Predictive Efficacy of a New Association Football League Format in Polish Ekstraklasa

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

[1] Interdisciplinary PhD Studies Program,
Institute of Computer Science, Polish Academy of Sciences,
Jana Kazimierza 5, 01-248 Warsaw, Poland, janek.lasek@gmail.com
[2] Systems Research Institute, Polish Academy of Sciences,
Newelska 6, 01-447 Warsaw, Poland, gagolews@ibspan.waw.pl
[3] Faculty of Mathematics and Information Science, Warsaw University of Technology,
Koszykowa 75, 00-662 Warsaw, Poland

Abstract. In this study we investigate the recently introduced competition format for the top association football division in Poland (similar to one used in, e.g., Belgium and Kazakhstan). We compare it to the double round-robin tournament which is the most prevalent league format among European leagues. In a simulation study we show that the new league format has better ability to determine the strongest competitor as the winner of a league as well as it yields higher correlation with theoretical latent teams' strength parameters in the model.

Keywords: Ekstraklasa, league format, predictive efficacy, round-robin tournament, simulation, football team ratings, tournament design.

1 Introduction

The issue of tournament design has several various dimensions. On one hand, one would like to design a competition so as to maximise the probability that the best team overall wins. On the other hand, the uncertainty of the outcome contributes to the excitement accompanying sport contests. Nothing is determined in advance and this very fact makes the competition interesting. Additionally, we are often facing time and space limitations that impose a set of constraints on both the schedule of the contest and the number of games that can be played. Moreover, there are economic factors that are of interest for a tournament organiser.

The design of sport contests have been of interest of authors in various research areas. For example, Appleton [1] 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 aspects. 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 this paper we focus on the predictive efficacy of the league format that was introduced in Polish *Ekstraklasa* – the top division of the football competition in Poland – as of the 2013/2014 season. We compare it with a standard round-robin league system that operated previously. The two tournament formats are compared with respect to their ability to produce the strongest contestant as the winner in a simulation study. Additionally, we investigate the level of agreement between the ranking of teams produced at the end of the competition and the one based on the teams' latent strength parameters in a simulation study.

This contribution is structured as follows. In Section 2, we present in detail the new tournament form. In Section 3, we describe a simulation experiment comparing the predictive efficacy of the two systems and in Section 4 we present the obtained results. Finally, the last section concludes the work.

2 League structure in Poland: Past and present

Over the years Polish *Ekstraklasa* operated on most occasions as a double round-robin tournament. In such a tournament, each team plays against all the other ones twice, home and away. It requires $2 \cdot \binom{n}{2}$ matches to be played, where n denotes the number of teams in a league. Since season 2005/06, 16 teams compete in the top division. However, as of season 2013/14 the competition format has changed. We shall now recall the rules of awarding points adopted with the introduction of the new system.

First of all, the season is divided into two phases. In the first phase, the teams compete in a standard double round-robin tournament (30 rounds). Next, the league table is divided into two groups: the championship ("top eight") and the relegation ("bottom eight") groups. At this stage, the number of points accumulated by the teams is divided by two (with possible rounding halves up) and the competition is extended to a single round-robin tournament within each group (with additional 7 rounds), which we call the final round. In this round, the points are awarded in a standard manner: 3 points for a win, 1 points for a draw and 0 points for a loss. The final ranking of teams is obtained by summing the points from the two rounds (within the championship and relegation group separately). This competition format requires $2 \cdot \binom{n}{2} + 2 \cdot \binom{n/2}{2}$ games to be played (with n even). Since the number of rounds in the final round is odd, some teams are playing one more game at home ground. These are the first four teams in each group after the first phase of a season. Moreover, in the final round, the top team in each group plays the second one at home.

As far as the top division leagues in the countries belonging to UEFA – the governing body for association football competition in Europe – are concerned, currently, the round-robin contest is most prevalent tournament format for domestic football competitions. In season 2014/15 the top division leagues in several countries operated in a general two-phase format resembling the one employed in Poland (with some special minor rules' modifications not discussed here), e.g., in Andorra, Belarus, Belgium, Bulgaria, Cyprus, Israel, Kazakhstan, Macedonia, Scotland and in Wales. In case of the leagues in Belgium and Kazakhstan, the points gained by the teams after the first phase are divided by two. Competition formats in these two countries are closely related to the league format which is currently in force in Poland. In fact, the Kazakh league is based on exactly the same rules but fewer teams are involved in the competition: 12 as compared to 16 in Poland.

3 Simulation experiment setup

Let us discuss the set up of an simulation experiment carried out for comparison of the new league system to the standard round-robin contest. We discuss the choice of model parameters. In the described computations we used data for four Polish league seasons (from 2011/12 to 2014/15) available at `http://www.90minut.pl/`. The data on European leagues' summary statistics were obtained from `http://www.football-data.co.uk/`.

Game outcome model. In the experiment, we employ an ordered logistic regression as the match results model [4]. The model depends on a single parameter per team – a rating – reflecting its latent strength. Teams' strength parameters are not directly observable – only the result of mutual games between the teams in a league are observed. Let r_i, r_j be ratings of two teams i and j and with team i playing at home ground. Let us denote with $d_{ij} = h + r_i - r_j$ the difference in the team ratings corrected for the home team advantage parameter h [13]. According to the model, if $R_{ij} \in \{H_{ij}, D_{ij}, A_{ij}\}$ is the set of 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:

$$R_{ij} = \begin{cases} H_{ij} & \text{if } d_{ij} + \epsilon \geq c, \\ D_{ij} & \text{if } d_{ij} + \epsilon \in (-c, c], \\ A_{ij} & \text{if } d_{ij} + \epsilon < -c, \end{cases} \tag{1}$$

where $c > 0$ is an intercept and ϵ is an i.i.d. (for all the games) random component. Under the assumption that the random component follows the logistic distribution with mean equal to 0 and scale parameter equal to 1, we have:

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

with $\mathbb{P}(\cdot)$ denoting the probability of a particular outcome.

The ordinal logistic regression is a simple model for match outcome. Each team is characterised by a single parameter indicating its overall strength. An alternative model could be the Poisson regression [2, 5]. However, in this model the teams are characterised by two parameters indicating their attacking and defence abilities. As a result, under such a model, a proper definition of "a better team" should be proposed. Under the ordinal logistic regression this definition is straightforward: better team is the one with higher rating.

Team ratings' distribution. In our simulations we assume that the team ratings are samples from a certain probability distribution. In an analogous 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 model (1) are shift invariant, hence we only focus on dispersion of the used distribution functions.

We also propose taking samples from the estimated team ratings from the last four game seasons (2011/12–2014/15). To this end, we estimate team ratings for these seasons. Next, we construct a kernel density estimator (KDE) based on these ratings with the Gaussian kernel. Random variate generation according to the obtained density estimate is done via sampling with replacement teams' from the teams' ratings and adding a Gaussian noise term with the standard deviation equal to the kernel's bandwidth σ_b. To estimate a team's strength parameters we use the ordered logistic regression model discussed above with elastic net regularisation [11]. Let us denote $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ the vector of teams' ratings for a given season. The likelihood function for the observed results is also dependent on a home team advantage parameter and intercepts, which we are not subject to regularisation. If $L(\mathbf{r})$ is the likelihood function of the observed results, to estimate team ratings we minimise:

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

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, $i = 1, 2, \ldots, m$, where 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.

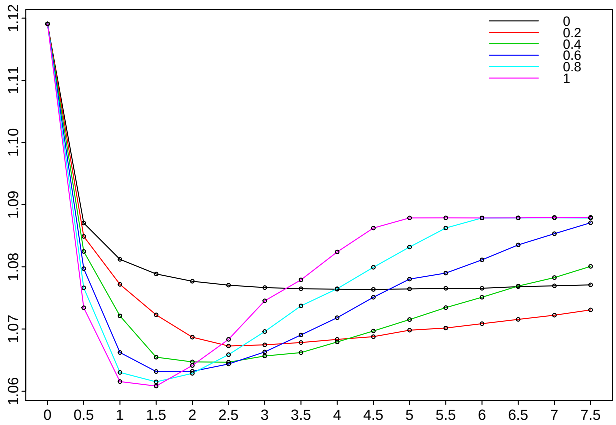

Fig. 1. Average logarithmic loss for test set for different choices of parameters λ (along x-axis) and α (coloured plots) for ratings in Ekstraklasa 2014/15 season.

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 Lasso regularisation yields the best performing model.

Model calibration. To set parameters for ratings distribution of the game outcome model discussed above, we look at the overall proportion of (H, D, A) results from European countries' leagues in season 2014/15 (Belgium, England, France, Germany, Greece, Italy, The Netherlands, Poland, Portugal, Scotland, Spain and Turkey). The proportion of the home team wins varies from 40% in Italy up to 53% in Greece. The fraction of draws ranges from 19% in Scottish Premier League to 31% in Italian *Serie A*. The highest fraction of away teams' wins is observed again in Scotland (36%) and the lowest in Greece (22%). The dispersion parameters of the assumed distribution functions are chosen so as to the simulated proportion of the results is approximately equal to the observed frequencies in the discussed football leagues. The intercept and the home team advantage parameters are set to $(c, h) = (0.6, 0.4)$. In this way, the probabilities of results for equally rated teams $r_j = r_j$ are equal $(\mathbb{P}(H), \mathbb{P}(D), \mathbb{P}(A)) = (0.45, 0.28, 0.27)$,

Table 1. Results for different distributions.

Kernel density-estimated with different bandwidths.

σ_h	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
π_1	0.443	0.458	0.481	0.514	0.546	0.580	0.608	0.634	0.657	0.674
π_2	0.473	0.488	0.509	0.540	0.575	0.608	0.636	0.662	0.682	0.703
τ_1	0.438	0.499	0.559	0.612	0.656	0.693	0.722	0.747	0.767	0.785
τ_2	0.458	0.521	0.581	0.633	0.677	0.712	0.741	0.765	0.784	0.801

Normal distributions of ratings with different σ.

σ	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**	**1.1**	**1.2**
π_1	0.342	0.423	0.490	0.542	0.583	0.616	0.643	0.666	0.684	0.698
π_2	0.365	0.451	0.517	0.568	0.611	0.642	0.668	0.690	0.707	0.724
τ_1	0.468	0.555	0.620	0.668	0.706	0.735	0.759	0.778	0.794	0.807
τ_2	0.489	0.577	0.641	0.689	0.725	0.753	0.776	0.794	0.810	0.822

Exponential distributions with different choices of rate μ.

μ	**0.8**	**1**	**1.2**	**1.4**	**1.6**	**1.8**	**2**	**2.2**	**2.4**	**2.6**
π_1	0.812	0.787	0.761	0.735	0.713	0.685	0.658	0.631	0.612	0.587
π_2	0.828	0.805	0.785	0.758	0.727	0.701	0.682	0.653	0.630	0.603
τ_1	0.741	0.698	0.660	0.625	0.593	0.563	0.536	0.512	0.489	0.468
τ_2	0.756	0.715	0.678	0.643	0.613	0.582	0.554	0.530	0.508	0.488

Pareto distributions with different choices of scale parameter s.

s	**0.25**	**0.3**	**0.35**	**0.4**	**0.45**	**0.5**	**0.55**	**0.6**	**0.65**	**0.7**
π_1	0.632	0.709	0.760	0.800	0.828	0.852	0.868	0.885	0.899	0.910
π_2	0.657	0.724	0.776	0.812	0.844	0.863	0.883	0.900	0.908	0.923
τ_1	0.422	0.478	0.527	0.568	0.602	0.633	0.658	0.681	0.701	0.720
τ_2	0.440	0.495	0.544	0.585	0.619	0.648	0.674	0.696	0.717	0.733

which approximately corresponds to the empirical averages observed for 2014/15 *Ekstraklasa* season equal to $(0.46, 0.27, 0.27)$.

Evaluation metrics. We employ two chosen metrics for comparison of the two discussed competition formats. The main aim of our study is to investigate which of the two league forms (the classic one or the new one) indicates the best team as the winner. For that, we calculate the percentage of simulations in which the best team won. Additionally, we compute Kendall's τ rank correlation coefficient for the two lists of teams: the first one based on pre-tournament ordering of teams according to their strength and the second one based on the final league ranking. We employ it as a measure of concordance between the two rankings. Other metrics studied in the literature can be well used. In this particular study we restrict our attention to the two given metrics and leave other ones for further research.

4 Results

Table 1 presents simulation results of estimation of the predictive efficacy of the two league formats for different choice of prior for ratings distributions. The outcomes are based on 100,000 simulations of possible scenarios. The proportion of wins of the highest ranked contestant is denoted with π_k with $k = 1, 2$ for the previous and new league format, respectively. Kendall's correlation coefficient's values are denoted by τ_k. For each entry we perform a test for equality of proportions (in case of π) and the Mann-Whitney-Wilcoxon test (for τ) to check if the results are significant (at the level of 0.05). All of the differences in Table 1 are significant.

5 Discussion and summary

Based on the results of our simulation experiment, we conclude that the newly introduced format has better predictive efficacy with respect to the both evaluation criteria: the selection of the strongest competition as well as the Kendall's correlation coefficient for the two list of teams − based on their true ratings and the final league ratings. The differences in the predictive efficacy of the two systems are significant though by not a large margin. We also note that the new system requires more games to be played in a given season. Under the presented simple model, more games played provides better results in the new league format. These results are in line with the principle that the more samples are obtained, the better are the estimates. On the other hand, the results are contradictory to common fans' believes experienced by the authors in the informal conversations: most of them claim that it is harder to win the league now than in the past. However, our simulation results show that the more the teams play, the higher the fraction of wins of the best contestant. The same applies to the Kendall's τ measure. We also observe that this fraction as well as Kendall's τ 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.

Limitations of the study. In this study, we analysed two league formats and their ability to rank stronger teams higher. We should stress the fact that the introduction of the new league format changed the rules of the competition. This may influence an individual team's behaviour and performance.[4] For example,

[4] Perhaps one of the most radical effects of rule changes on the competition is the example of Barbados–Grenada match in 1994 Caribbean Cup qualifiers. Due to introduction of a special rule that goals scored in extra-time counted double, the Barbados team deliberately scored an own goal by the end of the match for the extra-time to take place. This in turn allowed them to win by a margin of two goals by scoring effectively one goal, which allowed them to progress to the next round of the competition (see `https://en.wikipedia.org/wiki/Barbados_4-2_Grenada_(Caribbean_Cup_qualification)`, last access date July 23, 2015).

in mid-1990's, FIFA – the governing body for association football competition over the world – introduced 3-points for a win format, replacing an old 2-points for a win system. Several authors considered the influence of this change on the game. The conclusion is that such a change does have an impact on the strategy and tactics employed by the teams [6, 12], e.g., more attacking play under new rules. In our setting, we can possibly observe some motivation changes in the first part of the season. Roughly speaking, each game in the first round is worth 1.5 point for a win 0.5 point for a draw and 0 points for a loss. Due to this fact, we could observe less engagement in during the first part of the season and maximal motivation for its final part when the stakes are effectively doubled. On the other hand, based on our simulation results an indirect conclusion might be drawn that in order to win the league the team needs to play with maximal engagement regardless of the part of the season to win the league.

Future work. The observations implied from our analysis and the discussed limitations of the simulation model opens possibilities for further exploration of the topic of optimal tournament design. We note that a football league is a dynamic, evolving system. A team's shape is subject to fluctuations throughout the season. Also, unexpected events like players' injuries or transfers may influence a team's performance. Such a setting is suitable for more advanced analysis of league systems by Bayesian methods [3, 7]. Furthermore, as we observed that regularisation component leads to better predictive accuracy of ratings, again Bayesian models with regularising prior may be of use. There is still much room for improving the basic model to bring the simulation closer to reality. This interesting topic is left for further research.

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. 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)
2. Dixon, M.J. and Coles, S.G.: 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 (1997)
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. 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)
5. Maher, M.J.: Modelling Association Football Scores. Statistica Neerlandica, Vol. 36, No. 3, pp. 109–118 (1982)
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)