Estimation of Tournament Metrics for Association Football League Formats

Jan Lasek[1] and Marek Gągolewski[2,3]

[1] Institute of Computer Science, Polish Academy of Sciences,
ul. Jana Kazimierza 5, 01-248 Warsaw, Poland

[2] Warsaw University of Technology, Faculty of Mathematics and Information Science,
ul. Koszykowa 75, 00-662 Warsaw, Poland

[3] Systems Research Institute, Polish Academy of Sciences,
ul. Newelska 6, 01-447 Warsaw, Poland

Abstract. In this paper we study league design in terms of its ability to determine the strongest team as the winner of the competition. We compare a double round-robin tournament with a two-stage league format recently introduced in the Polish top division of association football. Based on simulations of a theoretical competition model with latent dynamic team strength parameters, we conclude that the new league format has better ability to produce the strongest team as the champion at the end of the season.

1 Introduction

The design of a tournament as several dimensions. As a basic requirement, the tournament should be scheduled so as to each involved team has the same chances to win it. On the other hand, one would like that the tournament were stable, so that it effectively produces the best team as the winner. On the other hand, some proportion of randomness in the results is allowed and appreciated as this produces the excitement accompanying the most important competitions. Yet another flavour of a design of a contest is the associated economic factors. Certain agreements due to, e.g., television rights impose special requirements on the competition.

The design of sport contests have been of interest of authors in various research areas. For example, Appleton [2] compares different competition formats according to their ability to indicate as a winner the best team involved in the competition. In a related study, Scarf et al. [9] examine different (also non-standard) tournament formats of the Champions League for association football and compare them according to several dimensions. The authors propose several so-called tournament metrics which aim to measure predictive efficacy of a contest. Ryvkin [8] investigates three popular competition formats – a contest, a binary elimination tournament and a round-robin tournament. Due to high complexity of the problem under study, the authors in their methodology employ simulations to determine different tournament metrics according to which they are later compared. All of these papers conclude that the round-robin format is the most effective to produce as the winner the best entrant of the competition. However, it requires relatively large number of games to be played. Proportion of the strongest competitor's victories in a series of simulations is one of the most basic and important tournament metrics considered in related studies. Apart from that, in economic literature, Szymanski [10] provides an overview of factors involved in designing a contest, both from the organiser's and participants' perspective. The author provides insights into incentives of the both involved sides in game–theoretic modelling of competition and tournament design. Also financial factors are discussed. This additionally stresses the fact that the discussed tournament design problem has many aspects. In our application, we focus solely on the predictive efficacy of a tournament format.

The aim of this paper is to evaluate two league competition formats for domestic football competition. We compare a double round-robin tournament with the league format introduced in *Ekstraklasa* – the top division of the football competition in Poland – as of the season 2012/2013. For the two tournaments, we estimate the probability of the strongest team win via a number of simulation experiments. Our major focus is the comparison of the predictive ability of the two formats in terms of the strongest team being crowned the champion of the league. This paper extends our previous contribution in this area in which we assumed a static model for football competition [4]. In current exposition, we propose a dynamic model for teams' ratings evolution in a more refined simulation setup.

This paper is structured as follows. In Section 2, we discuss the design of the two league formats being of our interest. In Section 3, we describe how to estimate tournament metrics via a theoretical model of a league in a simulation experiment. In Section 4 we discuss the obtained results. The last section concludes the work.

2 League formats

The majority of leagues in countries belonging to UEFA – the governing body for association football in Europe – operates as a double round-robin tournament. In such a tournament, the teams play against each other twice: home and away. With n teams competing in a league, this gives $2(n-1)$ rounds of play and in total $2 \cdot \binom{n}{2}$ matches.

In the double-stage league the season is divided into two phases. During the first phase, the teams compete as a standard double round-robin tournament. Next, the table is divided into two even groups: championship and relegation group. Moreover, the points gained in the first stage are divided by two (with possible rounding halves up). In the second phase, a single round-robin tournament is played within each group. At this stage, the points for a match are awarded in a standard manner: 3 points for a win, 1 point for a draw and 0 for a loss. This tournament format yields $2(n-1)+n/2-1$ rounds and $2\cdot\binom{n}{2}+2\cdot\binom{n/2}{2}$ matches (with n even).

As far as Polish Ekstraklasa is concerned, since season 2005/06, the league comprises of 16 teams. During seasons 2005/06 – 2012/13 it operated as a double round-robin tournament. There were 30 rounds of play with in total 240 matches. However, as of the season 2013/14, Ekstraklasa operates as a double-stage league described below. In this way, 37 rounds of games are played and in total 296 matches. Note that during the second phase of the season, 7 rounds are played. This means that the teams do not play an even number of games at home and away. The top 4 teams in each of the groups – championship and relegation – play 4 games at home and only 3 games away. This gives those teams some edge due to the home team advantage phenomenon [13–15]. In comparison to other league formats in the countries under UEFA, the league format currently in force in Poland is identical to the one used in Kazakhstan. However, in Kazakh league fewer teams are involved in the competition – 12 compared to 16 in Poland. Different two-stage league designs also operate in, e.g., the Belgian, Dutch or Scottish leagues at the top division level.

3 Comparison between league format in a simulation study

In this section we describe the way we estimate the probability of the strongest team win in a simulation experiment. In consecutive sections we present the components of the model.

3.1 Game outcome model

For the match outcome model we use the ordered logistic regression [1, 5]. The basic assumption behind the model is that each team is characterised by a single latent parameter – its rating or strength. Let r_i, r_j be ratings of two teams i and j, where team i is assumed to be playing at their home ground. Let us denote $d_{ij} = h+r_i-r_j$ as the difference in the team ratings accounted for the home team advantage parameter h [13–15]. According to the model, if $\{H_{ij}, D_{ij}, A_{ij}\}$ is the set if possible outcomes, with H_{ij} and A_{ij} denoting a home and away team win, respectively, and D_{ij} corresponding to a draw, we have that the probabilities of these outcomes are equal to:

$$\begin{aligned}
\mathbb{P}(H_{ij}) &= \frac{1}{1+e^{c-d_{ij}}},\\
\mathbb{P}(D_{ij}) &= \frac{1}{1+e^{-c-d_{ij}}} - \frac{1}{1+e^{c-d_{ij}}},\\
\mathbb{P}(A_{ij}) &= 1-\frac{1}{1+e^{-c-d_{ij}}},
\end{aligned}$$

where $c > 0$ is an intercept.

3.2 Time evolution of team skills

During the season, a team's shape changes. There are many factors which influence current team shape, e.g., fatigue, physical preparation or absence of key players. In our simulation we incorporate all these factors in seasonal shape drift modelled by a random walk. If we denote consecutive league rounds with $t = 1, 2, \ldots, T$, where T is the total number of rounds, we assume that the ith team's rating at round t, r_i^t, evolves according to the formula:

$$r_i^{t+1} \sim \mathcal{N}(r_i^t, \nu),$$

where $\mathcal{N}(r_i^t, \nu)$ denotes the normal distribution with mean r_i^t and standard deviation ν. Such a model for evolution of a team's shape parameters was considered by, e.g, Rue and Salversen [7] or Glickman [3].

3.3 Team strength over a season

Due to fluctuation in teams' strength over the season, it is not immediately obvious how to define the strongest team overall. For each team, its strength evolution path over the season needs to be aggregated to a single number so that their comparison is possible. Clearly, if a given team is the strongest one in each consecutive round of play, this team should be chosen as the strongest team overall. However, due to a random fluctuations in teams' shape, such a situation does not happen in general. To enable comparison between different teams strength over the season we suggest to compute an average team rating during the consecutive rounds

$$\bar{r}_i = \frac{1}{T} \sum_{t=1}^{T} r_i^t.$$

There are other possible choices as a median or maximum rating or a more complicated comparison function.

3.4 Team ratings' distribution

In our simulations we assume that the prior team ratings are sampled from a certain probability distribution. In a related study, Ryvkin [8] proposes using normal, exponential and Pareto distributions. We perform simulations under these distributions with different standard deviations σ for the normal distribution family, rate parameters μ for exponential distributions and scale parameters s for Pareto distributions. Note that the differences in ratings according to the game outcome model described in Section 3.1 are shift invariant, hence we only focus on dispersion of the used distribution functions.

Next to the three parametric distribution listed above, we also propose taking samples from the estimated team ratings from the last four game seasons (2011/12–2014/15). To this end, we calculate a kernel density estimator (KDE) based on the ratings estimated for each season individually with the use of the presented ordered logistic regression model (Subsection 3.1) with the Gaussian kernel. Random variate generation according to the obtained density estimate is done via sampling with replacement teams' ratings from and adding a Gaussian noise term with the standard deviation equal to the kernel's bandwidth σ_b as KDE is in fact a mixture distribution. To estimate a team's strength parameter based on season data we use the presented ordered logistic regression model with elastic net regularisation [11]. Let $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ denote the vector of teams' ratings for a given season. To estimate these parameters we use maximum likelihood principle. The match outcome probabilities as given in Section 3.1 are dependent on the home team advantage parameter h and the intercept c which we are not subject to regularisation. Let us denote by $L(D|\mathbf{r}, h, c)$ the likelihood function of the results observed in dataset D given model parameters $\mathbf{r}, h, c$. To estimate team ratings we minimise

$$-\log L(D|\mathbf{r}, h, c) + \lambda \cdot \left(\frac{1}{2}(1-\alpha)\|\mathbf{r}\|_2^2 + \alpha\|\mathbf{r}\|_1 \right),$$

where $|| \cdot ||_1$ and $|| \cdot ||_2$ are L_1 and L_2 norms, respectively, and $\alpha \in [0, 1]$ and λ are parameters for the regularisation component. Figure 1 depicts the estimates of the mean likelihood of predictions (logarithmic loss) given by $\frac{1}{m}\sum_{i=1}^{m} \log p_i$, where p_i is the probability of the final outcome of i-th game in data attributed by the model, $i = 1, 2, \ldots, m$, and m is the total number of matches in the test set. We use 60/40 train/test split for different choices of parameters (α, λ). The split is performed according to time: the model is trained on the first 60% matches in a given season and evaluated on the other 40% of games. The prediction error is minimised for parameter setup $(\lambda, \alpha) = (1.5, 1)$ for this particular season. The value of parameter $\alpha = 1$ means that the L_1 regularisation yields the best performing model.

3.5 Model calibration

For simulation purposes we choose the parameters of the ordinal logistic regression model, ratings distribution and its drift so that the proportion of (H, D, A) outcomes is approximately equal to the one observed in several European leagues – Belgium, England, France, Germany, Greece, Italy, The Netherlands, Poland, Portugal, Scotland, Spain and Turkey – in season 2014/15. The minimal and maximal values of these proportions are presented in Table 1.

Moreover, the intercept and the home team advantage parameters are set to $(c, h) = (0.6, 0.4)$. These values are roughly equal to their estimates obtained from data as described in Section 3.4 (rounded to 0.1). In this way, the probabilities of results for equally rated teams $r_i = r_j$ are equal $(\mathbb{P}(H), \mathbb{P}(D), \mathbb{P}(A)) = (0.45, 0.28, 0.27)$, which approximately corresponds to the empirical averages observed for 2014/15 *Ekstraklasa* season equal to $(0.46, 0.27, 0.27)$.

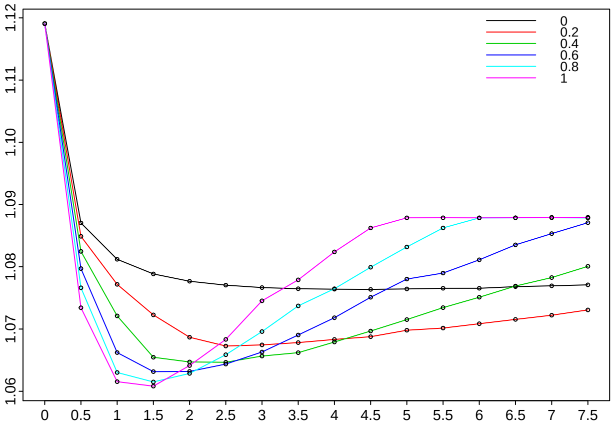

Fig. 1. Average logarithmic loss for test set for different choices of parameters λ (along x-axis) and α (coloured plots) for ratings in Ekstraklasa 2014/15 season.

Table 1. Statistics on proportions of home and away team wins and draws.

%	min	max
Home team wins	40% (Italy)	53% (Greece)
Draws	19% (Scotland)	31% (Italy)
Away team wins	22% (Greece)	36% (Scotland)

3.6 Evaluation metrics

Various metrics has been suggested for evaluation of the predictive efficacy of a tournament structure [2, 8, 9]. We employ a single metric to compare between league formats – the proportions of the strongest team win. This metric is computed as follows. We simulate a league competition for the two considered league formats. If the winner of the competition happens to be the team with the highest rating over a season according to the definition provided in Subsection 3.3, we say that the strongest team won the league. The experiment is repeated a large number of times to estimate the ability of a given league design to produce the strongest team as the winner of the tournament.

4 Results

Tables 2 and 3 present the proportions of the strongest team for the round-robin and the two-stage league formats, respectively. The prior ratings are derived from a KDE estimator of ratings based on previous season data. In rows we vary the parameter governing the spread of prior distribution and in columns the round to round team shape drift parameter. Here, the spread parameter is the

Gaussian kernel bandwidth σ_b. The estimation is based on 10,000 simulations of league competitions. In bold, we present significant differences in the corresponding entries of the tables based on a standard proportion test (at significance level of 0.05). The results are also presented in Figure 2. The tables for other distributions – normal, exponential and Pareto – are relegated to Appendix as similar qualitative conclusions apply to them.

Table 2. Simulation results for the double round-robin league format for ratings based on KDE.

σ_b/ν	*0.00*	*0.01*	*0.02*	*0.03*	*0.04*	*0.05*
0.1	0.449	0.454	0.454	0.457	0.458	0.462
0.3	0.480	0.478	0.482	0.485	0.491	0.496
0.5	0.549	0.546	0.546	0.546	0.549	0.553
0.7	0.609	0.603	0.601	0.601	0.605	0.606
0.9	0.661	0.652	0.654	0.652	0.647	0.652

Table 3. Simulation results for the two-stage league format for ratings based on KDE.

σ_b/ν	*0.00*	*0.01*	*0.02*	*0.03*	*0.04*	*0.05*
0.1	**0.470**	**0.472**	**0.477**	0.468	**0.483**	**0.486**
0.3	**0.507**	**0.511**	**0.502**	**0.510**	**0.516**	**0.519**
0.5	**0.572**	**0.573**	**0.577**	**0.575**	**0.579**	**0.576**
0.7	**0.631**	**0.641**	**0.634**	**0.636**	**0.641**	**0.635**
0.9	0.673	**0.681**	**0.678**	**0.682**	**0.682**	**0.682**

5 Summary and further work

Based on the results of simulation, we conclude that the two-stage system has better predictive capabilities to produce the strongest team as the winner of a competition. The differences are not large, albeit significant: they range from 1% to 3% with the median equal 2.5%. It should be stressed that the two-stage system involves a larger number of games to be played, which may yield better estimates of the strongest team's win. However, the games during the first stage of the season have a lower weight: 2 points for a win, while for a win during the second stage 3 points are awarded. In the double round-robin tournament each game is worth 3 points for a win. In this way, the weights of matches in different tournament formats differ and they cannot be compared directly.

The conclusion that two-stage league format is more efficient in determining the strongest team as the winner of the competition is somewhat contradictory

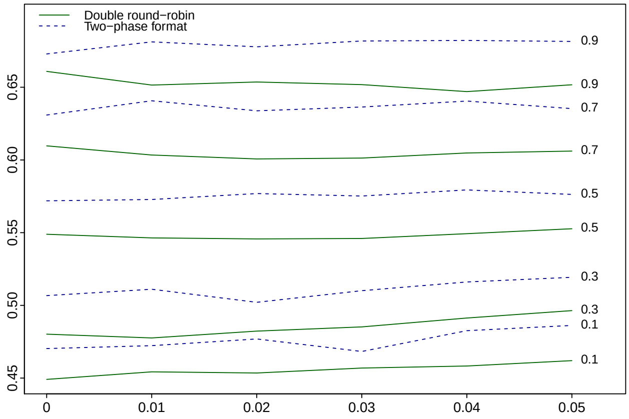

Fig. 2. Comparison of proportions (y-axis) on a plot for the two league formats according to drift parameter ν (y-axis) and different choices of bandwidth σ_b for KDE (on the right).

with opinion of football fans, that the authors experienced during informal conversations. Most of the surveyed fans believe that currently it is harder for the strongest team to win the championship. We also observe that the fraction of the top team's wins is increasing with the variance of ratings distribution. This means that the lower the competitive balance in the league, the higher proportion of better teams' wins is observed. The influence of drift parameter on the proportions is ambiguous: given the parameter governing spread of distribution, this proportion can increase, decrease or remain relatively unaffected with the increased values of parameter ν.

There are some limitations to the conducted study. First of all, we note that the new league format in a way can change the game. Since during the first stage the teams are effectively competing for 1.5 point for a win, we can possibly observe lower motivation/team commitment during this stage of competition. This observation is in analogy with the changes in football in mid-90 when FIFA – the governing body for association football competition over the world – introduced 3-points for a win rule. Previously, a team gained 2 points for a win. Various studies examined the effects of this rule on the competition [6, 12]. Based on these studies we may conclude that the change had influence on the teams strategies during the game. Secondly, we considered a random-walk for the evolution of team ratings during the season. Although such a model has been adopted in previous studies, the evolution of a team's shape can be a much more complex process, influenced by international cup games, players' injuries and transfers. Perhaps other processes underlying the evolution of team strengths can be studied, e.g., including shocks.

For further investigation of league designs we want to extend the presented study to other league formats as well as tournament metrics. We also want to address the mentioned limitations.

Acknowledgements. Jan Lasek would like to acknowledge the support by the European Union from resources of the European Social Fund, Project PO KL "Information technologies: Research and their interdisciplinary applications", agreement UDA-POKL.04.01.01-00-051/10-00 via the Interdisciplinary PhD Studies Program.

References

1. Aitchison, J. and Silvey, S.D.: The Generalization of Probit Analysis to the Case of Multiple Responses. Biometrika Vol. 44, No. 1–2, pp. 131–140 (1957)
2. Appleton, D.R.: May the Best Man Win? Journal of the Royal Statistical Society: Series C (The Statistician), Vol. 44, No. 4, pp. 529–538 (1995)
3. Glickman, M.E., Stern, H.S.: A State–Space Model for National Football League Scores. Journal of the American Statistical Association, Vol. 93, No. 441, pp. 25–35 (1998)
4. Lasek, J., Gągolewski, M.: Predictive Effcacy of a New Association Football League Format in Polish Ekstraklasa. Machine Learning and Data Mining for Sports Analytics Workshop at ECML/PPKD (2015)
5. Koning R.H.: Balance in Competition in Dutch Soccer. Journal of the Royal Statistical Society: Series C (The Statistician), Vol. 49, No. 3, pp. 419–431 (2000)
6. Moschini, G.: Incentives and Outcomes in a Strategic Setting: The 3-Points-For-a-Win System in Soccer. Economic Inquiry, Vol. 48, No. 1, pp. 65–79 (2010)
7. Rue, H. and Salvesen, O. Prediction and Retrospective Analysis of Soccer Matches in a League. Journal of the Royal Statistical Society: Series C (The Statistician), Vol. 49, No. 3, pp. 399–418 (2000)
8. Ryvkin, D.: The Selection Efficiency of Tournaments. European Journal of Operational Research, Vol. 206, No. 3, pp. 667–675 (2010)
9. Scarf, P., Yusof, M.M., Bilbao, M.: A Numerical Study of Designs for Sporting Contests. European Journal of Operational Research, Vol. 198, No. 1, pp. 190–198 (2009)
10. Szymanski, S.: The Economic Design of Sporting Contests. Journal of Economic Literature. Vol. 41, No. 4, pp. 1137–1187 (2003)
11. Zou, H. and Hastie, T.: Reguralization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society, Series B (Statistical Methodology), Vol. 67, pp. 301–320 (2005)
12. Borcas I. and Carrillo, J.D: Do the "The 3–Point–Victory" and "Golden Goal" Rules Make Soccer More Exciting? Journal of Sports Economics, Vol. 5, No. 2, pp. 169–185 (2004)
13. Pollard, R.: Home Advantage in Football: A Current Review of an Unsolved Puzzle. The Open Sports Sciences Journal, Vol. 1, No. 1, pp. 12–14 (2008)
14. Seckin, A. and Pollard, R.: Home advantage in Turkish professional soccer, Perceptual and Motor Skills, Vol. 107, No. 1, pp. 51–54 (2008)
15. Pollard, R. and da Silva, C. D. and Nisio, C. M.: Home advantage in football in Brazil: differences between teams and the effects of distance traveled, The Brazilian Journal of Soccer Science, Vol. 1, No. 1, pp. 3–10 (2008)

6 Appendix

Below results of simulations for other prior team ratings are presented. On most occasions, the two-stage league format yields significantly higher fraction of the strongest teams' championships in comparison to the double round-robin tournament.

Table 4. Simulation results for double round-robin league format for normally distributed prior ratings with standard deviation σ.

σ/ν	*0.00*	*0.01*	*0.02*	*0.03*	*0.04*	*0.05*
0.3	0.347	0.344	0.35	0.354	0.364	0.372
0.5	0.489	0.487	0.487	0.492	0.5	0.502
0.7	0.581	0.58	0.581	0.575	0.584	0.588
0.9	0.644	0.645	0.644	0.642	0.636	0.642
1.1	0.685	0.677	0.676	0.676	0.676	0.682

Table 5. Simulation results for two-stage league format for normally distributed prior ratings with standard deviation σ.

σ/ν	*0.00*	*0.01*	*0.02*	*0.03*	*0.04*	*0.05*
0.3	**0.363**	**0.364**	**0.368**	**0.38**	**0.394**	**0.409**
0.5	**0.512**	**0.51**	**0.514**	**0.522**	**0.519**	**0.53**
0.7	**0.608**	**0.609**	**0.601**	**0.608**	**0.613**	**0.607**
0.9	**0.663**	**0.668**	**0.671**	**0.669**	**0.666**	**0.669**
1.1	**0.708**	**0.707**	**0.708**	**0.707**	**0.706**	**0.704**

Table 6. Simulation results for double round-robin league format for exponentially distributed prior ratings with rate parameter μ.

μ/ν	*0.00*	*0.1*	*0.2*	*0.3*	*0.4*	*0.5*
0.8	0.814	0.811	0.8	0.796	0.79	0.77
1.2	0.768	0.757	0.746	0.739	0.73	0.73
1.6	0.704	0.706	0.702	0.7	0.702	0.702
2	0.653	0.659	0.66	0.674	0.68	0.694
2.4	0.601	0.615	0.635	0.656	0.669	0.683

Table 7. Simulation results for two-stage league format for normally exponentially prior ratings with rate parameter μ.

μ/ν	*0.00*	*0.1*	*0.2*	*0.3*	*0.4*	*0.5*
0.8	**0.828**	**0.822**	**0.818**	0.805	0.798	**0.798**
1.2	0.778	**0.775**	**0.768**	**0.765**	**0.76**	**0.764**
1.6	**0.722**	**0.722**	**0.721**	**0.735**	**0.737**	**0.756**
2	**0.677**	**0.68**	**0.691**	**0.716**	**0.73**	**0.745**
2.4	**0.628**	**0.632**	**0.67**	**0.711**	**0.728**	**0.742**

Table 8. Simulation results for double round-robin league format for Pareto-distributed prior ratings with scale parameter s.

s/ν	*0.00*	*0.05*	*0.1*	*0.15*	*0.2*	*0.25*
0.25	0.633	0.631	0.633	0.636	0.65	0.657
0.35	0.765	0.754	0.756	0.748	0.744	0.741
0.45	0.831	0.829	0.828	0.825	0.815	0.807
0.55	0.87	0.866	0.868	0.867	0.862	0.86
0.65	0.897	0.898	0.898	0.894	0.893	0.89

Table 9. Simulation results for two-stage league format for Pareto-distributed prior ratings with scale parameter s.

s/ν	*0.00*	*0.05*	*0.1*	*0.15*	*0.2*	*0.25*
0.25	**0.648**	**0.652**	**0.654**	**0.663**	**0.683**	**0.702**
0.35	**0.778**	**0.774**	**0.77**	**0.764**	**0.761**	**0.764**
0.45	**0.841**	**0.842**	0.836	0.829	0.823	**0.821**
0.55	0.879	**0.887**	**0.885**	**0.883**	**0.872**	0.87
0.65	**0.912**	**0.911**	**0.908**	**0.907**	**0.903**	**0.901**