	5	4
Int. Zaprešić	7 (0.445)	8 (0.292)	6 (0.151)	6	5
Slaven Belupo	6 (0.616)	7 (0.184)	5 (0.152)	7	7
Istra 1961	8 (0.438)	7 (0.292)	9 (0.198)	8	9
Osijek	9 (0.663)	8 (0.223)	7 (0.068)	9	8
NK Zagreb	10 (0.954)	9 (0.040)	8 (0.006)	10	10

Legend:
Naïve ranking – ranking after the winter part of the Season
Final ranking – ranking at the end of the Season

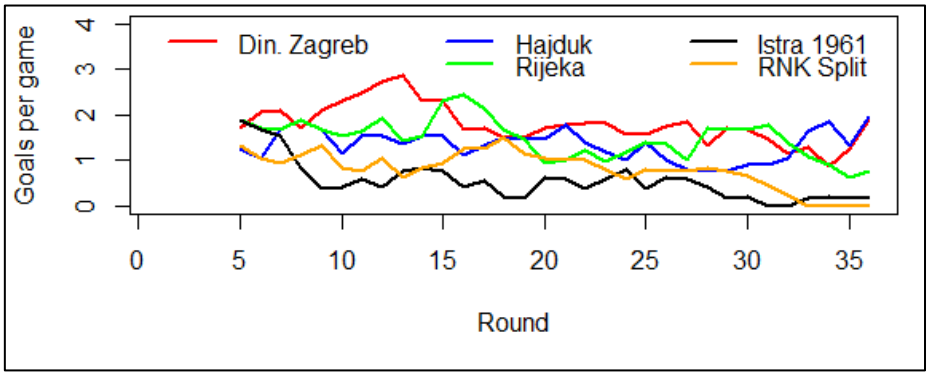

Figure 2 Expected number of goals per game by teams in Season 2015/16 based on time frame of last five games

Figure 2 presents variation in a team's performance for some of the teams in Season 2015/16. RNK Split team, which ended two positions lower than expected, had continuous decrease in performance in the second half of the Season. Istra 1961 reduced its performance during the second part of the Season as well, leading to its penultimate position.

Conclusions

We have introduced a simple probabilistic model of football match scores based on Poisson distribution. Parameters of the model were estimated from data on match scores in the first part of the season, and we used them to simulate rest of the season and the final rankings. Model was tested on data for Seasons 2014/15 and 2015/16. Prediction from the model was better than naïve prediction (i.e. extrapolation of the interim ranking after the first part of the Season). In Season 2015/16, even though Dinamo and Rijeka had a similar number of points at the end of the winter part of the Season (Dinamo 45, Rijeka 43) there was a large difference in their probabilities of

winning the championship. On the other hand, analysis correctly pointed out that NK Zagreb was at a high risk of ending the Season in the last position. Advantage of this method is that in addition to providing prediction for the final ranking it also provides probabilities of alternative positions. For teams that have large difference between the probabilities of the most likely and the second most likely position, we may have high confidence in the prediction of their final position, under the assumption that there are no changes in the relative quality of play of all teams. However, for teams for which probabilities of alternate positions are not very different, confidence of prediction is lower.

Demonstrated dynamical assessment of performance during a season could be used for detecting change in performance or validating effects of changes introduced in play. Such analyses could become a valuable tool for the team-management decision support. Model can be used for a priori impact assessment by running simulations of different management strategies based on their expected effects on match results.

In order to accommodate small amount of available data we had to make strong assumptions. This model does not accommodate variations in teams' performance. However, due to a small number of parameters it is possible to analyse part of the season, and thus monitor changes in teams' performances. It is also assumed that home advantage is the same for all teams and locations. In order to accommodate for variations in home advantage effects we would have to base our prediction on a larger set of matches. Similar argument could be put forward regarding independence of number of goals scored by the two teams in a match. Thus, we had to make a trade-off between model complexity and size of data. We show that even with these limitations it is possible to achieve reasonable prediction.

References

1. Arabzad, A.C. (2014). Football Match Results Prediction Using Artificial Neural Networks: The Case of Iran Pro League. International Journal of Applied Research on Industrial Engineering. Vol. 1, No. 3, pp 159–179.
2. Constantinou, A.C., Fenton, N.E., Neil, M. (2012). pi-football: A Bayesian network model for forecasting Association Football match outcomes. Knowledge-Based Systems, 36, pp. 322-339.
3. Dixon, M.J., Coles, S.G. (1997). Modelling Association Football Scores and Inefficiencies in the Football Betting Market. Journal of the Royal Statistical Society: Series C (Applied Statistics). Vol. 46, No. 2. pp. 265-280.
4. Dixon, M.J., Robinson, M. (1998). A birth process model for association football matches. Journal of the Royal Statistical Society: Series D (The Statistician), Vol. 47, No. 3, pp. 523–538.
5. Dobravec (2015). Forecasting the football world cup results using a matrix-factorization model. Elektrotehniški Vestnik, Vol. 82, No 1, pp 61–65.
6. Hill, I.D. (1974). Association football and statistical inference. Applied statistics, Vol. 23, pp. 203-208.
7. Karlis, D., Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. Journal of the Royal Statistical Society: Series D (The Statistician), Vol. 52, No. 3, pp. 381–393.
8. Koopman S.J., Lit, R. (2015). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League, Journal of the Royal Statistical Society: Series A (Statistics in Society), Vol. 178, No. 1, pp. 167–186.
9. Lee A. J. (1997). Modeling scores in Premier League: is Manchester United really the best. Chance, Vol. 10, pp. 15-19.
10. Maher M.J. (1982). Modelling association football scores. Statistica Neerlandica, Vol. 36, pp. 109–118.

11. Moneyball (2011). Directed by Bennett Miller [Film]. USA: Columbia Pictures.
12. Moroney, M.J. (1951). Facts from figures. London: Penguin
13. Muller, H-G., Stadtmuller, U. (2005). Generalized Functional Linear Models. The Annals of Statistics, Vol. 33 No. 2, pp. 774–805.
14. Munđar, D., Šamarija, D. (2016). Primjena simulacijskih modela za prognoziranje rezultata u bowlingu, Poučak, Vol. 64, pp. 73-79.
15. Munđar, D., Šimić, D. (2016) Croatian First Football League: Prediction of teams' ranking in the championship, Proceedings of the ISCCRO - International Statistical Conference in Croatia. Zagreb: Croatian Statistical Association, Vol 1, No. 1, pp. 211-217.
16. Owen, A.J. (2011). Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. IMA Journal of Management Mathematics, Vol. 22, No. 2, pp. 99-113.
17. Rotshtein, A., Posner, M., Rakytyanska, H. (2005). Prediction of results of Football games base of Fuzzy model with genetic and neuro tuning. Cybernetics and Systems Analysis, Vol. 41, No. 4, pp. 619-630.
18. Rue, H., Salvesen, O. (2000). Prediction and Retrospective Analysis of Soccer Matches in a League, Journal of the Royal Statistical Society. Series D (Statistician), Vol. 49, No. 3, pp. 399-418.
19. Sports Analytics Group, University of Toronto. Introduction to Sports Analytics, http://sportsanalytics.sa.utoronto.ca/2014/12/11/introduction-to-sports-analytics/, [03 January 2016]

About the authors

Dušan Munđar is a PhD student at Faculty of Organization and Informatics, University of Zagreb. His primary interests are in application of statistical and mathematical models in management and finance. He graduated in mathematical statistics and computing at the Faculty of Science, University of Zagreb, earned a postgraduate degree in management of business systems and a postgraduate degree in actuarial mathematics. He is an expert in public procurement specializing in methods for selection of economically most advantageous tender. Author can be contacted at *dusan.mundjar@foi.hr*.

Diana Šimić is a professor of research methods and statistics at Faculty of Organization and Informatics, University of Zagreb. She is a secretary general of the Croatian Biometric Society and a member of the American Statistical Association. She has authored more than 30 papers in journals, 20 papers in conference proceedings and 9 scientific monographs. Author can be contacted at *diana.simic@foi.hr*.