Slovak University of Technology in Bratislava
Faculty of Mechanical Engineering
17th Conference on Applied Mathematics **APLIMAT 2018**

Proceedings

COMPARISON OF HOME TEAM ADVANTAGE IN ENGLISH AND SPANISH FOOTBALL LEAGUES

MAREK Patrice (CZ), VÁVRA František (CZ)

Abstract. Home team advantage in sports is widely analysed phenomenon. This paper builds on results of recent research that – instead of points gained – uses goals scored and conceded to describe home team advantage. Using this approach, the home team advantage is random variable that can be described by trinomial distribution, and it is possible to use Jeffrey divergence and test for homogeneity of parallel samples to compare and test home team advantage of different leagues. The paper also introduces procedure that is based on identification of all distances between home advantages of leagues. These distances are later used to construct disconnected graph with components that contain leagues with similar home team advantage. Procedures are demonstrated on five top level English football leagues and two top level Spanish leagues from the 2007/2008 season to the 2016/2017 season.

Keywords: Home team advantage, Combined measure, Jeffrey divergence, Test for homogeneity of parallel samples, Premier League, La Liga

Mathematics subject classification: Primary 62P99; Secondary 62F03, 62F10

Home team advantage in sports is widely analysed phenomenon. Usually, the home team advantage is analysed by comparing the number of points that a team earned at its home ground with the total number of points that the team earned. This paper builds on results of recent research that – instead of points gained – uses goals scored and conceded to describe home team advantage.

We define random variable C *combined measure of home team advantage* that – according to results between teams – can take values -1 (away team advantage), 0 (no advantage), and 1 (home team advantage). Random variables $Z_r, r = -1, 0, 1$ describe number of cases (in a season) where it is possible to identify home team advantage ($r = 1$), away team advantage ($r = -1$), and no advantage ($r = 0$), i.e. Z_r sums number of cases where $C = r$. Vector (Z_{-1}, Z_0, Z_1) follows trinomial distribution.

To compare leagues, we use *Jeffrey divergence* – symmetric version of Kullback-Leibler distance – that allows to measure distances between probability distributions that describe home team advantage. Next, approach based on the *Test for homogeneity of parallel samples* allows to test hypothesis H_0 : *Two samples come from the same population* (or in our case: that home advantage in both tested

leagues is described by the same distribution) against alternative hypothesis H_1 : *Two samples do not come from the same population* (or in our case: that home advantage in both tested leagues come from different distributions). Moreover, χ^2 statistic used in this test is also used to measure distances between probability distributions that describes home team advantage. The paper also introduces procedure that is based on identification of all distances between home advantages of leagues. These distances are later used to construct disconnected graph with components that contain leagues with similar home team advantage. Procedures are demonstrated on five top level English football leagues and two top level Spanish leagues from the 2007/2008 season to the 2016/2017 season.

1 Introduction

Home team advantage in sports was investigated by many researchers, for exhaustive study see Pollard and Pollard [6] and Allen and Jones [1]. Home team advantage in these papers is based on comparing the number of points that a team earned at its home ground with the total number of points that the team earned. Marek and Vávra [5] used approach where – instead of points – home team advantage is based on number of goals scored and conceded. This approach is used in this paper because of its advantage that can be illustrated by the following example. Let us assume that two teams A and B played each other exactly two times in a season, and team A won both matches: 4–1 at the home ground and 2–1 at the home ground of team B. Using points in this situation will result in conclusion that team A won 50% of points at the home ground, i.e. that there is no home team advantage. However, using goals and their differences – for team A: 3 goals at the home ground and 1 goal at the away team ground – shows that team A played better at the home ground. The same holds true for the team B who lost by 1 goal at the home ground and by 3 goals at the away team ground, i.e. the better result was recorded at the home ground.

Marek and Vávra [5] defined three types of measures: *Active measure of home team advantage* that takes into account difference between goals scored at home and away; *Passive measure of home team advantage* that takes into account difference between goals conceded at home and away; and *Combined measure of home team advantage* – used in this paper and specified in Definition 1 – that takes into account both goals scored and conceded.

Definition 1 Combined measure of home team advantage *is random variable C that can take values $-1, 0,$ and 1. $C = -1$ for team T_1 if two matches between teams T_1 and T_2 in a season ended with a better result – measured by goal differences in matches – for team T_1 away from home (at T_2's ground). $C = 0$ for team T_1 if goal differences in both matches were exactly the same from T_1's point of view, and $C = 1$ for team T_1 if this team recorded better result – measured by goal differences in matches – at its own ground. With results $h_{T_1} : a_{T_2}$ at the home ground of team T_1 and $h_{T_2} : a_{T_1}$ at the home ground of team T_2 the value of random variable C for team T_1 is determined as*

$$C = \operatorname{sgn}((h_{T_1} - a_{T_2}) - (a_{T_1} - h_{T_2})). \tag{1}$$

2 Data and Methods

This section describes data and methods used for the estimation of parameters. It also introduces approach to compare home team advantage of different leagues.

2.1 Compared Leagues

Home team advantage in five top level English football leagues and two top level Spanish football leagues are compared by the methods described in the following subsections. Results of ten latest seasons – i.e. from the 2007/2008 season to the 2016/2017 season – used in the comparison are obtained from [3]. The following leagues are used for England:

- Premier League (level 1), ENG_1,
- English Football League Championship (level 2), ENG_2,
- English Football League One (level 3), ENG_3,
- English Football League Two (level 4), ENG_4,
- National League (level 5), ENG_5.

Next two leagues are used for Spain:

- La Liga (level 1), ESP_1,
- Segunda Divisíon (level 2), ESP_2.

Websites [7] and [4] were used for basic control of all data, e.g., total number of goals scored by each team in the whole season, and to fill in results that were occasionally not recorded in the original source. Following matches were excluded (and consequently the opposite matches): Blackpool vs. Huddersfield in 2014/2015 season of ENG_2 (because the match was abandoned after 48 minutes of play, and the final result 0–0 reflects only these 48 minutes) and all matches of Chester City in 2009/2010 season of ENG_5 (because Chester City was expelled during the season).

2.2 Parameters estimation method

A balanced schedule was used in all seasons, i.e. each team played each other team exactly two times, once as a home team and once as a visiting team. For example, the English Premier League consisted of 20 teams in all seasons and total number of matches in one season is always 380. The combined measure of home team advantage joins results of the first part of the season (where team A plays at its home ground against team B) with results of the second part of the season (where team A plays against team B at B's home ground). Therefore, 190 observations of combined measure are available for one season in the English Premier League.

Let $Z_r, r = -1, 0, 1$ are random variables which describe number of cases (in a season) where it is possible to identify home team advantage ($r = 1$), away team advantage ($r = -1$) and no advantage ($r = 0$), i.e. Z_r sums number of cases where $C = r$. Vector (Z_{-1}, Z_0, Z_1) follows trinomial distribution with parameters p_{-1}, p_0, p_1, and K with probability function

$$P(k_{-1}, k_0, k_1) = \frac{K!}{k_{-1}!k_0!k_1!} p_{-1}^{k_{-1}} p_0^{k_0} p_1^{k_1}, \tag{2}$$

where K is total number of observations of combined measure in a season, p_{-1}, p_0, p_1 are probabilities of occurring home team advantage ($r = 1$), away team advantage ($r = -1$) and no advantage ($r = 0$). k_{-1}, k_0, k_1 ($k_{-1} + k_0 + k_1 = K$) are observations of appropriate advantage.

Maximum likelihood estimator of parameters $p_r, r = -1, 0, 1$ is

$$\hat{p}_r = \frac{k_r}{K}, \quad r = -1, 0, 1. \tag{3}$$

The results of the highest leagues in England and Spain in the 2016/2017 season are used to illustrate previous procedure. Observed counts of combined measure of home team advantage are in Table 1, and estimated parameters are later presented in Table 2

League	k_{-1}	k_0	k_1
ENG_1	56	34	100
ESP_1	52	41	97

Tab. 1. Observed counts of combined measure in 2016/2017 season.

League	$\hat{p}_{-1}$	$\hat{p}_0$	$\hat{p}_1$
ENG_1	0.295	0.179	0.526
ESP_1	0.274	0.216	0.511

Tab. 2. Estimated probabilities in 2016/2017 season.

2.3 Methods of Comparison

Two approaches were adopted to compare home team advantage in different leagues. The first one is *Jeffrey divergence* (JD), i.e. symmetric version of *Kullback-Leibler distance*, which allows to measure distance between two probability distributions P and Q

$$JD(P, Q) = \sum_x \left(p(x) \ln \frac{p(x)}{q(x)} + q(x) \ln \frac{q(x)}{p(x)} \right). \tag{4}$$

More about Jeffrey divergence can be found in Deza and Deza [2] (page 185). Equation 2 allows to measure distance between probability distributions of (Z_{-1}, Z_0, Z_1) for each two leagues. These distances can be compared to find leagues with similar, or, on the contrary, different home team advantages. Values of JD are frequently small, therefore – for better readability in the following parts – we multiply them by 100. Jeffrey divergence (multiplied by 100) of data presented in Table 2 is 0.894.

The second approach is based on the *Test for homogeneity of parallel samples* that allows to test hypothesis H_0 : *Two samples come from the same population* (or in our case: that home advantage in both tested leagues is described by the same distribution) against alternative hypothesis H_1 : *Two samples do not come from the same population* (or in our case: that home advantage in both tested leagues come from different distributions). χ^2 statistic can be – for the case with two leagues – computed as

$$\chi^2 = \sum_{r=-1}^{1} \frac{\left(n_{1,r} - \frac{n_{1,.} n_{.,r}}{n_{.,.}} \right)^2}{\frac{n_{1,.} n_{.,r}}{n_{.,.}}} + \sum_{r=-1}^{1} \frac{\left(n_{2,r} - \frac{n_{2,.} n_{.,r}}{n_{.,.}} \right)^2}{\frac{n_{2,.} n_{.,r}}{n_{.,.}}}, \tag{5}$$

	$C=-1$	$C=0$	$C=1$	Total
Sample 1	$n_{1,-1}$	$n_{1,0}$	$n_{1,1}$	$n_{1,.}$
Sample 2	$n_{2,-1}$	$n_{2,0}$	$n_{2,1}$	$n_{2,.}$
Total	$n_{.,-1}$	$n_{.,0}$	$n_{.,1}$	$n_{.,.}$

Tab. 3. Observed numbers in given sample and class.

where appropriate observed numbers ($n_{1,-1}, n_{1,0}$, etc.) are obtained according to Table 3.

Assymptotical distribution of χ^2 statistic for the data in Table 3 is χ^2 distribution with two degrees of freedom (i.e. number of classes reduced by one). More about this test and its more general form can be found in [8] (pages 398–402).

χ^2 statistic for the data presented in Table 1 is 0.847, and corresponding p-value is 0.655. The hypothesis H_0 that home team advantage in ENG_1 (Premier League) and ESP_1 (La Liga) in the season 2016/2017 is described by the same distribution cannot be rejected, i.e. we cannot reject that there is no difference in home team advantage (measured by combined measure) in these two leagues in the 2016/2017 season.

Method based on χ^2 statistic can be used to test whether the home team advantage in two leagues differ or not. Next, both methods – the first based on JD and the second based on χ^2 – can be used to identify groups of leagues with similar home team advantage. In this case, we will proceed according to these steps:

- construct graph, where nodes represent leagues, and edges represent distance (JD or χ^2) between leagues,
- remove the edge with the highest distance from the graph,
- repeat the previous step until the original full graph becomes disconnected, i.e. until two components are obtained.

Previous procedure will identify two groups of leagues where distance (JD or χ^2) between any league from the first group and the second group is always equal or greater than the last removed distance of the graph. This procedure can be further used to obtain more than two components.

3 Results

The first part will present full results for data summed over all ten seasons for each league. Next part will present major findings for single seasons.

3.1 Categorization by Leagues

In this part, we do not distinguish among seasons; therefore, the findings are valid for leagues in general (or, more precisely, for the last ten seasons).

Table 4 presents observed counts of combined measure for all leagues over all seasons, i.e. from the 2007/2008 season to the 2016/2017 season. ENG_4 recorded the lowest value of $\hat{p}_1$ and ESP_1 recorded the highest value. The second highest value of $\hat{p}_1$ was recorded by ESP_2. Both ESP_1 and ESP_2 recorded two lowest values of $\hat{p}_{-1}$. Based on these numbers, we can formulate hypothesis that home team advantage in Spain is higher than in England.

JD divergence (multiplied by 100) is used to measure distances between each two leagues, and results are presented in Table 5. For example, the distance between the English Premier League and Spanish

League	$k=-1$	$k=0$	$k=1$	Total	$\hat{p}_{-1}$	$\hat{p}_0$	$\hat{p}_1$
ENG_1	520	346	1 034	1 900	0.274	0.182	0.544
ENG_2	796	510	1 453	2 759	0.289	0.185	0.527
ENG_3	845	503	1 412	2 760	0.306	0.182	0.512
ENG_4	922	518	1 320	2 760	0.334	0.188	0.478
ENG_5	845	494	1 398	2 737	0.309	0.180	0.511
ESP_1	484	304	1 112	1 900	0.255	0.160	0.585
ESP_2	600	438	1 272	2 310	0.260	0.190	0.551

Tab. 4. Observed counts of combined measure for all ten seasons with appropriate estimated probabilities.

La Liga is quantified by $100 \cdot JD(ENG_1, ESP_1) = 0.721$. This gives us brief look at difference between home team advantage in studied leagues. The procedure described in section 2.3 is used to obtain disconnected graph (where nodes are leagues, and edges are distances between leagues). After removing edges – the last removed edge is between ENG_1 and ESP_1 – we obtain six leagues in one group (white colour in Table 5) and ESP_1 in the second group (grey colour in Table 5).

League	ENG_1	ENG_2	ENG_3	ENG_4	ENG_5	ESP_1	ESP_2
ENG_1		0.140	0.566	2.072	0.636	0.721	0.111
ENG_2	0.140		0.152	1.138	0.196	1.398	0.421
ENG_3	0.566	0.152		0.484	0.004	2.226	1.080
ENG_4	2.072	1.138	0.484		0.442	4.752	2.892
ENG_5	0.636	0.196	0.004	0.442		2.299	1.191
ESP_1	0.721	1.398	2.226	4.752	2.299		0.724
ESP_2	0.111	0.421	1.080	2.892	1.191	0.724	

Tab. 5. JD divergence for leagues over all ten seasons.

Last view is based on χ^2 statistic. Table 6 presents p-values for leagues over all ten seasons. All p-values are based on χ^2 distribution with two degrees of freedom; therefore, values of χ^2 are not presented, and values in Table 6 can be easily used to identify which leagues are more distant (the lower p-value, the higher distance). The procedure to obtain disconnected graph was also used for distances based on χ^2. The result is exactly the same as in the previous case, i.e. ESP_1 belongs to one group and the rest of leagues to the other group.

League	ENG_1	ENG_2	ENG_3	ENG_4	ENG_5	ESP_1	ESP_2
ENG_1		0.456	0.042*	<0.001*	0.029*	0.033*	0.561
ENG_2	0.456		0.350	<0.001*	0.260	<0.001*	0.071
ENG_3	0.042*	0.350		0.036*	0.973	<0.001*	0.001*
ENG_4	<0.001*	<0.001*	0.036*		0.048*	<0.001*	<0.001*
ENG_5	0.029*	0.260	0.973	0.048*		<0.001*	0.001*
ESP_1	0.033*	<0.001*	<0.001*	<0.001*	<0.001*		0.024*
ESP_2	0.561	0.071	0.001*	<0.001*	0.001*	0.024*	

Tab. 6. p-values for leagues over all ten seasons.

Asterisk in Table 6 is used to identify p-values lower that 0.05, i.e. pairs of leagues where – on 5% level of significance – the hypothesis H_0 that home team advantage in these two leagues is described by the same distribution can be rejected, and the alternative H_1 that home team advantage in these two leagues is described by different distributions can be accepted. Test of ESP_1 with any other league results in rejection of H_0 and acceptance of H_1, i.e. that there is difference in home team advantage. We can interpret this result – taking into account values from Table 4 – that home team advantage in ESP_1 is significantly stronger than in other tested leagues.

Next league with all p-values lower that 0.05 is ENG_4. In this case we can interpret this result – taking into account values from Table 4 – that home team advantage in ENG_4 is significantly weaker than in other tested leagues. This result is supported by the fact that if the procedure of finding components in Table 5 and Table 6 is be used to find three components then one component is ESP_1 (higher home team advantage), the second component is ENG_4 (lower home team advantage), and the third component is group of rest leagues.

3.2 Categorization by Leagues and Seasons

Previous part suggests that the highest difference is between ESP_1 and ENG_4 (for example, the highest JD divergence in Table 5); therefore, these two leagues are chosen to present evolution of $\hat{p}_1$ in Figure 1. Moreover, this figure contains evolution of $\hat{p}_1$ for the English Premier League ENG_1. Other leagues are omitted to maintain readability of the figure. Point estimate of p_1 for ESP_1 is in all seasons above point estimate of p_1 for ENG_4 and usually above point estimate of p_1 for ENG_1.

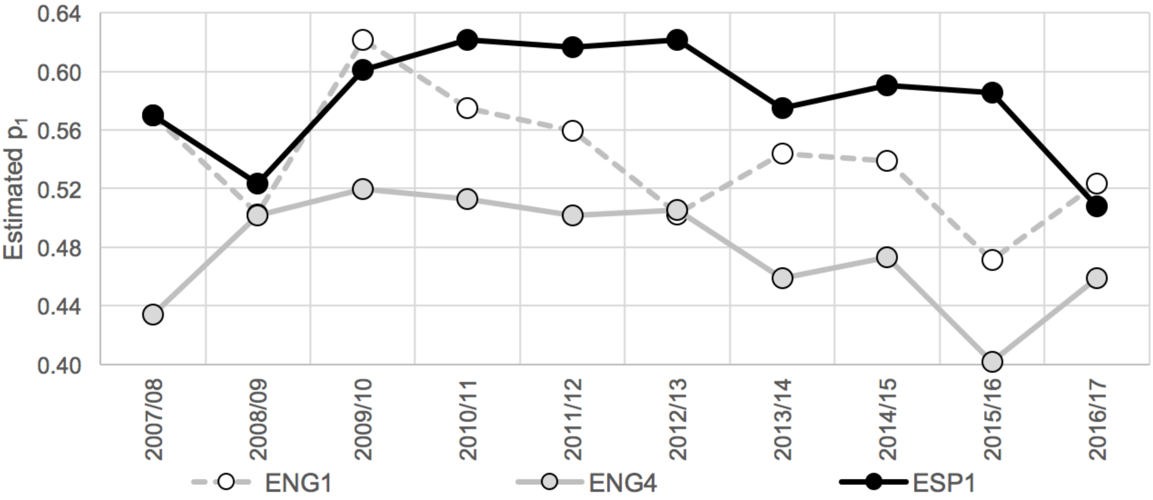

Fig. 1. Evolution of $\hat{p}_1$ for ENG_1, ENG_4, and ESP_1.

The test for homogeneity of parallel samples (see Section 2.3) can be used identify differences in home team advantage between leagues in single seasons. Here, we will present only results for the test of ESP_1 against ENG_1 and ENG_4. Hypothesis H_0 that home team advantage in ESP_1 and ENG_4 is described by the same distribution cannot be rejected in three seasons: 2008/2009, 2009/2010, and 2016/2017. The hypothesis H_0 can be rejected in the rest of tested seasons, and its alternative H_1 that home team advantage in ESP_1 and ENG_4 is described by different distributions can be accepted. The same test where ESP_1 and ENG_1 are used is not able to reject H_0 in all tested seasons. These results suggest that when single seasons are used then – to identify difference in home team advantage – the difference has to be strong. Therefore, we will use the procedure based on searching components of graph (see section 2.3).

Components obtained in every season are presented in Table 7. Results are based on distance measured by χ^2 statistic that was defined in Equation 5. For example, a graph with two components is obtained in the 2010/2011 season. The first component contains $ENG_2, ENG_3, ENG_4,$ and ENG_5, and the second component contains $ENG_1, ESP_1,$ and ESP_2. The interpretation is based on values of $\hat{p}_r, r = -1, 0, 1$. The first component contains leagues where all leagues have value of $\hat{p}_{-1}$ higher than leagues of the second component. The situation in the case of $\hat{p}_1$ is the opposite, and all leagues of the second component have value of $\hat{p}_1$ higher than leagues of the first component. We emphasize that words *higher* and *lower* are here used only to distinguish between two components (i.e. the meaning is relative), and it should not be interpreted as components with high or low home team advantage. Other leagues – except ENG_3 – are consequently classified as leagues with higher home team advantage.

Season	Lower home adv.	Higher home adv.
2007/2008	ENG_4	all others
2008/2009	all others	ESP_2
2009/2010	all others	ENG_1
2010/2011	all others	$ENG_1, ESP_{1,2}$
2011/2012	all others	ESP_1
2012/2013	all others	$ESP_{1,2}$
2013/2014	$ENG_{2,3,4}$	all others
2014/2015*	ENG_3	all others
2015/2016**	ENG_4	$ENG_{1,2,5}, ESP_{1,2}$
2016/2017	all others	ESP_2

Tab. 7. Obtained components in seasons (based on χ^2).

Table 7 contains two remarks. The first remark (*) is for the 2014/2015 season which is the only season where components obtained by the same procedure but with distances based on JD divergence (Equation 4) are different. Components in the case of JD divergence are ESP_1 with higher home team advantage and *all others* with lower home team advantage. The second remark (**) belongs to the 2015/2016 season where obtained components are ENG_3 and *all others*. This is the only season where the conclusion is not made in favour of higher or lower home team advantage, and it is possible to observe both situations at the expense of the lower value of $\hat{p}_0$. ENG_3 has the second highest value of $\hat{p}_{-1}$ and also of $\hat{p}_1$. Therefore, the procedure continued to obtain third component (ENG_4) which can be classified as a league with lower home team advantage.

4 Conclusion

This paper introduced procedure for comparison of the home team advantage in different leagues. The procedure uses combined measure of home team advantage. This measure is based on both goals scored and conceded; and therefore, it is able to detect home team advantage in a better way than usual approach based on points gained. Two procedures were proposed – the first is based on determining distances between leagues, and the second uses classical statistical hypothesis testing. Distances can be further used to obtain groups of leagues with similar home team advantage based.

Procedures applied on five top level English football leagues and two top level Spanish leagues suggest that home team advantage in Spain is higher that in England. More specifically, La Liga (level 1 Spanish league) has the highest home team advantage among the analysed leagues, and English

Football League Two (level 4 English league) has the lowest home team advantage among analysed leagues. More convincing results are obtained when all ten analysed seasons are used together.

Acknowledgement

This publication was supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports.

References

[1] M. S. ALLEN and M. V. JONES. The home advantage over the first 20 seasons of the English Premier League: Effects of shirt colour, team ability and time trends. *International Journal Of Sport And Exercise Psychology*, 12(1):10–18, 2014.

[2] E. DEZA and M. M. DEZA. *Dictionary of Distances*. Elsevier, 2006.

[3] England Football Results and Betting Odds. Premiership Results and Betting Odds, 2017. http: // www.football-data.co.uk/englandm.php.

[4] Liga de Fútbol Profesional. The official website for Laliga, 2017. http://www.laliga. es/en/ laliga-santander.

[5] P. MAREK and F. VÁVRA. Home Team Advantage in English Premier League. In *Mathsport International 2017 Conference Proceedings*, pages 244–254, 2017.

[6] R. POLLARD and G. POLLARD. Long-term trends in home advantage in professional team sports in North America and England (1876–2003). *Journal of Sports Sciences*, 23(4):337–350, 2005.

[7] Premier League Football News, Fixtures, Scores & Results. Premier League Football Scores, Results & Season Archives, 2017. https://www.premierleague.com/results?co= 1&se=1&cl=-1.

[8] C. R. RAO. *Linear Statistical Inference and its Applications*. John Wiley & Sons, 2002.

Current address

MAREK Patrice, Ing., Ph.D.
European Centre of Excellence NTIS – New Technologies for Information Society
Faculty of Applied Sciences, University of West Bohemia
Univerzitní 8, 301 00 Plzeň, Czech Republic
E-mail: patrke@kma.zcu.cz

VÁVRA František, doc. Ing., CSc.
European Centre of Excellence NTIS – New Technologies for Information Society
Faculty of Applied Sciences, University of West Bohemia
Univerzitní 8, 301 00 Plzeň, Czech Republic
E-mail: vavra@kma.zcu.cz