Metodološki zvezki, Vol. 17, No. 1, 2020, 55–76

Home team advantage in the English Premier League

Patrice Marek[1] František Vávra[2]

Abstract

The home team advantage in association football is a well known phenomenon. The aim of this paper is to offer a different view on the home team advantage. Usually, in association football, every two teams – team A and team B – play each other twice in a season. Once as a home team and once as a visiting, or away team. This gives us two results between teams A and B which are combined together to evaluate whether team A, against its opponent B, recorded a result at its home ground – in the comparison to the away ground – that is better, even, or worse. This leads to a random variable with three possible outcomes, i.e. with trinomial distribution. The combination and comparison of home and away results of the same two teams is the key to eliminate problems with different squad strengths of teams in a league. The bayesian approach is used to determine point and interval estimates of unknown parameters of the source trinomial distribution, i.e. the probability that the result at home will be better, even, or worse. Moreover, it is possible to test a hypothesis that the home team advantage for a selected team is statistically significant.

1 Introduction

Home team advantage in sports is a well known phenomenon. It is frequently used in models that estimate the probability of win, draw, and loss in a match. The use of home team advantage in modelling and predicting sports results can be traced back to Maher (1982) who used one parameter to include this information. Home team advantage was later used in many papers that studied different kind of sports, e.g., in association football by Dixon and Coles (1997), in water polo by Karlis and Ntzoufras (2003), and in ice hockey by Marek et al. (2014).

[1] European Centre of Excellence NTIS – New Technologies for Information Society, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic; patrke@kma.zcu.cz

[2] European Centre of Excellence NTIS – New Technologies for Information Society, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic; vavra@kma.zcu.cz

The main causes of home advantage in association football – crowd effect, referee bias, travel effect, familiarity with local conditions, territoriality, special tactics, etc. – are analysed and discussed in detail by Pollard and Gómez (2014).

Home team advantage as a self-standing phenomenon has been studied by many authors, e.g. by Leite and Almeida (2018) in Portuguese futsal; Rooney and Kennedy (2018) in Gaelic football; Jones (2015) in Major League Baseball; Pollard and Gómez (2015) in several college and professional team sports.

Exhaustive analysis of home team advantage was studied by Pollard and Pollard (2005). Their paper offers an interesting summary of previous research on this phenomenon and analysis of more than 400 000 matches in many different sports played between years 1876 and 2003. They quantified home team advantage in association football as "*the number of points obtained by the home team expressed as a percentage of all points obtained in all games played*". This definition is usually used in research papers (including papers mentioned in the previous paragraph) and was introduced by Pollard (1986).

The same definition of home team advantage was used by Allen and Jones (2014) in their analysis of the English Premier League in the seasons 1992/1993–2011/2012. Their results showed that 60.77% of total points were won in home matches. Note that Allen and Jones (2014) used data with deducted points[3], i.e. they used 19 points for Portsmouth in the 2009/2010 season, the correct number of points obtained in real games was 28; and they used 39 points for Middlesbrough in the 1996/1997 season, where the correct number of points obtained in real games was 42. Using the correct number of points slightly changes the overall result to 60.80%.

The problem with deducted points suggests that using points can cause some complications. The next problem can be illustrated when the same data (seasons 1992/1993–2011/2012) is used, but only two points are awarded for a win. Then the proportion of points obtained in home matches declines from 60.80% to 59.83%. Pollard and Ruano (2009) mentioned another problem of this method when it is applied to individual teams where some adjustments have to be made. They mention two main reasons for the necessity of adjustment. The first is that the overall home team advantage for all teams in a given season will affect the value of home team advantage of each individual team, and the second is that a team's ability influences the magnitude of home team advantage, i.e. strong teams usually win both games at home and away; therefore, they do not achieve a high calculated value for home team advantage.

The main limitation of the previous – and widely used – approach is that it is not easily applicable to individual teams, its results are influenced by the point system and, above all, it understands home team advantage only in the terms of points obtained. This

[3]In some cases, a team that violates rules and policies of a given league is affected by deduction of points. This means that even though the team has achieved a certain number of points (P) in real matches, some points (D) are deducted for breaking the rules. The final number of points is therefore $P - D$. In rare cases, it is also possible that the total number of points will be negative.

means that if we are dealing with a very strong team that wins almost all of its matches, the method will not identify home team advantage, as this team will gain half the points on the home ground. A similar problem will be with weak teams, where, in addition, the results will be very sensitive to each point gained.

Our paper offers a slightly different view on home team advantage and – instead of points – home team advantage is based on the number of goals scored and their differences in paired matches. Similar approach based on goals was already used in Clarke and Norman (1995), where all matches from one season were analysed together. Their approach does not offer a method for testing a hypothesis about individual home team advantage as they use models for predicting match results where they use special parameter for home team advantage (as it is common in these models). Nevertheless, using goals is a good idea, and its advantage can be demonstrated on results of a team that played the same opponent at home and away and won both matches. Let us assume that the result at home was a 3–0 win, and the result away was a 2–1 win. Obviously, the better result was recorded at the home ground; however, based on points obtained, it is not possible to distinguish between these results as the team is always awarded by three points for a win.

The method described in our paper allows us to distinguish between these two results and to identify home team advantage. It also offers a new approach how to measure home team advantage for a single team, and to observe changes during time. Moreover, it allows us to perform statistical testing of the hypothesis that a single team has home team advantage. Our concept is based on the analysis of the phenomenon *"A better result is achieved on the home ground than on the opponent's ground"*, i.e. if a team wins on the home ground it wins "more" than on the opponent's ground and if a team loses on the home ground it loses "less" than on the opponent's ground. A random variable with three possible values will be used to distinguish these situations – a value of 1 will represent a better result on the home ground, -1 a worse result on the home ground and 0 a situation where it is not possible to decide. Our approach provides an alternative and exact view of the far intuitively understood home team advantage.

The derived procedure can even be used by managers of teams with limited knowledge of statistics as all necessary functions are available in MS Excel.

2 Data and methods

Methods presented in this paper are demonstrated on the English Premier League results from the 1992/1993 season to the 2016/2017 season. These results were obtained from England Football Results and Betting Odds (2017). Data for the first English Premier League season (1992/1993) was obtained from the official website Premier League Football News, Fixtures, Scores & Results (2017). This website was also used for the basic control of all data, e.g., total number of goals scored by each team in the whole

season.

The Premier League consisted of 22 teams in the first three analysed seasons and of 20 teams in the rest of seasons. A balanced schedule was used in all seasons, i.e. each team played each other team exactly two times in a season, once as a home team and once as a visiting team. This means that for each team there are 19 opponents (21 in the first three seasons) with two results in a season. These two results are combined together and used to measure home team advantage which is evaluated according to the following Definition 1. A proposed method can be used to evaluate home team advantage in the whole league over a long-term period of time or in one season. The same method can also be used to evaluate home team advantage of a single team over a long-term period of time or in one season. Application of the method for a single season data is preferred as there can be significant changes of teams that form the league and changes in squad members of teams between seasons, and these changes can cause some interpretation problems.

Definition 1 *Let T_1 and T_2 be two teams that have played together twice in one season, h_{T_i} is number of goals scored by team T_i at its home ground, and a_{T_i} is number of goals scored by team T_i away from home ($i = 1, 2$). Let the results of these two matches be*

$h_{T_1} : a_{T_2}$ *at T_1's ground (with difference viewed by T_1 as $d_{T_1T_2} = h_{T_1} - a_{T_2}$) and*

$h_{T_2} : a_{T_1}$ *at T_2's ground (with difference viewed by T_1 as $d_{T_2T_1} = a_{T_1} - h_{T_2}$).*

Then we define active, passive*, and* combined measure of home team advantage *as follows.*

Active measure of home team advantage *for team T_1 is random variable A that can take values: 1 (for $h_{T_1} > a_{T_1}$), -1 (for $h_{T_1} < a_{T_1}$), and 0 (for $h_{T_1} = a_{T_1}$), i.e. random variable A is determined as*

$$A = \text{sgn}(h_{T_1} - a_{T_1}), \tag{2.1}$$

and it indicates whether team T_1 has scored more goals at home, away from home, or whether the number of goals scored was the same.

Passive measure of home team advantage *for team T_1 is random variable P that can take values: 1 (for $a_{T_2} < h_{T_2}$), -1 (for $a_{T_2} > h_{T_2}$), and 0 (for $a_{T_2} = h_{T_2}$), i.e. random variable P is determined as*

$$P = \text{sgn}(h_{T_2} - a_{T_2}), \tag{2.2}$$

and it indicates whether team T_1 has conceded less goals at home, away from home, or whether the number of goals conceded was the same.

Combined measure of home team advantage *for team T_1 is random variable C that can take values: 1 (for $d_{T_1T_2} > d_{T_2T_1}$), -1 (for $d_{T_1T_2} < d_{T_2T_1}$), and 0 (for $d_{T_1T_2} = d_{T_2T_1}$), i.e. random variable C is determined as*

$$C = \text{sgn}(d_{T_1T_2} - d_{T_2T_1}), \tag{2.3}$$

and it indicates whether the goal difference viewed by T_1 was better at home, away from home, or whether the both differences were the same.

All three measures are defined so that value 1 means that the result was better at home, 0 means that there was no difference, and -1 means that the better result was recorded away from home. Obviously, active measure for team T_1 is passive measure for team T_2. The combination of results between the two same teams – as used in Definition 1 – eliminates the fact that teams in the league are of different quality. All three random variables can take the same values with similar interpretation; therefore, in the following parts, the combined measure C is used, and it can be easily substituted by A or P measures to obtain results for the other two measures.

Example 1 *We will demonstrate use of Definition 1 for Chelsea (T_1) and Everton (T_2) in the 2016/2017 season. The first match was played in Chelsea with result $h_{T_1} : a_{T_2} = 5 : 0$ and the second one in Everton with result $h_{T_2} : a_{T_1} = 0 : 3$, i.e. both matches were won by Chelsea, and from the view of points, home team advantage would not be identified. Differences viewed by Chelsea (T_1) were $d_{T_1T_2} = 5$ and $d_{T_2T_1} = 3$. For Chelsea we obtained*

- $A = \text{sgn}(h_{T_1} - a_{T_1}) = \text{sgn}(5 - 3) = 1$,
- $P = \text{sgn}(h_{T_2} - a_{T_2}) = \text{sgn}(0 - 0) = 0$, *and*
- $C = \text{sgn}(d_{T_1T_2} - d_{T_2T_1}) = \text{sgn}(5 - 3) = 1$.

As can be seen from the example, home team advantage (viewed by combined measure) was identified in this case as Chelsea won at the home ground by 5 goals and away by 3 goals. Active measure also indicates home team advantage in scoring goals. From the view of Everton we would get $A = 0, P = 1$, and $C = 1$ as Everton conceded less goals at the home ground and lost by a lower difference at the home ground. For each team we get as many realisations of measures in a season as there are opponents in the league (This is caused by the fact that our observation is a pair of matches, not a single match.).

The English Premier League uses a balanced schedule in all seasons with exactly two matches between each two teams. Let L denote number of teams in a league (for our data $L = 22$ or $L = 20$) then for each team in a season, there are $K = L - 1$ opponents. Random sample $C_1, C_2, \ldots, C_K$ – combined measures of home team advantage – is obtained as one season's results of given team and its opponents. C_i's are considered to be identically distributed because there are no big changes in a team during one season. Therefore, probabilities p_{-1}, p_0 and p_1 of possible outcomes $-1, 0$ and 1 are considered constant in a season. The meaning is that during a season home team advantage of a team is stationary. The second assumption is that C_i's are independent. The interpretation is that matches with one opponent do not probabilistically influence matches with other opponents.

Remark 1 *Assumption that $C_i,\ i = 1, 2, \ldots K$, are i.i.d. may not be strictly true in reality. However, it can be expected that violation of this assumption is not strong; therefore, it is used in the same sense in majority of studies that deal with sports. Without this simplification it would be impossible to use statistics for sports as every single match could be played under slightly different conditions (for example, in different weather conditions). Moreover, undermentioned methods are robust, and this simplification should not result in any problems with interpretation of obtained findings.*

Let $Z_r,\ r = -1, 0, 1$, be random variable which describes number of cases in a season where it is possible to observe home team advantage ($r = 1$), away team advantage ($r = -1$) and no advantage ($r = 0$). Obviously, for K matches in a season $Z_1 + Z_0 = K - Z_{-1}$. Vector (Z_{-1}, Z_0, Z_1) follows trinomial distribution with parameters K and p_{-1}, p_0, p_1, and probability mass function under this notation is

$$P(k_{-1}, k_0, k_1) = \frac{K!}{k_{-1}!k_0!k_1!} p_{-1}^{k_{-1}} p_0^{k_0} p_1^{k_1}, \tag{2.4}$$

where K is total number of opponents in a season for one team, 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.

Bayesian inference is used to estimate unknown parameters and confidence intervals. Non-informative distribution of parameters p_{-1}, p_0 and p_1 is set to be uniform, i.e. it does not matter where a team plays a match, and probability in Equation (2.4) is used as conditional probability of observation under given parameters, i.e. $P(k_{-1}, k_0, k_1|p_{-1}, p_0, p_1)$. This leads to posterior probability density of parameters p_{-1}, p_0, p_1 given by

$$\begin{gathered} P(p_{-1}, p_0, p_1|k_{-1},k_0, k_1) = \frac{\Gamma(K+3)}{\Gamma(k_{-1}+1)\Gamma(k_0+1)\Gamma(k_1+1)} p_{-1}^{k_{-1}} p_0^{k_0} p_1^{k_1}, \\ p_{-1}, p_0, p_1 \geq 0, \quad p_{-1} + p_0 + p_1 = 1, \end{gathered} \tag{2.5}$$

where K is total number of opponents in a season for one team, and k_{-1}, k_0, k_1 ($k_{-1} + k_0 + k_1 = K$) are observations of corresponding advantage. K, k_{-1}, k_0, and k_1 are positive integers and therefore, gamma function can be represented also by factorials, e.g., $\Gamma(K + 3) = (K + 2)!$. Equation (2.5) is probability density function of a Dirichlet distribution Dir$(\alpha_1 = k_{-1} + 1, \alpha_2 = k_0 + 1, \alpha_3 = k_1 + 1)$. Bayesian estimator of probabilities in Equation (2.4) is given (using squared-error loss function) as a mean value of this Dirichlet distribution, i.e.

$$\hat{p_r} = \frac{k_r + 1}{K + 3}, \quad r = -1, 0, 1. \tag{2.6}$$

If p_{-1}, p_0, p_1 follows Dirichlet distribution Dir$(\alpha_1 = k_{-1} + 1, \alpha_2 = k_0 + 1, \alpha_3 = k_1 + 1)$, $k_{-1} + k_0 + k_1 = K$, then marginal distribution of $p_r,\ r = -1, 0, 1$, is Beta$(\alpha = k_r + 1, \beta = K - k_r + 2)$, see (Pitman, 1993, p. 473). This can be used to find individual $(1 - \alpha_l - \alpha_u)$-confidence intervals $(\hat{p}_{r,l}, \hat{p}_{r,u})$ for each p_r which are given by quantiles of Beta distribution

$$\hat{p}_{r,l} = \text{Beta}^{-1}(\alpha_l, k_r + 1, K - k_r + 2) \tag{2.7}$$

and

$$\hat{p}_{r,u} = \text{Beta}^{-1}(1 - \alpha_u, k_r + 1, K - k_r + 2). \tag{2.8}$$

Remark 2 *These individual confidence intervals can be used for simultaneous confidence interval of all three parameters. Based on Bonferroni inequality, they form together at least a* $(1 - 3(\alpha_l + \alpha_u))$*-simultaneous confidence interval.*

For testing hypothesis it is necessary to obtain $P(p_1 > p_{-1})$ from Equation (2.5). Using results of (Omar and Joarder, 2012, p. 932) and observed values of k_1 and k_{-1} this probability is estimated as

$$P(p_1 > p_{-1}) = 1 - I_{1/2}(k_1 + 1, k_{-1} + 1), \tag{2.9}$$

where $I_{1/2}(k_1 + 1, k_{-1} + 1)$ is regularized incomplete beta function.

Remark 3 $P(p_1 > p_{-1})$ *in this paper is an estimate based on observed values of* k_1 *and* k_{-1}*. However, for better readability, the word* estimate *is omitted in the following text.*

$P(p_1 > p_{-1})$ is the probability of occurrence of home team advantage, i.e. it can be used as a measure of home team advantage (the higher value of $P(p_1 > p_{-1})$, the higher home team advantage). Hypothesis that home team advantage is real can be accepted if $P(p_1 > p_{-1}) \geq 1 - \alpha$.

3 Results

As mentioned before, the English Premier League results from the 1992/1993 season to the 2016/2017 season were analysed by the proposed method. A total of 9 746 matches were played in these seasons, and, allowing for promotion and relegation, there were 47 teams that played in at least one season in the English Premier League. Out of these teams, only six have played in each season – Arsenal, Chelsea, Everton, Liverpool, Manchester United, and Tottenham. Note that in its first three seasons the English Premier League consisted of 22 teams and of 20 teams in the following seasons.

3.1 Global results

First, results that do not distinguish among seasons nor teams are presented, i.e. all 9 746 matches are analysed at once. The hypothesis that home team advantage exists was tested (see Equation (2.9)). The hypothesis is accepted when $P(p_1 > p_{-1}) \geq 0.95$. Results for the active measure (A), passive measure (P), and combined measure (C) are listed in Table 1. We recall that in a match between teams T_1 and T_2, the active measure for team T_1 is the passive measure for team T_2; therefore, if summed up, over all seasons and teams, the numbers are the same. For a combined measure, a pair of combined matches forms a single observation. Therefore, in Table 1, the combined measure has half the number of observations compared to the active and passive measure.

Results confirm – for all three measures – the expected conclusion that it is possible to accept the hypothesis that home team advantage exists.

Table 1: Global results for English Premier League.

Measure	$l = -1$	$l = 0$	$l = 1$	$P(p_1 > p_{-1})$
Active ($A = l$)	2 669	2 572	4 505	>0.999
Passive ($P = l$)	2 669	2 572	4 505	>0.999
Combined ($C = l$)	1 288	879	2 706	>0.999

3.2 Categorization by seasons or teams

The same conclusion as in the previous part can be made when each season – without distinguishing amongst teams – is tested separately. The lowest obtained $P(p_1 > p_{-1})$ is 0.997, and it is obtained for the 2012/2013 season for active and passive measures. This means that in every single season it is possible to accept – for all three measures – the hypothesis that home team advantage exists.

Results obtained for single teams over all their played seasons offer first cases where it is not possible to accept hypothesis about home team advantage. If combined measure is used, then the hypothesis about home team advantage is not accepted for 5 teams presented in Table 2. In four out of five cases these teams played less than three seasons, and only Crystal Palace played 8 seasons (twice in a season with 21 opponents and six times in a season with 19 opponents).

For active measure, the hypothesis about home team advantage is not accepted for 9 teams and for passive measure for 12 teams. Teams for which the hypothesis about home team advantage was not accepted for at least one measure are presented in Table 3. Bold font is used for those results where the hypothesis was not accepted.

Table 2: Results for combined measure for teams over all played seasons.

Measure	Seasons	$C = -1$	$C = 0$	$C = 1$	$P(p_1 > p_{-1})$
Blackpool	1	6	2	11	0.881
Bournemouth	2	14	6	18	0.757
Cardiff	1	5	3	11	0.928
Crystal Palace	8	54	33	69	0.911
Swindon	1	6	6	9	0.773

Table 3: $P(p_1 > p_{-1})$ for all measures for teams over all played seasons.

Team	Seasons	Combined	Active	Passive
Barnsley	1	0.985	0.998	**0.928**
Blackpool	1	**0.881**	**0.685**	**0.696**
Bournemouth	2	**0.757**	0.956	**0.500**
Bradford	2	0.990	0.997	**0.945**
Cardiff	1	**0.928**	**0.910**	**0.598**
Coventry	9	>0.999	**0.884**	>0.999
Crystal Palace	8	**0.911**	**0.423**	**0.941**
Hull	5	0.995	0.991	**0.950**
Nottingham Forest	5	0.990	**0.453**	0.988
Oldham	2	0.999	0.975	**0.912**
Reading	3	0.985	**0.617**	**0.908**
Swindon	1	**0.773**	**0.760**	**0.500**
West Brom	11	0.999	>0.999	**0.925**
Wigan	8	0.963	**0.580**	0.992
Wolves	4	0.979	**0.744**	**0.874**

3.3 Categorization by seasons and teams

This part contains the final and the most detailed decomposition where single team data is analysed only in a single season. This allows us to avoid interpretation problems that can be caused by analysis of one team over several seasons where some significant changes in team members (or managers) can be made. The problems when a season is analysed without distinguishing teams (i.e. it is a mixture of all teams) are also eliminated. However, we have to expect that in many cases it will not be possible to accept hypothesis about home team advantage because the uniform distribution was considered as the prior distribution of $p_{-1}, p_0,$ and p_1 in Equation (2.4), and we possess only 19 – or, in the case of the first three seasons, 21 – observations in one season so the results have to be extremely in favour for home team advantage.

Table 4 contains the number of teams for each season where it is possible to accept – based on combined measure – the hypothesis about home team advantage. The highest number was recorded in the 2009/2010 season (17 teams out of 20), and the lowest number was recorded in the 2015/2016 season (2 teams out of 20). Numbers of teams where the hypothesis about home team advantage is accepted are in all seasons between 1 and 9 when active or passive measure are used.

Table 4: Numbers of teams for which the hypothesis about home team advantage was accepted (combined measure).

Season	Teams	Season	Teams	Season	Teams
1992/93	11	2001/02	8	2010/11	10
1993/94	5	2002/03	8	2011/12	9
1994/95	12	2003/04	7	2012/13	4
1995/96	8	2004/05	10	2013/14	5
1996/97	4	2005/06	10	2014/15	5
1997/98	9	2006/07	8	2015/16	2
1998/99	6	2007/08	12	2016/17	7
1999/00	13	2008/09	5		
2000/01	9	2009/10	17		

Table 5 contains results of all three used measures for the English Premier League season 2016/2017. All results in this table are sorted by values of $P(p_1 > p_{-1})$ obtained for combined measure. An asterisk is used for those teams where it is possible to accept hypothesis about home team advantage when a combined measure is used. Two asterisks are used for those teams where the hypothesis about home team advantage can be accepted for all three measures.

Active and passive measures can be used as auxiliary measures, e.g., to identify that home team advantage of Chelsea (see Table 5) is caused mainly by its ability to score

against the same opponent more goals at home ground than away rather than to concede with the same opponent less goals at home ground than away. Observations of active measure for Chelsea are: once $A = -1$, seven times $A = 0$, and eleven times $A = 1$ and for passive measure: six times $P = -1$, seven times $P = 0$, and six times $P = 1$.

Active and passive measures can also be used as main measures. An example situation where combined measure does not confirm general home team advantage is Swansea in Table 5. Nevertheless, active measure of Swansea indicates that this team has the ability to score with the same opponent more goals at home ground than away; obtained values are: three times $A = -1$, six times $A = 0$, and ten times $A = 1$, and for comparison, observed values of passive measure are: seven times $P = -1$, four times $P = 0$, and eight times $P = 1$.

Table 5: Results for the 2016/2017 season

Team	Combined	Active	Passive
Everton**	0.999	0.962	0.962
Leicester**	0.996	0.971	0.994
Burnley*	0.982	0.989	0.941
Tottenham*	0.975	0.941	0.954
Chelsea*	0.971	0.998	0.500
Liverpool*	0.962	0.696	0.849
Watford*	0.952	0.928	0.941
Stoke	0.941	0.849	0.954
West Brom	0.928	0.834	0.954
Hull	0.928	0.994	0.773
Bournemouth	0.916	0.927	0.598
Middlesbrough	0.806	0.910	0.407
Sunderland	0.760	0.685	0.500
Swansea	0.760	0.971	0.598
Crystal Palace	0.685	0.402	0.941
Arsenal	0.676	0.685	0.928
West Ham	0.402	0.105	0.500
Southampton	0.315	0.090	0.773
Man City	0.240	0.212	0.605
Man United	0.227	0.500	0.500

The team with the highest home team advantage in 2016/2017 season was Everton (see Table 5). This team is one of those six teams that have played in each season of the English Premier League, and the course of $P(p_1 > p_{-1})$ for Everton in all seasons is shown in Figure 1 (seasons where it is possible to accept hypothesis that home team advantage exists, i.e. where $P(p_1 > p_{-1}) \geq 0.95$, are denoted by full bullets (●)).

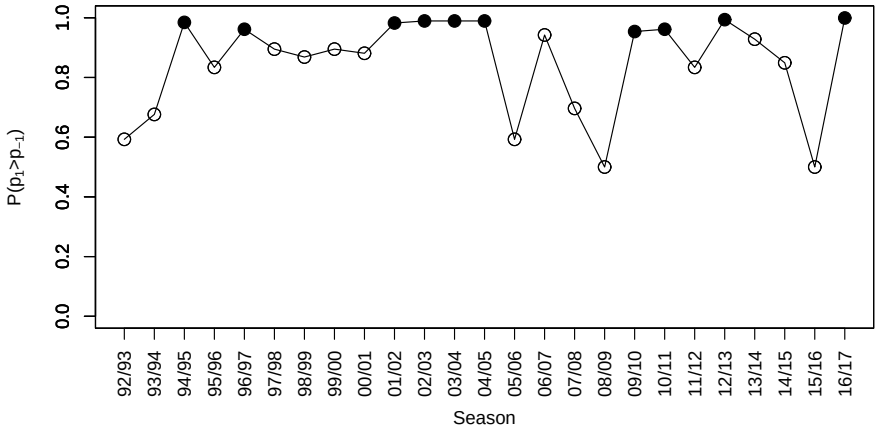

Figure 1: Evolution of $P(p_1 > p_{-1})$ for Everton.

It is also possible to estimate probability $\hat{p}_1$ (using Equation (2.6)) and 95% confidence interval $(\hat{p}_{1,l}, \hat{p}_{1,u})$ (using Equations (2.7) and (2.8)). Figure 2 shows these estimates for Everton during all seasons.

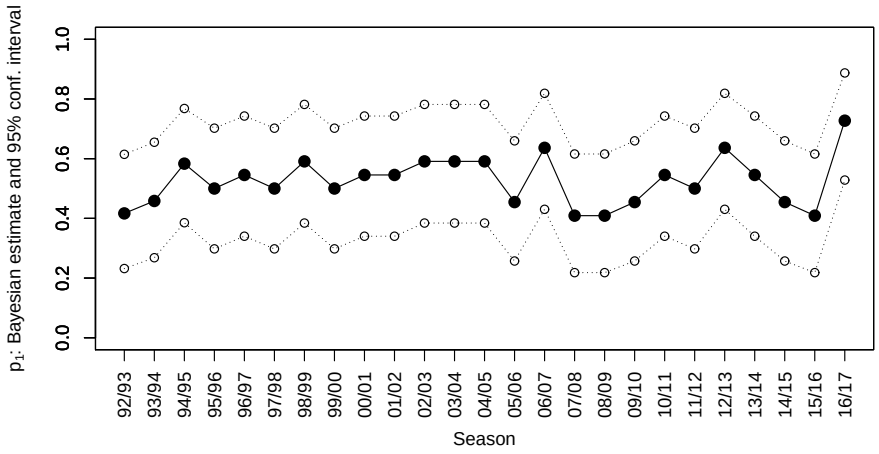

Figure 2: Evolution of Bayesian estimate and confidence interval for p_1 for Everton.

Manchester United is the team with the lowest home team advantage on the opposite side of Table 5. Value $P(p_1 > p_{-1}) = 0.227$ suggests that it is even possible to talk about home team disadvantage. This team also played all seasons of the English Premier League, and the course of $P(p_1 > p_{-1})$ for Manchester United in all seasons is shown in Figure 3 (seasons where it is possible to accept hypothesis that home team advantage exists, i.e. where $P(p_1 > p_{-1}) \geq 0.95$, are denoted by full bullets (●)).

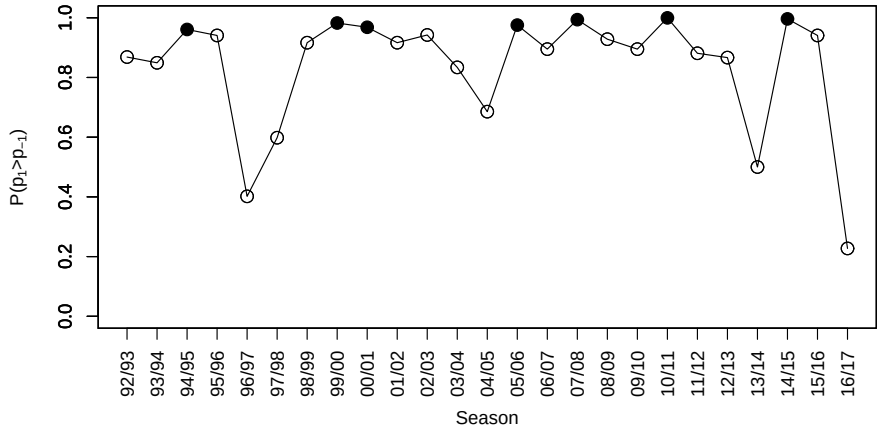

Figure 3: Evolution of $P(p_1 > p_{-1})$ for Manchester United.

Figure 4 shows estimated $\hat{p}_1$ and 95% confidence interval $(\hat{p}_{1,l}, \hat{p}_{1,u})$ for Manchester United during all seasons.

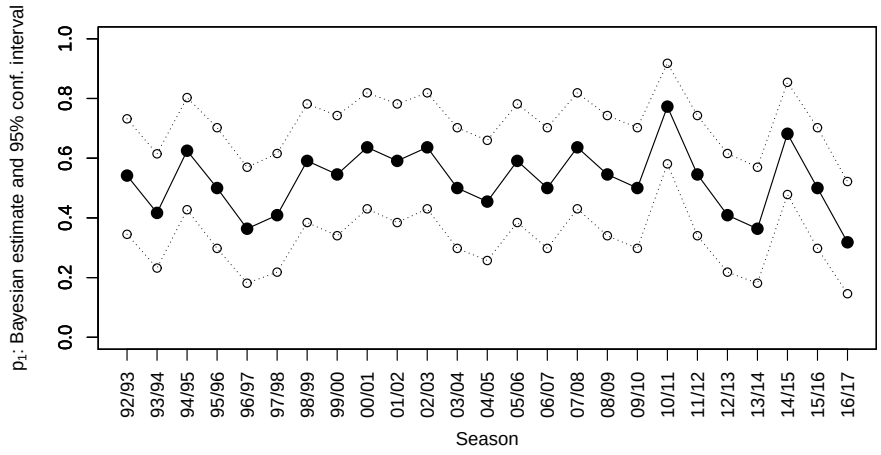

Figure 4: Evolution of Bayesian estimate and confidence interval for p_1 for Manchester United.

Results of Everton and Manchester United, once again, show that home team advantage does not mean good results, and the word *advantage* should not be understood to be beneficial. These teams are on opposite ends of Table 5; however, these teams finished season very similarly – Everton in 7th place and Manchester United in 6th place. Gained points indicates the same, Everton got 43 of its 61 points (70.5%) from home matches and Manchester United 34 of its 69 points (49.3%). Points are used only for illustration, as stated before, using them can be problematic – this can be illustrated for

Liverpool who gained 41 of its 76 points (53.9%) from home matches. Nevertheless, probability $P(p_1 > p_{-1})$ for combined measure for Liverpool is 0.962 (see Table 5), i.e. it is possible to accept hypothesis about home team advantage. This is based on observed numbers for combined measure that were: four times $C = -1$, four times $C = 0$, and eleven times $C = 1$.

Evolution of $P(p_1 > p_{-1})$ for all teams that played at least once between the 2012/13 season and the 2016/2017 season is presented in Table 6. Bold font is used for those results where it is possible to accept hypothesis that home team advantage exists. Norwich in the 2013/2014 season (18th place, 33 points) is another interesting example that home team advantage does not ensure good results. It only ensures that with the same opponent the result at home is better than result away from home; however, both can mean loss. Norwich, in the 2013/2014 season, recorded: three times $C = -1$, once $C = 0$, and 15 times $C = 1$. For example, Norwich lost 0–1 to Manchester United at home ground and 0–4 in Manchester. Obviously, 0–1 is better result than 0–4, and therefore $C = 1$, as described in Definition 1.

The last presented results are extreme values obtained for all used data. Five lowest values of $P(p_1 > p_{-1})$ are presented in Table 7 and five highest values in Table 8. These tables also contain numbers of cases where combined measure of home team advantage (C) took value of $-1, 0,$ or 1 in the referred season. It can be seen that $P(p_1 > p_{-1})$ is in many cases close to 1 but it is usually far from 0.

4 Discussion

The methods presented in this paper were demonstrated on the English Premier League data between the 1992/1993 season and the 2016/2017 season. Firstly, all of the data was tested, without distinguishing amongst seasons or teams. Hypothesis about home team advantage was accepted for all three used measures, i.e. active measure that deals with only goals scored, passive measure that deals only with goals conceded, and combined measure that uses a combination of both previous measures. These results confirm the expected conclusion that is widely accepted in literature.

Subsequently, data was divided according to seasons, teams, or both of them. First, each season was analysed separately, and for all three measures it was possible to accept the hypothesis about home team advantage. Conclusion about existence of home team advantage in each season between the 1992/1993 season and the 2011/2012 season was obtained also by Allen and Jones (2014) who used the common definition – based on points obtained at home – published by Pollard (1986). We also used this method to analyse the remaining seasons (2012/2013–2016/2017), and for these seasons, the hypothesis of the existence of home advantage was also confirmed.

Next, data for single teams over all the analysed seasons was used. This can cause some interpretation problems, and conclusions have to be made with caution. The main

Table 6: Evolution of $P(p_1 > p_{-1})$ for all teams in the seasons 2012/13–2016/17.

Team	Season				
	12/13	13/14	14/15	15/16	16/17
Arsenal	0.895	**0.962**	**0.975**	0.895	0.676
Aston Villa	0.760	0.685	0.820	0.788	—
Bournemouth	—	—	—	0.304	0.916
Burnley	—	—	0.685	—	**0.982**
Cardiff	—	0.928	—	—	—
Chelsea	0.834	0.849	0.849	0.696	**0.971**
Crystal Palace	—	0.696	0.212	0.227	0.685
Everton	**0.994**	0.928	0.849	0.500	**0.999**
Fulham	0.760	**0.962**	—	—	—
Hull	—	0.928	0.928	—	0.928
Leicester	—	—	0.895	0.605	**0.996**
Liverpool	0.773	0.748	0.834	0.696	**0.962**
Man City	**0.952**	**1.000**	0.820	0.928	0.240
Man United	0.867	0.500	**0.996**	0.941	0.227
Middlesbrough	—	—	—	—	0.806
Newcastle	0.760	0.820	**0.975**	**0.998**	—
Norwich	**0.994**	**0.998**	—	0.895	—
QPR	0.500	—	**0.996**	—	—
Reading	0.788	—	—	—	—
Southampton	0.788	0.849	0.881	0.868	0.315
Stoke	0.773	**0.975**	0.916	0.834	0.941
Sunderland	0.773	0.500	0.402	0.916	0.760
Swansea	0.788	0.928	0.867	**0.952**	0.760
Tottenham	0.500	0.895	0.928	0.613	**0.975**
Watford	—	—	—	0.806	**0.952**
West Brom	0.928	0.941	0.676	0.315	0.928
West Ham	**0.994**	0.788	**0.952**	0.500	0.402
Wigan	0.304	—	—	—	—

Table 7: Five lowest obtained values of $P(p_1 > p_{-1})$.

Team	Season	$P(p_1 > p_{-1})$	$C = -1$	$C = 0$	$C = 1$
Hull	2008/09	0.038	11	4	4
Norwich	1993/94	0.072	11	5	5
Blackburn	2003/04	0.166	10	3	6
Wolves	2011/12	0.166	10	3	6
Crystal Palace	1997/98	0.180	11	1	7

Table 8: Five highest obtained values of $P(p_1 > p_{-1})$ (more decimal places of estimates are shown only for illustration, all results can be considered as equivalent).

Team	Season	$P(p_1 > p_{-1})$	$C = -1$	$C = 0$	$C = 1$
Blackburn	2009/10	0.99999	0	4	15
Leeds	1992/93	0.99998	1	2	18
West Ham	1997/98	0.99998	1	0	18
Arsenal	1997/98	0.99993	1	2	16
Bolton	2005/06	0.99993	1	2	16

problem is that 25 seasons were analysed, and teams that appear at the beginning and the end of this time interval only have their names in common. However, if there is a general home team advantage, it should lead to acceptance of the hypothesis about home team advantage even in this case. Results showed that if combined measure is used, then it is not possible to accept the hypothesis only for 5 teams out of 47 teams analysed (out of these team, most seasons were played by Crystal Palace: 8). For active measure the hypothesis was not accepted for 9 teams (most seasons played by Coventry: 9) and for passive measure for 12 teams (most seasons played by West Brom: 11). Nevertheless, $P(p_1 > p_{-1})$ was usually high, and the lowest value 0.757 for combined measure was obtained for Bournemouth who played only two seasons. More surprising results were obtained for active and passive measure where the lowest value 0.423 was obtained for Crystal Palace in active measure. Next three lowest values were obtained for Nottingham Forest (0.453 in active measure, 5 seasons), Bournemouth (0.500 in passive measure, 2 seasons), and Swindon (0.500 in passive measure, 1 season). Together, if the number of seasons are considered, results of Crystal Palace show that they do not match other results. The reason is probably that Crystal Palace is not able to score against the same opponent more goals at home than away.

The final part with data divided by teams and seasons shows the main advantage of a newly presented method. There is no need for adjustment that is necessary in other methods, and moreover, obtained results in one season can be compared with results

from another season. Next, it is also possible to confirm the hypothesis about home team advantage for strong teams that usually win both matches in a season and obtain the same number of points from both, e.g., Liverpool in 2016/2017 season.

Each team was tested in each season to identify whether it is possible to accept the hypothesis about the home team advantage. Results of combined measure are diverse – from two teams with the home team advantage in the 2015/2016 season to 17 teams in the 2009/2010 season – and with no clear trend. Similar results are obtained when active and passive measures are used. There is also no clear trend, and number of teams for which the hypothesis about home team advantage is accepted varies between 1 and 9. Therefore, based on these results, it is not possible to see any clear change in home team advantage during the time.

Detailed results were presented for the 2016/2017 season. There are only two teams – Everton and Leicester – for which it is possible to accept hypothesis about home team advantage for all three measures. Clearly, these teams played significantly better at home than away, and they had strong home team advantage (or strong away team disadvantage). The lowest value of $P(p_1 > p_{-1})$ in combined measure was obtained for Manchester United who recorded better results against 9 opponents playing away and with six opponents playing at home. Manchester United also recorded $P(p_1 > p_{-1}) = 0.5$ for active and passive measures. This indicates that Manchester United in 2016/2017 season was affected very little by playing at home and combined measure suggests that there could be even some home team disadvantage. However, historical records (see Figure 3) show that Manchester United usually records higher home team advantage.

Result also show that it is rare to obtain $P(p_1 > p_{-1}) \leq 0.2$. For combined measure, this was achieved only in five cases (see Table 7) out of 506 possibilities (for active measure in 21 cases and for passive measure in 16 cases). On the contrary, $P(p_1 > p_{-1}) \geq 0.8$ was recorded in 361 cases for combined measure, 269 cases for active measure, and 270 cases for passive measure. These results are not surprising as the hypothesis about home team advantage is widely accepted. The interesting point of this method is, that it offers unique possibility to compare home team advantage between teams and seasons; therefore, it is possible to see huge differences in home team advantage among teams (see Table 6).

Based on our research, there is currently no direct analysis focused on home advantage of individual teams in the Premier League. It is possible to find papers related to other leagues that deal with home advantage from the point of view of individual teams, but even in this case, the analysis is not performed separately for each season, but teams are examined for all seasons played in the league. Examples of these articles are: Armatas and Pollard (2014) who analysed Greek football league, Goumas (2017) who analysed home advantage for individual teams in UEFA Champions League, and Pollard et al. (2017) who analysed home advantage in the Iranian football league.

Since the most leagues use a balanced match schedule, the procedure described in this paper can be applied to these leagues without a change. An example of a league where this procedure cannot be directly applied is the Scottish Premiership, which uses an unbalanced schedule. Nevertheless, even in this league it would be possible to apply this procedure when limited to first 22 rounds of the season where each team plays each other team exactly two times.

At the end of this section, we would like to note that home team advantage can also be seen as synonymous with away team disadvantage. Thus, the term "advantage" should not automatically be taken as a positive phenomenon. This can also be demonstrated on the example of a team that does not gain a single point in a season. Yet, it can still have a very strong home ground advantage, as it loses less at home ground. Similarly, a team that gains all its possible points in a season can have a strong home disadvantage as it wins less at home.

5 Conclusion

This paper offers an alternative approach for identification of home team advantage in results. The new method is based on goals scored rather than on points gained. This allows us to distinguish matches that look identical when points are used; for example, a 0–2 loss is not as bad as a 1–5 loss. Three measures of home team advantage were defined: active, passive, and their combination. Later, the Bayesian estimator and confidence intervals for probabilities of appropriate states – home team advantage, no advantage, and away team advantage – were derived. The last theoretical part contains test of home team advantage.

The new method was presented on the English Premier League, and results suggest that home team advantage is real; however, it cannot be taken for granted. There are also differences among three used measures, and home team advantage found in one of them does not imply home advantage in the other two measures.

The main advantage of this newly presented method is the possibility to analyse single teams without need for adjustment, and obtained results are easily comparable. Disadvantage is that this method can be used only when a balanced schedule is used (or at least when each played match can be easily paired with another match at the opposite ground).

Acknowledgment

This work was supported by the Czech Ministry of Education, Youth and Sports under Grant LO1506. We would like to thank the anonymous reviewers and the editor for their helpful comments and suggestions that contributed to improve this paper.

References

[1] Allen, M.S. and Jones, M.V. (2014): 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.

[2] Armatas, V. and Pollard, R. (2014): Home advantage in Greek football. *European Journal of Sport Science*, **14**(2), 116–122.

[3] Clarke, S.R. and Norman, J.M. (1995): Home ground advantage of individual clubs in English soccer. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **44**(4), 509–521.

[4] Dixon, M.J. and Coles, S.G. (1997): Modelling association footbal scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **46**(2), 265–280.

[5] England Football Results and Betting Odds (2017): Premiership Results and Betting Odds. `http://www.football-data.co.uk/englandm.php`.

[6] Goumas, C. (2017): Modelling home advantage for individual teams in UEFA champions league football. *Journal of Sport and Health Science*, **6**, 321–326.

[7] Jones, M.B. (2015): The home advantage in major league baseball. *Perceptual and Motor Skills*, **121**(3), 791–804.

[8] Karlis, D. and Ntzoufras, I. (2003): Analysis of sports data by using bivariate Poisson models. *The Statistician*, **52**(3), 381–393.

[9] Leite, W.S.S. and Almeida, C.H. (2018): Competitive-level and midterm effects on the magnitude of home advantage in Portuguese futsal. *International Journal of Performance Analysis in Sport*, **18**(1), 184–194.

[10] Maher, M.J. (1982): Modelling association football scores. *Statistica Neerlandica*, **36**(3), 109–118.

[11] Marek, P., Šedivá, B. and Ťoupal, T. (2014): Modeling and prediction of ice hockey match results. *Journal of Quantitative Analysis in Sports*, **10**(3), 357–365.

[12] Omar, M.H. and Joarder, A.H. (2012): Some mathematical characteristics of the Beta density function of two variables. *Bulletin of the Malaysian Mathematical Sciences Society*, **35**(4), 923–933.

[13] Pitman, J. (1993): *Probability*. New York, NY: Springer.

[14] Pollard, R. (1986): Home advantage in soccer: A retrospective analysis. *Journal of Sports Sciences*, **4**(3), 237–248.

[15] Pollard, R., Armatas, V. and Zamani Sani, S.H. (2017): Home advantage in professional football in Iran: Differences between teams, levels of play and the effects of climate. *International Journal of Science Culture and Sport*, **5**, 328–339.

[16] Pollard, R. and Gómez, M. (2014): Components of home advantage in 157 national soccer leagues worldwide. *International Journal of Sport and Exercise Psychology*, **12**(3), 218–233.

[17] Pollard, R. and Gómez, M.A. (2015): Comparison of home advantage in college and professional team sports in the United States. *Collegium Antropologicum*, **39**(3), 583–589.

[18] Pollard, R. and Pollard, G. (2005): 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.

[19] Pollard, R. and Ruano, M.A.G. (2009): Home advantage in football in South-West Europe: Long-term trends, regional variation, and team differences. *European Journal of Sport Science*, **9**(6), 341–352.

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

[21] Rooney, L. and Kennedy, R. (2018): Home advantage in Gaelic football: The effect of divisional status, season and team ability. *International Journal of Performance Analysis in Sport*, **18**(6), 917–925.

A Performing analysis in Excel

The analysis described in this paper can be performed in MS Excel even without advanced statistical knowledge. In the next steps, it is assumed that in one season each two teams play together exactly twice in the league – once at home and once away. The procedure is valid for one season and can be applied to other seasons in the same way.

The first step is to obtain data, for example from the official website of the analysed league. Next, for each match in the season, we need to find its counterpart (i.e. for the match between teams T_1 and T_2, where T_1 is a home team, we need to find the result from the match between teams T_2 and T_1, where T_2 is a home team).

The second step is to calculate Active, Passive, and Combined measure for home teams according to the Definition 1. For example, to compute Combined measure for Burnley from the pair of matches between Burnley and Swansea (see Table 9) we use

```
=sign(0-1-(2-3)),
```

where instead of numbers we use references to the relevant cells.

Table 9: Sample of result from season 2016/2017

Home Team	Away Team	First Match		Second Match	
		Home	Away	Away	Home
Burnley	Swansea	0	1	2	3
Crystal Palace	West Brom	0	1	2	0
Everton	Tottenham	1	1	2	3

The third step is to use contingency tables to obtain aggregate values for each team. This leads to results presented in Table 10 that can be used to compute point estimates, e.g., $\hat{p_1}$ for Arsenal we use Equation (2.6) as

```
=(10+1)/(19+3),
```

where instead of numbers we use references to the relevant cells (19 is sum in the row).

Confidence interval $(\hat{p}_{r,l}, \hat{p}_{r,u})$, defined in Equation (2.7) and Equation (2.8), can be computed using data from Table 10, e.g., for Arsenal we compute symmetric 95% interval for p_1, i.e. $(\hat{p}_{1,l}, \hat{p}_{1,u})$ by

```
=BETA.INV(0.025; 10+1; 19-10+2),

=BETA.INV(0.975; 10+1; 19-10+2),
```

where instead of numbers we use references to the relevant cells (19 is sum in the row).

Finally, $P(p_1 > p_{-1})$ from Equation (2.9) can be computed using data from Table 10, e.g., for Arsenal we obtain

```
=1-BETA.DIST(0.5; 10+1; 8+1; 1),
```

where instead of numbers we use references to the relevant cells.

The similar procedure can be used to analyse data for a whole season or several seasons. In all these cases we use the same procedure based on aggregate data.

As an electronic attachment to this paper, we have prepared an Excel file that can be easily used to analyse data from a single season.

Table 10: Combined measure from season 2016/2017 (sample)

Team	Combined measure		
	$C = -1$	$C = 0$	$C = 1$
Arsenal	8	1	10
Bournemouth	6	1	12
Crystal Palace	7	3	9