ELEKTROTEHNIŠKI VESTNIK 91(4): 157–183, 2024
ORIGINAL SCIENTIFIC PAPER

A quantitative analysis of home advantage in top football and basketball leagues

Timur Kulenović[1], **Jure Demšar**[1,2,†]

[1]*Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana*
[2]*Faculty of Arts, Department of Psychology, MBLab, University of Ljubljana, Aškerčeva cesta 2, 1000 Ljubljana*
[†] *E-mail: jure.demsar@fri.uni-lj.si*

Abstract. Home advantage is a phenomenon in sports where teams tend to perform better when playing at home. In this work, we explore home advantage in professional basketball and football leagues. Our results suggest that in basketball all leagues (NBA, Euroleague, Eurocup, ABA, and Slovenian league) seem to exhibit a significant degree of home advantage, it seems to be the highest in the ABA league and the lowest in the NBA. In football, four of the leagues (Premier League, Serie A, Bundesliga, and Ligue 1) showed moderate presence of home advantage, whereas La Liga showed significantly higher levels. When analysing seasonal trends, we usually observed lower home advantage in the seasons affected by the COVID-19 pandemic. We proposed a novel metric to measure home advantage at the individual game level, this metric allowed us to investigate connections between potential factors (crowd impact, referee bias, and travel fatigue) and home advantage. The crowd attendance seems to have a positive correlation with home advantage, while counter-intuitively the opposite seems to often hold for the referee bias. A side product of our work is also an extensive and carefully curated dataset, which we made publicly available for the whole research community.

Keywords: home advantage, football; basketball, Bayesian statistics; data science

Kvantitativna analiza prednosti domačega terena v najmočnejših nogometnih in košarkarskih ligah

Prednost domačega igrišča v športu je pojav, ko ekipe dosegajo boljše rezultate na domačem prizorišču v primerjavi z rezultati, ki jih dosegajo na gostujočem prizorišču. V delu obravnavamo prednost domačega igrišča v košarki in nogometu, pri čemer smo se osredotočili na pet profesionalnih lig v vsakem športu. Naš cilj je bil kvantificirati prednost domačega igrišča in jo primerjati med ligami.
Zbrali in obdelali smo množico podatkov ter ustvarili podatkovno zbirko, ki je javno dostopna širši uporabi. Za potrebe analize smo predstavili novo metriko, ki meri prednost domačega igrišča na ravni tekme, kar jo razlikuje od obstoječih metrik, ki prednost domačega igrišča merijo na ravni skupine tekem oz. sezone.
Za kvantifikacijo prednosti domačega igrišča in oceno negotovosti smo uporabili Bayesovske hierarhične modele. Rezultati kažejo, da je med izbranimi košarkarskimi ligami prednost domačega igrišča največja v ligi ABA, najnižja pa v ligi NBA. Pri nogometu smo ugotovili, da z večjo prednostjo domačega igrišča izstopa španska liga, v preostalih štiri ligah pa so rezultati podobni. Raziskali smo tudi sezonske trende, pri čemer smo pri nekaterih ligah opazili manjšo prednost domačega igrišča v času epidemije covida 19.
Proučili smo tudi povezavo med potencialnimi dejavniki in prednostjo domačega igrišča, kjer smo uporabili novoustvarjeno metriko. Tu se rezultati nekoliko razlikujejo med ligami, v glavnem pa je vpliv gledalcev pozitivno koreliran, sodniška prisotnost pa negativno korelirana s prednostjo domačega igrišča. Kljub temu izpostavljamo obstoj dejavnikov, ki jih v raziskavo nismo vključili, vendar verjetno pomembno prispevajo k prednosti domačega igrišča.

Ključne besede: prednost domačega igrišča, nogomet, košarka, Bayesova statistika, podatkovna znanost

Received 13 May 2024
Accepted 1 July 2024

1 INTRODUCTION

Sports analytics is a scientific field where state-of-the-art data science methodology is applied to the domain of sports. This is particularly interesting to managers and owners of sports teams since they can facilitate the insights gained through data analysis for informed data-driven decision-making where each team manager strives to make optimal decisions in every aspect of their work to gain a competitive advantage. A better understanding of the home advantage would allow teams to exploit it more efficiently.

Home advantage is a phenomenon in sports that describes the benefit that the hosting team has over the visiting team. It appears to be present since the start of organised football [1] and it occurs in most team sports. The reasons for this occurrence have been attributed to several aspects, such as referee bias, crowd effects, travel effects, familiarity with the playing field, territoriality, specific tactics, rule factors, and psychological factors [1]. A lot of existing research focused on statistical analyses of the development of the phenomenon through

time and on differences in the magnitude of home advantage between different sports. However, to the best of our knowledge, it is still not exactly clear how different previously mentioned factors affect the home advantage.

A lot of research on the phenomenon of home advantage has been already conducted, particularly in football. Pollard [2] has introduced the home advantage ratio in football as the ratio between the number of points won by the home team and the number of points won in total. When we can claim with high certainty that the ratio is higher than 50%, then we say that home advantage is present. In the 2015/2016 season, the average value of the ratio for the top 10 European leagues was estimated to be $58.25 \pm 2.95\%$. Results from older seasons yield a higher ratio, meaning that home advantage seems to be in decline [3]. Furthermore, a substantial decline was also found in English football from 1974 to 2018. Interestingly, an approximately equal decline seems to be present across different divisions. Even though the advantage seems to be in decline, it is still present in the last couple of years [4].

When analysing 19 European football leagues between the 2007/2008 season and the 2016/2017 season, the Greek league had the highest, while the English Football League Two (fourth rank) had the lowest home advantage. Based on the results, a hypothesis that home advantage is reduced in the lower-level leagues can be stated and connected to the crowd effect [5]. By analysing football matches of national teams Pollard and Armatas [6] found out that, besides crowd size, the altitude and the number of time zones crossed by the visiting team were significantly related to the number of points won by home teams. Additionally, every 100 meters of altitude difference is connected with the jump of expected probability for the home team winning the match by 1.1 percentage points [7]. Pollard and Armatas [6] also found out that significantly more red cards were issued, and more penalties were awarded against the away team.

To isolate some potential factors of home advantage, such as players' familiarity with the stadium and travel fatigue, Ponzo and Scoppa analysed same-stadium derbies [8]. They concluded that there is crowd support's effect on the home advantage generated through the encouragement of players' performance. Furthremore, the crowd tends to affect the referee's decision in favour of the home team. In the football Champions League and Europa League, referees issued 25% and 10% more yellow cards to away teams than to home teams. The higher level of home team bias in the Champions League appeared mainly due to higher crowd densities [9]. Referee bias has been more thoroughly researched on English Premier League matches by Boyko et al. [10]. They found out that the referee bias is not omnipresent and varies between referees. Similar to [8], in order to eliminate some factors, Boudreaux et al. [11] performed an analysis on the case of matches between the basketball teams of Los Angeles Lakers and Los Angeles Clippers, who both play their home matches in the Staples Center in Los Angeles. Due to crowd effects, it was estimated that the increase in the likelihood of home team victory is between 21 and 22.8 percentage points. Additionally, Sors et al. [12] have also considered the level of competitive anxiety of referees. They concluded that the crowd noise does not seem to affect the referees' decisions unless we consider the anxiety because external factors might more easily influence the decisions of referees with high anxiety. Neural networks have also been used to analyze home advantage in the NBA.The conclusions were that attendance, altitude, and market size were not connected to home advantage. However, the style of play seems to be connected as teams that made more two-point and free-throw shots saw larger advantages at home. Another interesting study was conducted by Gomez and Pollard [13], they state that in some European countries basketball teams from capital cities have a significantly lower home advantage than teams from other cities. Pollard et al. [14] also completed a comprehensive analysis of home advantage in different sports and countries. They state that basketball and handball have the highest home advantage, which was the most prominent in Bosnia and Herzegovina, with other Balkan nations also well above average.

The fact that a large number of sports matches from 2019 to 2021 have been played without spectators, due to COVID-19, gives us a unique opportunity for analysis, allowing us to isolate certain factors of home advantage. For example, many matches were forcefully played without attendance, which removes the home crowd effect. Indeed, some research on this matter has already been conducted, with the results showing that the home advantage has dropped significantly in games without spectators [15], [16], [17], [18]. A significant drop in home advantage ratio was observed in the German Bundesliga (a 10% drop), turning home advantage into home disadvantage [15]. In contrast, no change was observed in the second and third divisions of German football [18]. Sors et al. [17] also revealed a reduced home advantage and the absence of referee bias in ghost games (games without spectators). These results support the claim that, among all the factors contributing to home advantage and referee bias, crowd attendance has a relevant role. Thus, it seems like fans can significantly affect the outcomes of football matches [16], [17].

The main objective of this study was to conduct a thorough statistical analysis of home advantage in basketball and football. We selected five football competitions and five basketball competitions, collected the data, and then calculated and evaluated the level of the

home advantage in each league using different metrics and Bayesian models. We compared the amount of home advantage between the leagues and also analysed how it changes through time. Next, we chose three potential factors that could impact the degree of home advantage. Since these factors are not directly measurable, we had to come up with proxy variables. We quantified the connection between the proxy variables and home advantage in each league. In the process of searching for necessary data to conduct the complete analysis of this work, we found out that there are some limitations in the granularity and availability of the data. Consequently, we collected, cleaned, and published the basketball and football data of high granularity. We appraise publishing this data as a valuable contribution not only to this work, but also for potential future research in sports analytics.

2 METHODS

In this section we first describe how we acquired the data that was later used in our statistical analyses. Next, we introduced the metrics that we used for quantifying home advantage. In the final part, we describe the models behind the Bayesian statistical analyses supporting our results.

2.1 Data acquisition

To conduct a thorough analysis of home advantage in basketball and football, as outlined in the Introduction section, we had to acquire an extensive dataset containing very granular match data. Some sport-related datasets are publicly available. However, the objective of this analysis required relatively specific information (such as the number of spectators or number of fouls in the match) that was not present in these datasets.

Furthermore, for most of the leagues that we wanted to include, free official APIs were not available. Therefore, we had to use the web scraping technique (web harvesting or web data extraction) to obtain the necessary data. We did this time-consuming process with Python libraries Selenium [19], Requests [20] and BeautifulSoup [21].

The collected datasets exceed the requirements of this thesis in terms of the granularity of the data and the purpose of obtaining the data was not only to get the required data, but our objective was also to create an extensive dataset of sports data that would be publicly available. The collected datasets should ease the process of obtaining the data for other researchers working on basketball or football-related projects.

First, we scraped an extensive amount of data for the following basketball competitions:

- League of Adriatic Basketball Association (ABA),
- Eurocup (EC),
- Euroleague (EL),
- Slovenian Basketball League (SLO) and
- National Basketball Association league (NBA).

A high-level description of the datasets is presented in Table 1. The dataset, along with the source code, is available in an open GitHub repository https://github.com/timurkulenovic/basketball-dataset [22].

For the ABA and NBA leagues, data is available before the season of 2007/2008. However, to use the same time-frame for all leagues, we did not use any data before the season of 2007/2008. Table 2 shows the number of matches we included in our work.

Second, we obtained the data for the following football leagues:

- Premier League (English league),
- Bundesliga (German league),
- Ligue 1 (French league),
- Serie A (Italian league) and
- La Liga (Spanish league).

These football datasets with the source code behind their acquisition are also available at our GitHub repository https://github.com/timurkulenovic/football-dataset [23]. The datasets are not as granular as the basketball datasets. However, they still contain much information for each game. The first available season is 2009/2010 (there is some data available before, but it is limited), and the last available season is 2022/2023. We used all of the seasons for the analysis. Table 3 presents summary data about each used league.

2.2 Metrics

In this analysis, we use two home advantage metrics, the first one quite common in related work. We call this metrics HA_{SEASON} and it denotes the percentage of the team's points won at home throughout the whole season. This metrics can be used to home advantage at either the season, the league, or the sport level. However, to conduct our analysis in the desired way, we needed another metric that measures the home advantage on the level of a single game. Consequently, we propose the HA_{GAME} metric, which serves as a measure of the home advantage on a game level and, therefore, enables us to quantify the home advantage for every game.

2.2.1 Home advantage on level of the whole season: As mentioned, HA_{SEASON} metric measures home advantage on a team level in the scope of one season. It represents the amount of the team's points gained at home divided by the team's total points:

$$HA_{SEASON} = \frac{Points_H}{Points_H + Points_A}, \qquad (1)$$

where $Points_H$ denotes the points that the team gained at home and $Points_A$ denotes the points that the team gained away in the scope of a season. In football, the team gets 3 points for a win, 1 point for a draw, and 0 points for a loss. In basketball, the winner gets 2 points and the loser gets 0 points.

	ABA	Eurocup	Euroleague	SLO	NBA
First season available	2001/2002	2007/2008	2007/2008	2007/2008	2000/2001
Main Info	✓	✓	✓	✓	✓
Box Score	✓	✓	✓	✓	✓
Play by play	✓	✓	✓	✓	
Score evolution	✓	✓	✓	✓	
Shots	✓	✓	✓		
Team comparison		✓	✓	✓	
Venues	✓	✓	✓	✓	✓

Table 1.: **Description of the basketball dataset.** Cells with the checkmarks denote that data is available for the corresponding league. The last season available is 2022/2023 for all competitions.

	ABA	Eurocup	Euroleague	SLO	NBA
2007/2008	196	326	227	217	1316
2008/2009	185	150	184	222	1315
2008/2010	185	156	184	218	1312
2010/2011	185	155	185	143	1311
2011/2012	184	156	184	158	1074
2012/2013	185	156	249	158	1314
2013/2014	184	362	249	179	1319
2014/2015	194	306	247	189	1311
2015/2016	189	306	246	159	1316
2016/2017	191	146	255	201	1309
2017/2018	142	184	256	185	1312
2018/2019	143	186	256	154	1312
2019/2020	125	168	252	83	1142
2020/2021	166	185	324	138	1165
2021/2022	197	189	295	157	1317
2022/2023	202	195	324	153	1314
Total	2853	3326	3917	2714	20459

Table 2.: **Number of basketball matches (by season and league) used in our work.** We excluded games played at neutral locations. Note that a smaller amount of games in 2019/2020 is due to the COVID-19 outbreak.

	Premier League	La Liga	Ligue 1	Serie A	Bundesliga
Games per season	380	380	380	380	306
Total games	5320	5320	5218	5319	4284

Table 3.: **Number of football matches used in our work.** Data for each league consists of 14 seasons – from 2009/2010 to 2022/2023. Season 2019/2020 in Ligue 1 ended without all the matches being played. One Serie A match in season 2012/2013 was not played.

2.2.2 Home advantage on level of an individual game: One of the goals of this work is to quantify the effect of different factors on home advantage. To fulfil this goal, we proceeded with quantifying the home advantage for each game. This requires defining a metric HA_{GAME} that quantifies home advantage on a game level. The gist of the metric is to compare the expected point difference δ_{EXP} between two teams and the observed point (or goal) difference δ_{OBS} between two teams. The main idea is that the difference between δ_{OBS} and δ_{EXP} yields the quantified home advantage of the game. Obtaining δ_{OBS} is straightforward. We simply calculate the difference between points scored by the home team and points scored by the away team. Next, we need a value that measures the expected point difference δ_{EXP}. To obtain it, we use the averages of score difference in a season for home and away teams. The expected difference δ_{EXP} is the difference between the home team's average and the away team's average.

Rewriting the idea in a bit more concise manner, HA_{GAME} measures home advantage on a game level between the home team A and the away team B in season i. $S(T, i)$ denotes team T's average score difference in season i:

$$\begin{aligned} \delta_{EXP} &= S(A, i) - S(B, i), \\ \delta_{OBS} &= Points_A - Points_B, \\ HA_{GAME} &= \delta_{OBS} - \delta_{EXP}. \end{aligned} \tag{2}$$

Let's take an example from the 2022/2023 Premier League season, when Arsenal scored 88 goals and conceded 43 goals on 38 matches, $S(Arsenal, 2022/2023) = \frac{88-43}{38} = 1.184$, whereas Tottenham scored 70 goals and conceded 63 goals, $S(Tottenham, 2022/2023) = \frac{70-63}{38} = 0.184$. The match between Arsenal and Tottenham game ended 3-1. The expected difference δ_{EXP} was $1.184 - 0.184 = 1$, while the observed difference δ_{OBS} was 3 - 1 = 2.

Finally, HA_{GAME} is $2 - 1 = 1$.

2.3 Factors influencing the home advantage

One of the goals of our study was to quantify the factors that are considered to have some effect on the home advantage and analyze the level of their correlation with the home advantage. Based on the availability of the data, we selected three factors: **referee bias**, **crowd impact** and **travel fatigue**. None of these factors is directly measurable, so we had to introduce proxy variables that try to serve in place of them. Furthermore, we must be aware that there are several other confounding factors that are not included in our data and thus not included in the analysis. Because of this, we must be very cautious when making any kind of claims about the direct influence (causation) of a certain factor on the home advantage. For example, it is known that teams that are playing well and achieving good results have a higher attendance, so higher attendance might be the effect of a team doing well and not vice-versa. The goal is not to build a high-performance model that would predict the home advantage with high accuracy but to check the connection between the chosen factors and the home advantage.

2.3.1 Referee bias: When discussing the home advantage in football, basketball or other similar sports, a commonly mentioned factor is the referee bias. It is widely believed that the referees sometimes, intentionally or not, help certain teams with unfair decisions. One of the reasons for this is the pressure from the crowd. Therefore, we expect the referee bias to be more in favour of the home teams rather than the away teams. It is, however, not straightforward to quantify the level of referee bias, as there is no objective variable that would measure the referee bias. Hence, we created a metric called $RBIAS$ that quantifies the referee bias based on the called fouls. The idea of this variable is similar to the concept of HA_{GAME}. We compare the observed difference of committed fouls F_{OBS} with the expected difference F_{EXP}, which is based on the teams' averages in a season. As such, $RBIAS$ measures the referee bias on a game level between the home team A and the away team B in the season i. The notation $C(T, i)$ denotes the team T's average of the foul difference in season i and is calculated as the difference between the team's total drawn fouls and the team's total committed fouls. The given difference is then divided by the number of the games that the team played in season i. With $DFouls_A$ and $DFouls_B$ we denote the observed number of drawn fouls in the game for teams A and B, respectively:

$$\begin{aligned} F_{EXP} &= C(A, i) - C(B, i), \\ F_{OBS} &= DFouls_A - DFouls_B, \\ RBIAS &= F_{OBS} - F_{EXP}. \end{aligned} \tag{3}$$

To illustrate the idea, let's take an example from the 2007/2008 Euroleague game between Tau Ceramica (home team) and Union Olimpija (away team). In the season 2007/2008 Tau Ceramica drew 480 fouls and committed 448 fouls in 23 games, while Union Olimpija drew 253 fouls and committed 345 fouls in 14 games:

$$\begin{aligned} C(Tau\ Ceramica, 2007/2008) &= \frac{480 - 448}{23} = 1.39, \\ C(Union\ Olimpija, 2007/2008) &= \frac{253 - 345}{14} = -6.57. \end{aligned} \tag{4}$$

The expected difference F_{EXP} was $1.39 - (-6.57) = 7.96$. In the match, Tau Ceramica drew 26 fouls and Union Olimpija drew 16. The observed difference δ_{OBS} was 26 - 16 = 10. Finally, $RBIAS = 10 - 7.96 = 2.04$.

We must be aware that this newly introduced variable measures the number of fouls teams draw and commit compared to the expected number of fouls. We assume that the average of the variable being above 0 originates from the referee bias. However, it could also be that teams draw more fouls at home, because they play in such a way at home due to some other factors. Furthermore, for the assessment we only used fouls (fouls in football and personal fouls in basketball). We could expand this variable to consider personal fouls, unsportsmanlike fouls, technical fouls (in basketball), yellow cards, red cards, and penalty kicks (in football). However, we wanted it to quantify something similar in both sports, so we kept the simple definition that only includes standard fouls.

2.3.2 Crowd impact: Another factor that is perceived to have an important impact on the home advantage is the crowd effect. It is believed that loud and supportive chants motivate the home team to play better. Again, it is difficult to quantify the crowd effect directly, so we used the attendance number information. We introduce the ATT metric, which represents the ratio between the number of spectators in the game and the maximum number of spectators ever recorded in the arena where the game was played. ATT is represented as a number on $[0, 1]$ interval that describes how full the venue is:

$$ATT = \frac{n_{spectators}}{\max\limits_{games\ in\ arena}(n_{spectators})}. \tag{5}$$

2.3.3 Travel fatigue: The third factor included in our analysis is the travel fatigue that players experience when they have to travel to an away venue. Similar to the other factors, there is no simple way of quantifying travel fatigue. Instead, we used the air distance the away team had to travel from their hometown to the away arena. To have the variable in the same range for all the leagues, we normalise the variable by dividing the distance by the league's median travel distance:

$$DIST = \frac{\text{air distance}}{\text{median over air distances in the league}}. \quad (6)$$

This way, we obtain a value that tries to quantify how much fatigue the travel caused normalised by the league's standards. There are some shortcomings of this approach. We assume that teams always travel to the away arena from their home location, which is not always true because teams on tight schedules often travel from the location of their previous away game. Furthermore, teams use different means of transportation (for example, a bus instead of a plane) for the same distance. Consequently, the same air distance can cause different levels of travel fatigue.

2.4 Statistical modelling

We used Bayesian statistics to analyse the results. All analyses were conducted using Stan – a state-of-the-art platform for executing modern Bayesian statistical analyses [24]. We used Stan's default (non-informative) priors in all analyses.

To distinguish reported Bayesian probabilities from frequentist p-values we denote them with a capital P. Unlike p-values, the reported probabilities directly describe the probability by which we can claim that our hypotheses are true or not. The probability that the opposite of our claim is true can be calculated as $1 - P$. We used Monte Carlo Standard Error (MCSE) to estimate uncertainty in our quantifications. Since MCSE was in all cases lower than 1%, we decided to omit it for the sake of brevity.

When comparing leagues between each other through the HA_{SEASON} and HA_{GAME} metrics, we used the following Bayesian hierarchical normal model:

$$\begin{aligned} HA &\sim N(\mu_{HA_{season_i}}, \sigma^2_{HA_{season_i}}), \\ \mu_{HA_{season_i}} &\sim N(\mu_{HA_{league}}, \sigma^2_{HA_{league}}), \end{aligned} \quad (7)$$

where HA is either HA_{SEASON} or HA_{GAME}. With this model, we obtain the $\mu_{HA_{league}}$ posterior distribution and n_{season} posterior distributions of $\mu_{HA_{season_i}}$ for each of the five leagues. We first used the HA values to get the season-level parameters. These were then used to obtain the league-level parameters.

To analyze $RBIAS$, we use the two-level hierarchical normal model, with the leagues on the first level and the seasons on the second level:

$$\begin{aligned} RBIAS &\sim N(\mu_{RBIAS_{season_i}}, \sigma^2_{RBIAS_{season_i}}) \\ \mu_{RBIAS_{season_i}} &\sim N(\mu_{RBIAS_{league}}, \sigma^2_{RBIAS_{league}}). \end{aligned} \quad (8)$$

We analyze the impact of $RBIAS$, ATT, and $DIST$ on HA_{GAME} by using Bayesian linear regression, which can be formalized as:

$$\begin{aligned} HA_{GAME} \sim & N(\alpha + \beta_{RBIAS}\text{RBIAS} + \\ & \beta_{ATT}\text{ATT} + \beta_{DIST}\text{DIST}, \sigma^2). \end{aligned} \quad (9)$$

All three independent variables, as well as the dependent variable, are calculated at the game level. Therefore, each game corresponds to one data point when fitting the linear regression model. Since the variable $DIST$ was created using the league-level information, we modelled each league separately. We have five separate linear regression fits with one fit for each league. The intercept α is included because it can be interpreted as the value of HA_{GAME} if all the included home advantage factors were not present (i.e., they are all 0). By observing how close to 0 it is, the value of the intercept gives insight into how much of the HA_{GAME} can be attributed to the three included factors.

3 Results

In the first part of our results section, we describe our findings when comparing the HA_{SEASON} and HA_{GAME} metrics across leagues and across seasons (years) for both basketball and football. In the second part, we investigate how various factors (crowd impact, referee bias, and travel fatigue) might influence home advantage.

3.1 Comparing leagues and seasons in basketball

The results of our analysis for the HA_{season} metric in basketball are visualised in Figure 1. Our analysis shows that teams achieved 59%-67% of their wins at home. In the large majority of the cases, teams achieved more than half of their wins at home, but there are cases in each league where teams achieve more wins away than at home, i.e., $HA_{SEASON} < 0.5$. In the Eurocup and Euroleague, we observe a bit higher variance of HA_{SEASON}. It could be because teams played fewer games in one season of these competitions, the competition was organized into groups of 6 teams, because of this each team played only ten games.

The differences between the leagues might not be immediately clear from the histograms but are more evident in the posterior distributions of $\mu_{HA_{SEASON}}$, displayed on the left side of Figure 1. NBA and the Slovenian league semm to have the smallest amount of home advantage, while ABA again has the highest level of home advantage. The distributions for the Eurocup and the Euroleague seem to be quite aligned, which is confirmed by quantified comparisons between the leagues in Table 4, where $P(\mu_{HA_{SEASON_{Eurocup}}} > \mu_{HA_{SEASON_{Euroleauge}}}) = 0.586 \pm 0.011$. Furthermore, from the probabilities in the table, we can conclude that we have one group of leagues (NBA and the Slovenian league) that very likely has lower $\mu_{HA_{SEASON}}$ than another group (ABA, Eurocup and

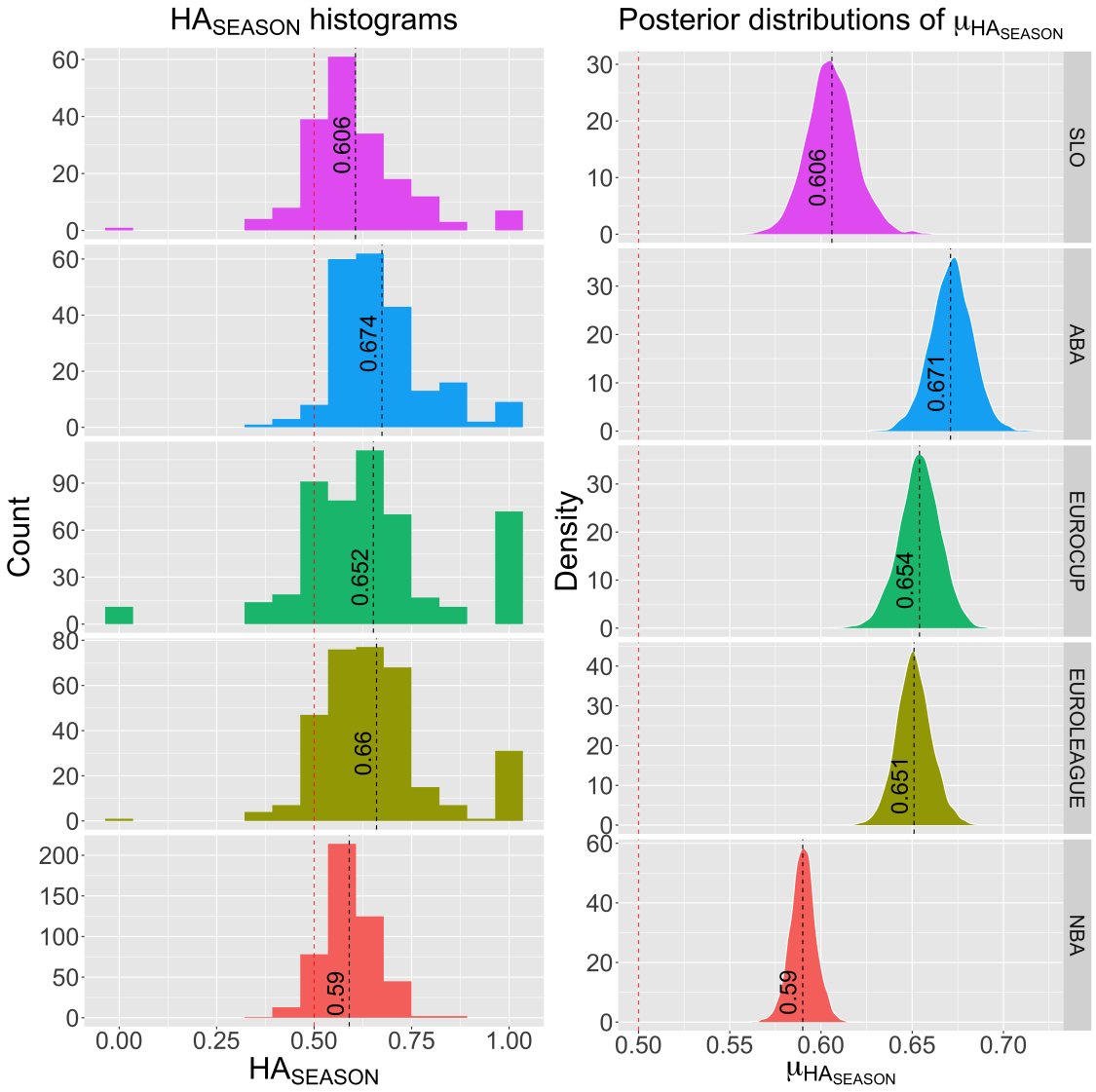

Figure 1.: **Histograms for the HA_{SEASON} metric (left) and posterior distributions calculated with our Bayesian anaysis of the $\mu_{HA_{SEASON}}$ parameter (right) for basketball leagues.** The black dashed line represents the distribution mean. The red line represents the value around which the distribution would be centered (0.5) if there was no home advantage. In the left part of the figure, we see that there are cases in all the leagues when teams achieve more points in away games than in home games. However, the right part shows that the means are certainly above 0.5. ABA is likely to have the highest value of the metrics, followed by Euroleague and Eurocup, while NBA and the Slovenian league seem to have the lowest $\mu_{HA_{SEASON}}$ value.

	SLO	ABA	EC	EL	NBA
SLO	-	≈ 0	0.005 ± 0.001	0.006 ± 0.002	0.857 ± 0.006
ABA	≈ 1	-	0.865 ± 0.007	0.913 ± 0.005	≈ 1
EC	0.995 ± 0.001	0.135 ± 0.007	-	0.586 ± 0.011	≈ 1
EL	0.994 ± 0.002	0.087 ± 0.005	0.414 ± 0.011	-	≈ 1
NBA	0.144 ± 0.006	≈ 0	≈ 0	≈ 0	-

Table 4.: **Comparison of $\mu_{HA_{SEASON}}$ for basketball leagues.** Each cell represents $P(\mu_{HA_{SEASON_i}} > \mu_{HA_{SEASON_j}}) \pm$ MCSE. ABA, Eurocup and Euroleague surely have higher $\mu_{HA_{SEASON}}$ than NBA and the Slovenian league. Furthermore, ABA is likely to have a higher value than Eurocup and Euroleague, but the probability is not very close to 1.

Euroleague) – $P(\mu_{HA_{SEASON_{ABA,Eurocup,Euroleague}}} > \mu_{HA_{SEASON_{NBA,SLO}}}) = 0.9981 \pm 0.0005$. Strong claims about intra-group comparisons cannot be made.

NBA and the Slovenian league consistently have the lowest home advantage and ABA has the highest $\mu_{HA_{SEASON}}$ in almost every season, while the Eurocup and Euroleague consistently seem to be somewhere in between. The drop during the COVID-19 seasons is visible for NBA, ABA and Euroleague, but it is not very significant for the Slovenian league and Eurocup.

The second metric, HA_{GAME}, is distributed normally with mean values ranging from 2.7 to 4.4, as shown in the histograms in the left part of Figure 3. This means that home teams gain a 2.7–4.4 point higher score difference on average compared to the expected score difference. Note that the variances of these distributions are relatively high, i.e., it is not rare that HA_{GAME} is negative. However, the posterior distributions of $\mu_{HA_{GAME}}$ parameters in the right part of the figure suggest that we are extremely confident that the mean is positive. Once again, ABA has the highest amount of home advantage, followed by the Eurocup and Euroleague, while NBA and the Slovenian league have the least amount of it. This consistency in conclusions between the established HA_{SEASON} metric and the newly introduced HA_{GAME} metrics gives as an assurance that our new metric is a viable descriptor of home advantage. This is important for more detailed analyses that follow in the second part of this work.

By observing the results in Table 5, we can claim that ABA has higher $\mu_{HA_{GAME}}$ value than the Eurocup and the Euroleague with probabilities 0.962 and 0.935, respectively. The Eurocup and the Euroleague seem to have quite equal amounts of home advantage according to this metric. The probabilities of these two competitions having higher $\mu_{HA_{GAME}}$ than the Slovenian league is fairly high (0.881 for the Eurocup and 0.938 for the Euroleague). Nevertheless, we can be confident with a 0.953 probability, that even the Slovenian league has a higher value than NBA. We are confident that ABA is the league with the highest $\mu_{HA_{GAME}}$ and that NBA is the one with the lowest value.

We show the visualisation of $\mu_{HA_{GAME}}$ over the seasons in Figure 4. Overall, the posterior distributions are very similar to those in Figure 2. ABA seems to be consistently the league with the highest home advantage, while NBA seems to have the lowest home advantage. One visible discrepancy of $\mu_{HA_{GAME}}$ compared to the other metric is that ABA has a lower value than the Euroleague and Eurocup in 2007/2008, as this is not the case with $\mu_{HA_{SEASON}}$.

To sum up, both metrics yield similar results. ABA very likely has the highest amount of home advantage, the Eurocup and Euroleague follow, while the Slovenian league and NBA have the least amount of home advantage. The results also show that this order stays consistent over the seasons without any major discrepancies across the metrics.

3.2 Comparing leagues and seasons in football

In the left part of Figure 5, we observe that the HA_{SEASON} values are approximately normally distributed with a mean around 0.59 except for La Liga, which has a mean above 0.61. In the right part of the Figure 5 we visualise the posterior distributions of $\mu_{HA_{SEASON}}$. Once more, La Liga has the most prominent home advantage, whereas the other four leagues seem to be more or less similar to each other in terms of their $\mu_{HA_{SEASON}}$ posterior distributions.

We show the probabilities of comparisons between the leagues in Table 6. The second row consists of the values estimating the probability that La Liga has higher $\mu_{HA_{SEASON}}$ than the other leagues. The lowest probability in this row is 0.938 (for Ligue 1), which confirms that La Liga very likely has the highest $\mu_{HA_{SEASON}}$. The probabilities that estimate the relationships between the other leagues are too small to be able to claim anything with high confidence.

We visualise the $\mu_{HA_{SEASON}}$ posterior distributions over the seasons in Figure 6. Spanish La Liga is quite consistent in having the highest home advantage over the seasons, which also explains the higher probabilities in Table 6. Bundesliga seems to have lower values in the first four seasons but was consistently somewhat in the middle for the rest of the seasons. We also observe a drop in the COVID-19 seasons 2019/2020 and 2020/2021. $P(\mu_{HA_{SEASON}})$ drops below 0.5 for some of the leagues in 2020/2021. Except for these two seasons, the distributions are positioned approximately around the same values over the seasons.

As shown in Figure 7, HA_{GAME} in football approximately follows the normal distribution with a mean ranging from 0.315 to 0.429 for different leagues. The interpretation of this is that home teams in football gain a bit less than half of a goal of the advantage on average. It happens quite often that HA_{GAME} is below zero. Nevertheless, based on the visualisation in the right column of Figure 7, we can be confident that $\mu_{HA_{GAME}}$ (the expected mean) is positive for all football leagues. Again, as with the HA_{SEASON} metric, $\mu_{HA_{GAME}}$ for La Liga stands out a bit when comparing leagues' posterior distributions for the parameter.

Looking at the second row in Table 7, we see that the probabilities of La Liga having higher $\mu_{HA_{GAME}}$ than the other leagues are above 0.9, where the lowest out of these four is the probability that La Liga has a higher value than the Bundesliga (0.926). Furthermore, there is 0.897 chance that the Bundesliga has more home advantage than Serie A, which seems to have the lowest $\mu_{HA_{GAME}}$.

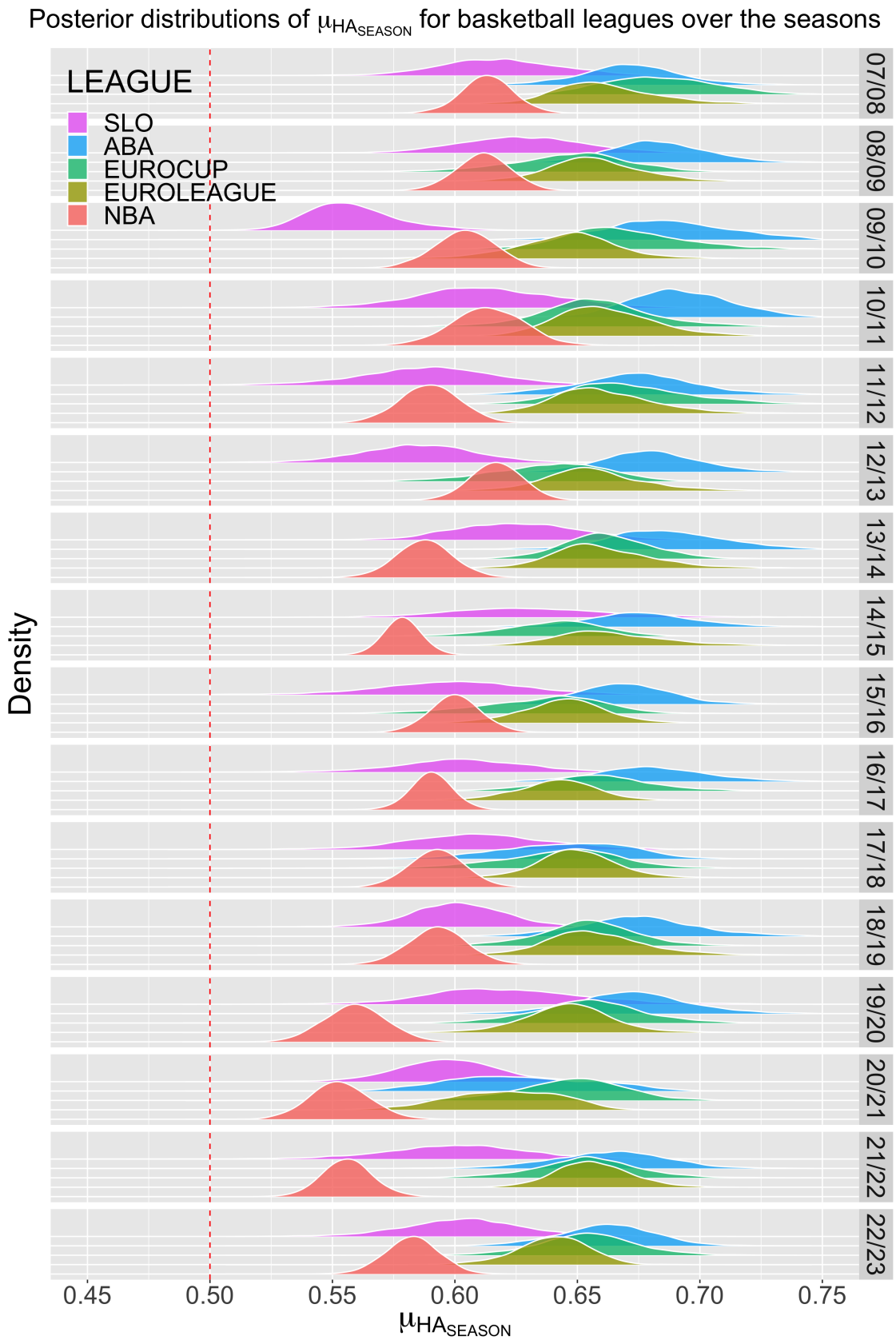

Figure 2.: $\mu_{HA_{SEASON}}$ **posterior distributions for basketball leagues over the seasons.** The order of the leagues seems to be fairly consistent over the seasons, with ABA having the highest value and NBA and the Slovenian league with the lowest value.

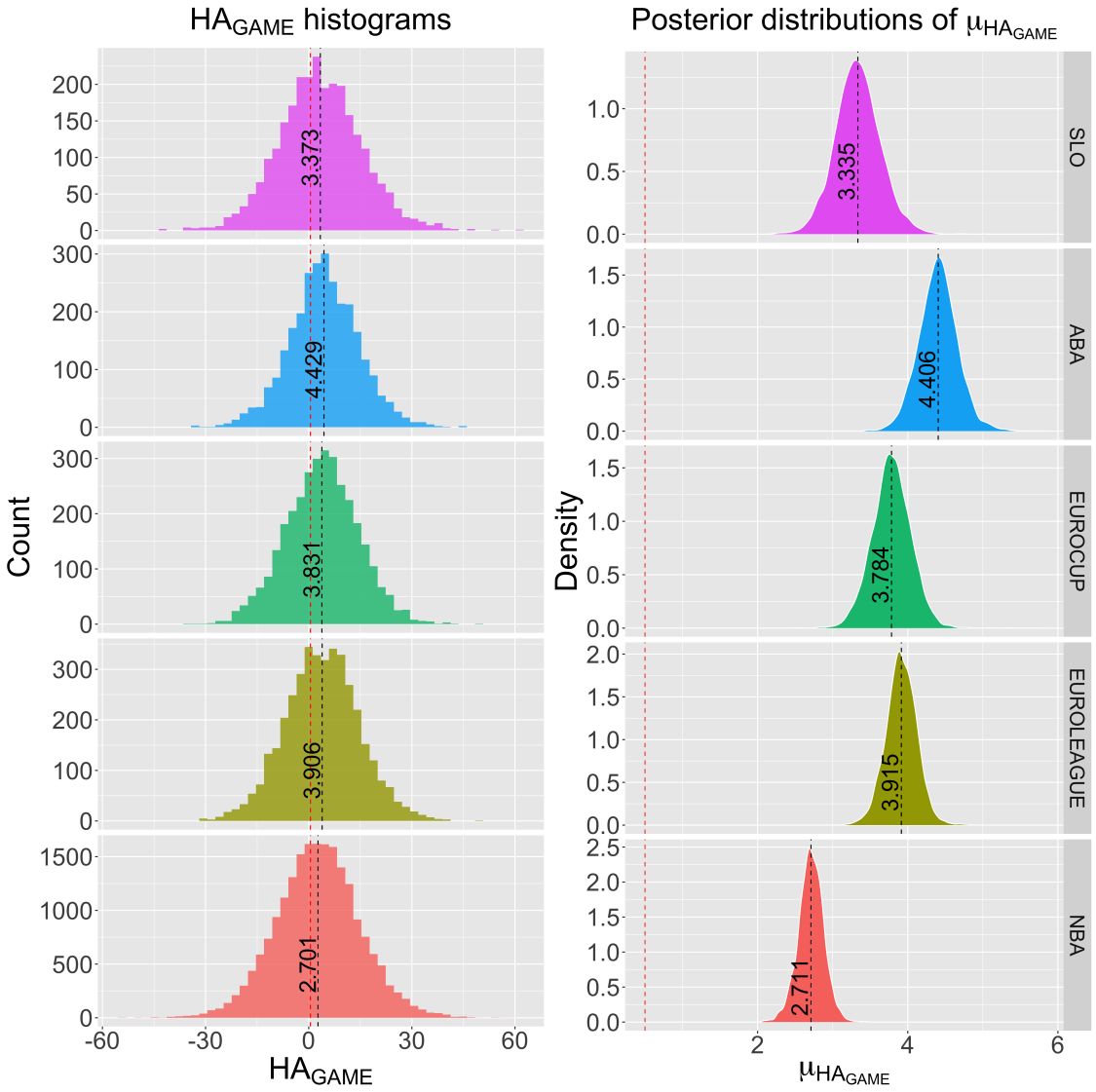

Figure 3.: **Histograms for HA_{GAME} and posterior distributions of $\mu_{HA_{GAME}}$ for basketball leagues.** The black dashed line represents the distribution mean. The red line represents the value around which the distribution would be centered (0) if there was no home advantage. HA_{GAME} is distributed normally with positive means but with frequent negative values. On the right side, we see that it is certain that the mean parameters of the distributions are above 0.

	SLO	ABA	EC	EL	NBA
SLO	-	0.005 ± 0.001	0.119 ± 0.005	0.062 ± 0.004	0.953 ± 0.003
ABA	0.995 ± 0.001	-	0.962 ± 0.003	0.935 ± 0.004	≈ 1
EC	0.881 ± 0.005	0.038 ± 0.003	-	0.352 ± 0.008	0.999 ± 0.001
EL	0.938 ± 0.004	0.065 ± 0.004	0.648 ± 0.008	-	≈ 1
NBA	0.047 ± 0.003	≈ 0	0.001 ± 0.001	≈ 0	-

Table 5.: **Comparison of $\mu_{HA_{GAME}}$ for basketball leagues.** Each cell represents $P(\mu_{HA_{GAME_i}} > \mu_{HA_{GAME_j}})$ $\pm$ MCSE. High values in the NBA column show that all other leagues very likely have higher $\mu_{HA_{GAME}}$, while high values in the ABA row indicate that this league has the most prominent home advantage when concerning this metric.

The fact that La Liga has the highest $\mu_{HA_{GAME}}$ is mostly due to the high values in seasons from 2010/2011 until 2015/2016 as seen in Figure 8. After 2015/2016 La Liga does not have a prominent home advantage

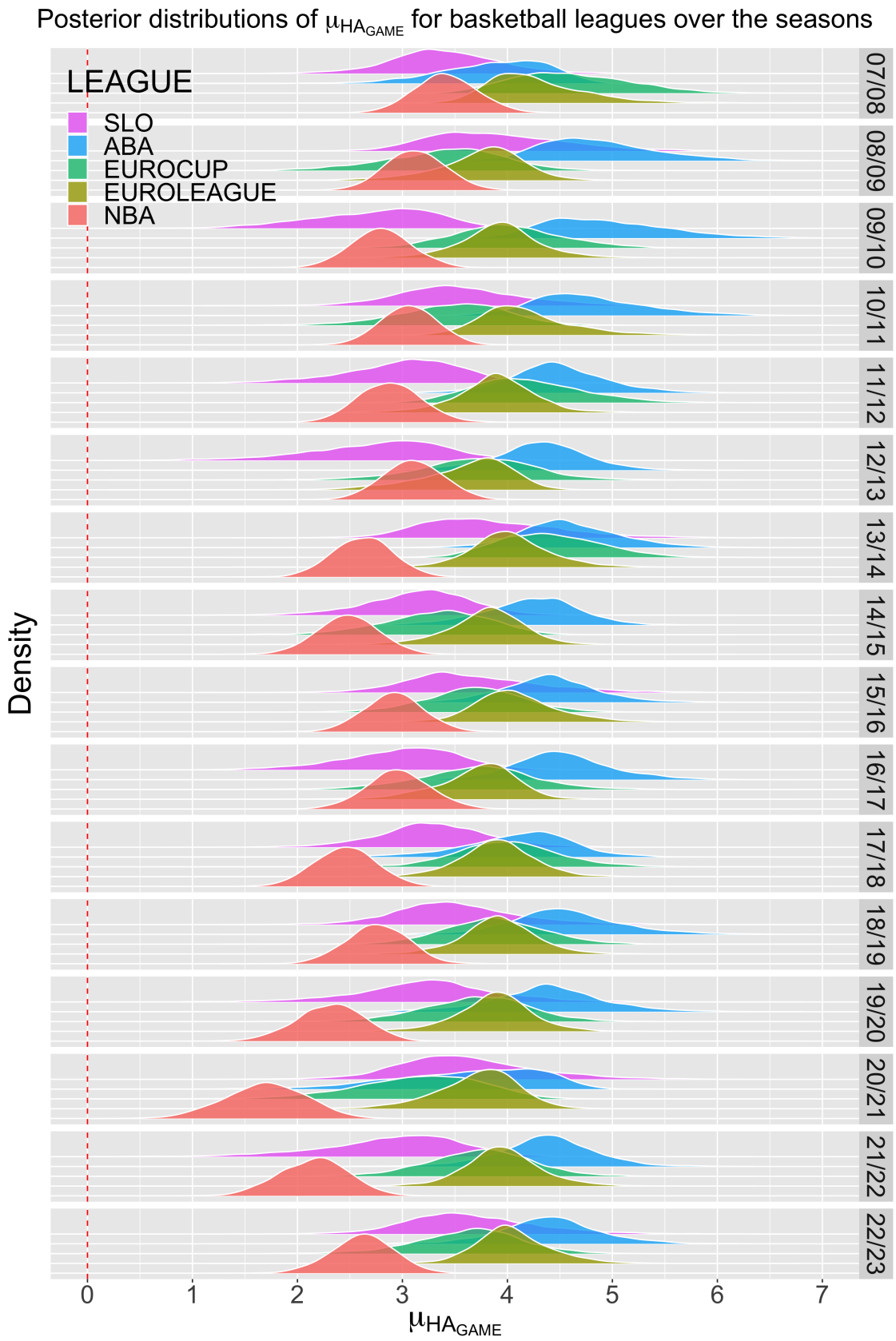

Figure 4.: **Posterior distributions of $\mu_{HA_{GAME}}$ for basketball leagues over the seasons.** Most values of the distributions lie between 2 and 5, with some exceptions, most notably with ABA having the majority of the distribution values over 5 in the seasons 2008/2009, 2009/2010, and 2010/2011. Otherwise, the changes do not seem to correlate with time, so there are no clearly visible trends. The league order mostly stays consistent over the seasons, just like with the $\mu_{HA_{SEASON}}$ metric.

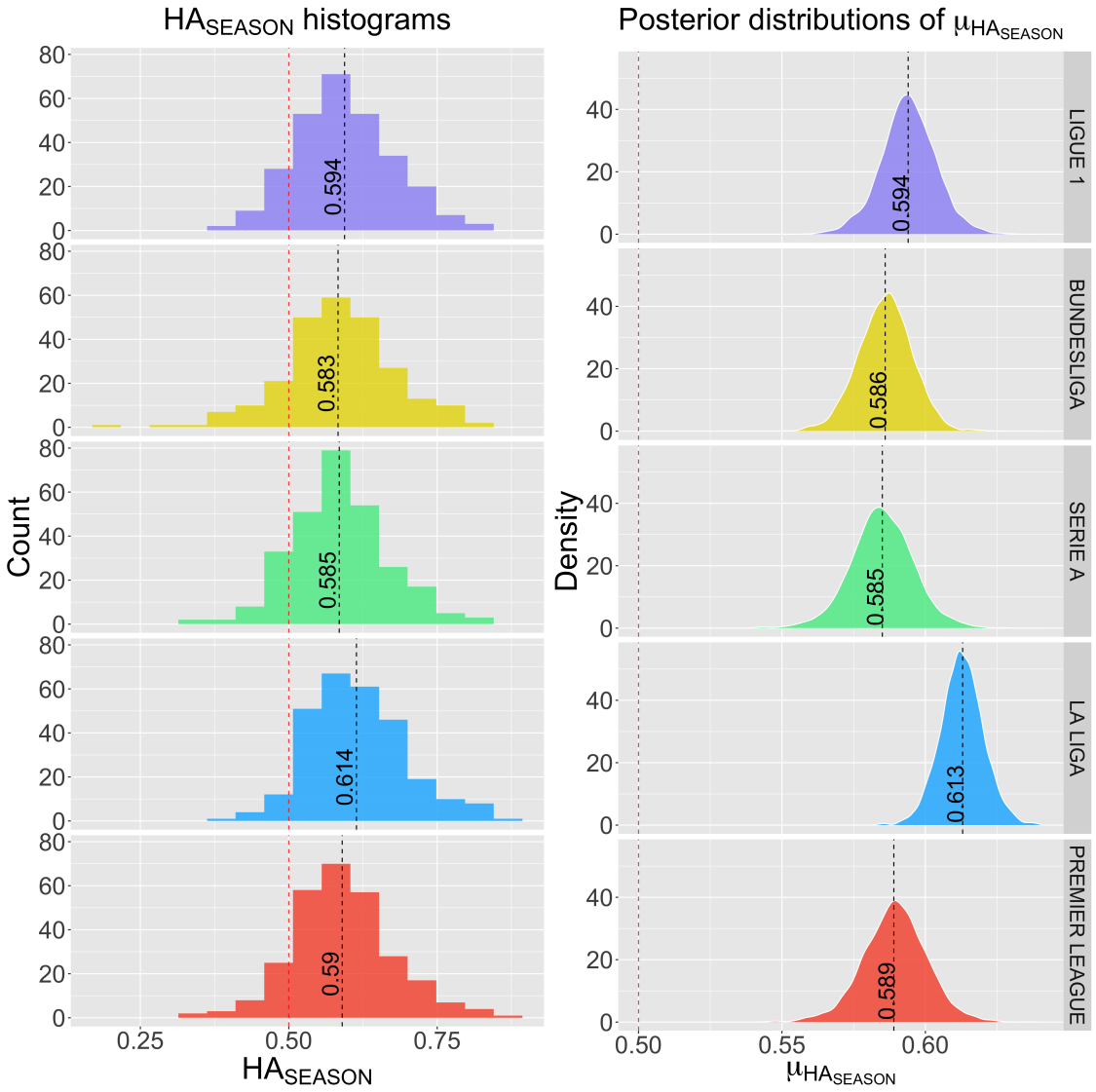

Figure 5.: **Histograms for HA_{SEASON} and posterior distributions of $\mu_{HA_{SEASON}}$ for football leagues.** The left column shows distributions of HA_{SEASON}, which seem normally distributed around a similar mean. In the right column, we show the posterior distributions of $\mu_{HA_{SEASON}}$ parameter, where we observe that La Liga very likely has a higher mean value than the other four leagues.

	ENG	SPA	ITA	GER	FRA
ENG	-	0.039 ± 0.004	0.618 ± 0.008	0.602 ± 0.008	0.375 ± 0.008
SPA	0.961 ± 0.004	-	0.979 ± 0.002	0.985 ± 0.002	0.938 ± 0.004
ITA	0.382 ± 0.008	0.021 ± 0.002	-	0.465 ± 0.008	0.250 ± 0.007
GER	0.398 ± 0.008	0.015 ± 0.002	0.535 ± 0.008	-	0.272 ± 0.008
FRA	0.625 ± 0.008	0.062 ± 0.004	0.750 ± 0.007	0.728 ± 0.008	-

Table 6.: **Comparison of $\mu_{HA_{SEASON}}$ for football leagues.** Each cell represents $P(\mu_{HA_{SEASON_i}} > \mu_{HA_{SEASON_j}})$ ± MCSE. High values in the second row confirm, with a high probability, that La Liga seems to be the league with the highest overall $\mu_{HA_{SEASON}}$.

compared to the other leagues. We stated that the Bundesliga is likely the league with the second most home advantage and from these distributions, we can, in large part, attribute this to three most recent seasons, when the Bundesliga had the highest $\mu_{HA_{GAME}}$ value. Also, we observe a drop in the values in 2020/2021. Not only is the shift in 2020/2021 very clear, but also the probabilities of $\mu_{HA_{GAME}}$ for Ligue 1 and Premier

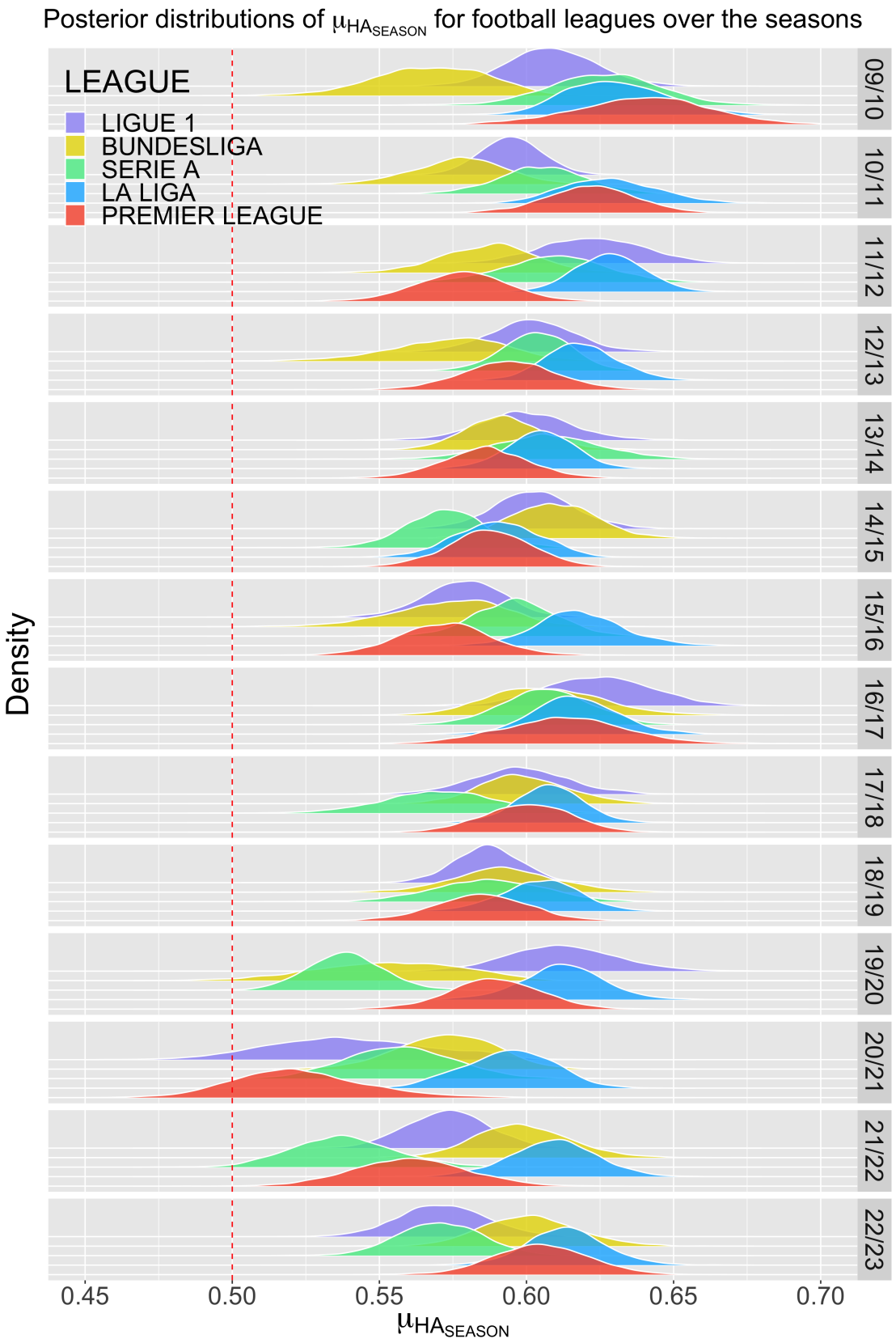

Figure 6.: **Histograms for HA_{SEASON} and posterior distributions of $\mu_{HA_{SEASON}}$ for football leagues over the seasons.** La Liga distributions are consistently distributed over the higher values. We can be sure that $\mu_{HA_{SEASON}}$ values are over 0.5, except for Premier League and Ligue 1 in 2020/2021.

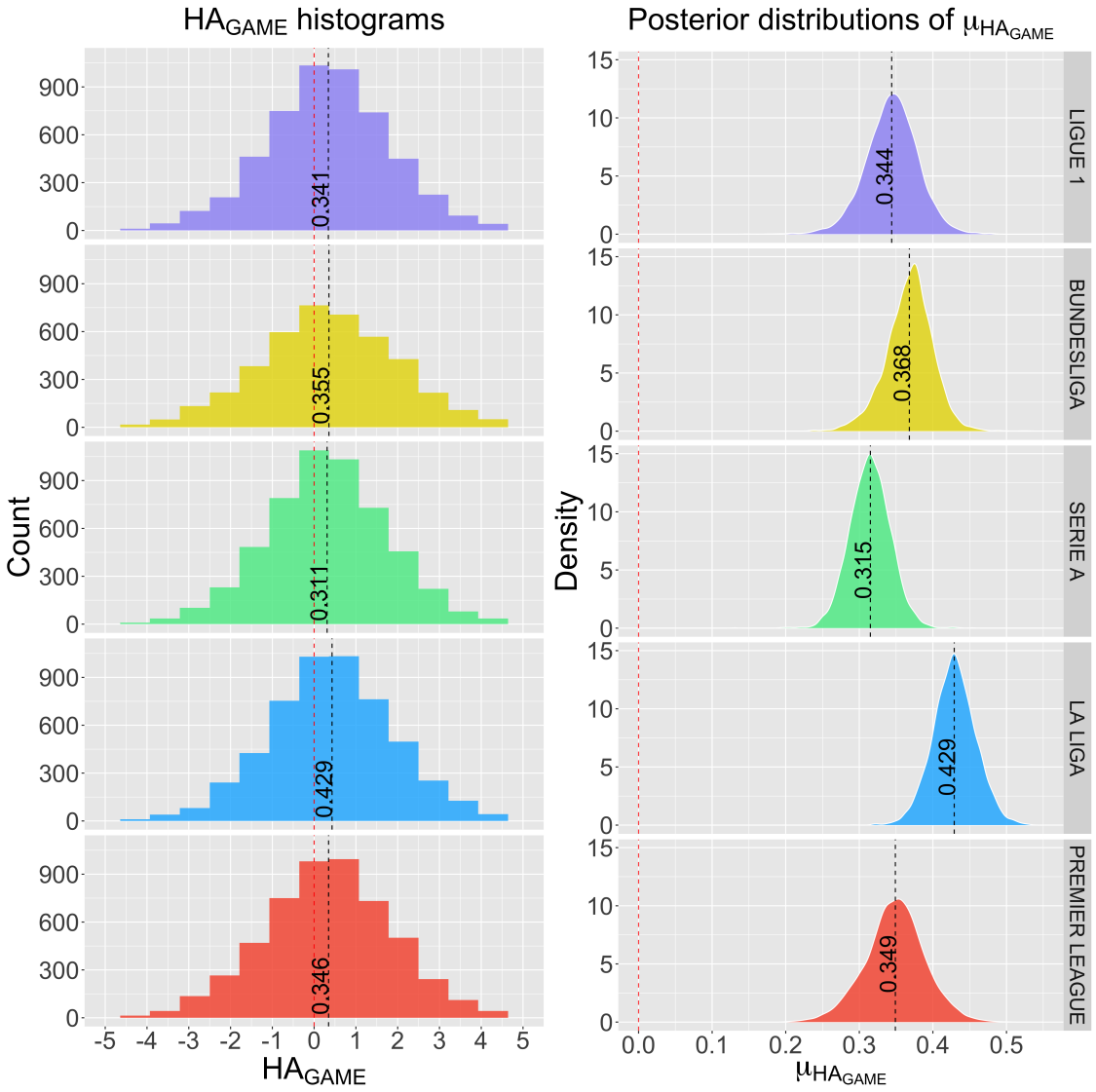

Figure 7.: **Histograms for HA_{GAME} and posterior distributions of $\mu_{HA_{GAME}}$ for football leagues.** La Liga is very likely to have the highest home advantage according to $\mu_{HA_{GAME}}$ as well. Serie A is likely to have the lowest value, however, the confidence in this claim is not very high.

	ENG	SPA	ITA	GER	FRA
ENG	-	0.052 ± 0.004	0.772 ± 0.007	0.341 ± 0.007	0.536 ± 0.007
SPA	0.948 ± 0.004	-	0.997 ± 0.001	0.926 ± 0.004	0.969 ± 0.003
ITA	0.228 ± 0.007	0.004 ± 0.001	-	0.103 ± 0.005	0.243 ± 0.007
GER	0.659 ± 0.007	0.074 ± 0.004	0.897 ± 0.005	-	0.699 ± 0.007
FRA	0.464 ± 0.007	0.031 ± 0.003	0.757 ± 0.007	0.302 ± 0.007	-

Table 7.: **Comparison of $\mu_{HA_{GAME}}$ for football leagues.** Each cell represents $P(\mu_{HA_{GAME_i}} > \mu_{HA_{GAME_j}}) \pm$ MCSE. Spanish La Liga is very likely the league with the highest HA_{GAME} value (second row). There also seems to be an important difference between Bundesliga and Serie A, as $P(\mu_{HA_{GAME_{GER}}} > \mu_{HA_{GAME_{ITA}}}) = 0.897$.

League drop below 1, similar to the drop in $\mu_{HA_{SEASON}}$.

We conclude that Spanish La Liga is the league with the highest home advantage based on the values from both metrics. However, this is mostly due to La Liga's higher home advantage in certain seasons only. We cannot claim that this league consistently has the highest values over the seasons. Furthermore, we see that the differences between basketball leagues are of lower degree compared to the differences between football leagues. The order of football leagues over the seasons

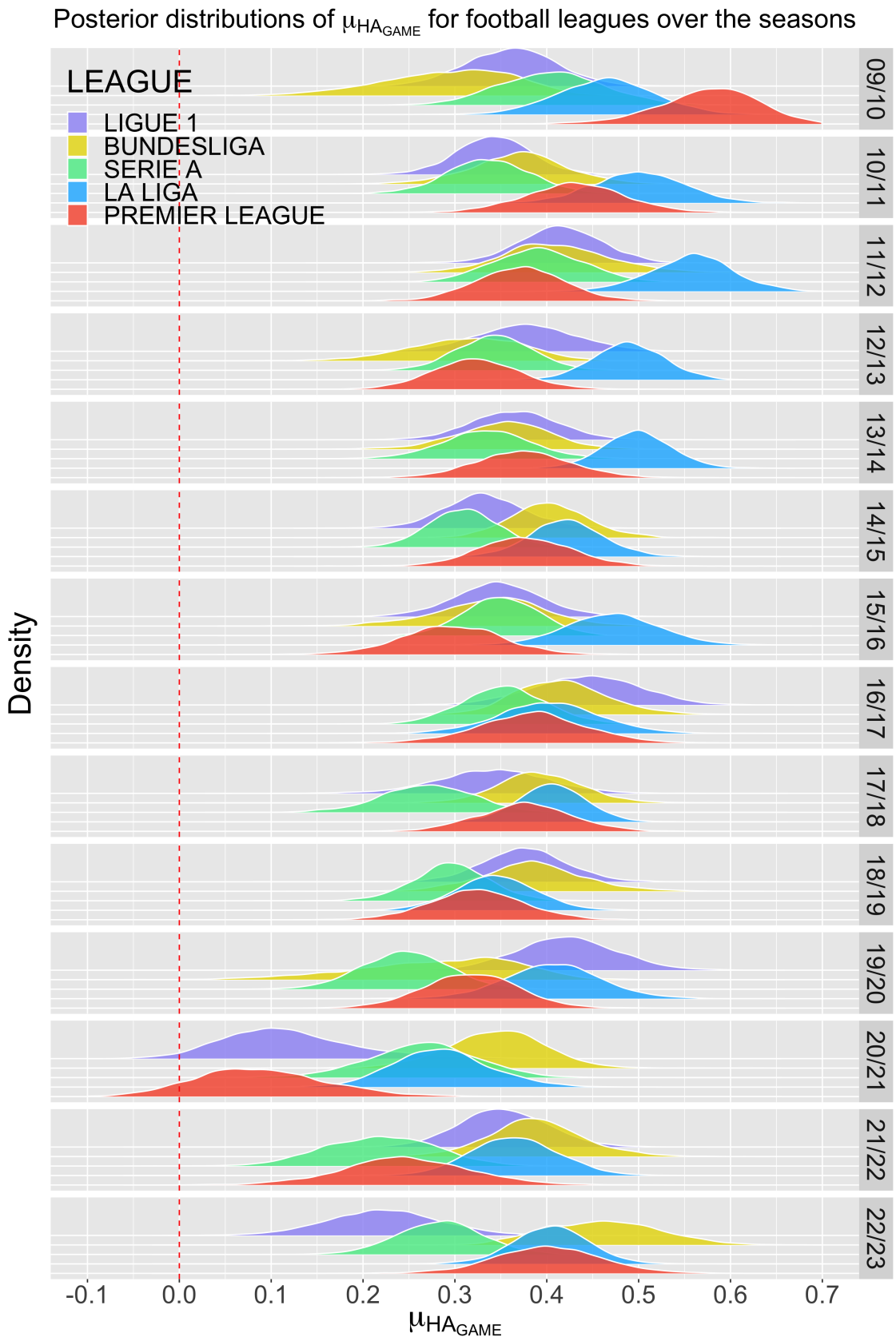

Figure 8.: **Histograms for HA_{GAME} and posterior distributions of $\mu_{HA_{GAME}}$ for football leagues over the seasons.** La Liga likely had the highest $\mu_{HA_{GAME_{season}}}$ from 2010/2011 to 2015/2016, while Serie A very often had the lowest value.

is very inconsistent, whereas in basketball, the order did not change much over time. Again, we can see consistencies between conclusions reached when using HA_{SEASON} and HA_{GAME} which gives validity to our newly introduced metric.

3.3 Correlation between metrics

In sections that follow, we will present the analysis of the home advantage factors, where we used the HA_{GAME} metric. As we stated, this is a newly proposed metric that tries to measure home advantage on a game level, previous sections already showed that conclusions reached through both metrics are consistent. To validate it even further and gain additional insights, we checked its correlation with the widely used HA_{SEASON} metric that measures home advantage on the season level. Since HA_{GAME} and HA_{SEASON} have different scopes of definition, we grouped HA_{GAME} by season and (home) team and calculated the team's average in a single season – $\overline{HA}_{GAME}$. Next, we calculated Pearson's correlation coefficient between HA_{SEASON} and $\overline{HA}_{GAME}$. We obtained the coefficient for each league separately. The results are presented in Table 8. Coefficients for football leagues range from 0.74 to 0.83. Hence, this suggests a strong correlation between the metrics. The coefficients for basketball leagues are a bit lower, but still confidently positive. For four leagues, the values are above 0.55, which still confirms a relatively high correlation. The correlation for the Slovenian league is lower though (0.287).

3.4 Influence of various factors on home advantage in basketball

In Figure 9, we display the distribution of the $RBIAS$ over basketball leagues. In the left column of the figure we observe that the $RBIAS$ values are normally distributed with means close to 0, however posterior distrubutions obtained through our Bayesian analysis (displayed in the right column of the figure) show that we are very confident that the means are above 0. This means that teams seem to be getting more foul calls to their benefit when they are playing at home court, but as we mentioned, we cannot claim that this exclusively happens due to the bias of the referees. Home teams in NBA and Euroleague are getting more than 0.6 additional calls per game. While ABA, Eurocup, and Slovenian League follow with averages of 0.57, 0.44 and 0.2, respectively. The probability that μ_{RBIAS} of Slovenian league is above 0 is still high – $P(\mu_{RBIAS_{SLO}} > 0) = 0.978 \pm 0.003$.

In Figure 10, we visualise the $\mu_{RBIAS_{season}}$ posterior distributions over time to have better insight if there are any changes through the seasons. The order of the leagues matches the order in Figure 9 and stays more or less the same over the seasons. We also notice that $\mu_{RBIAS_{season}}$ values for NBA are somewhat declining over the seasons. A drop in the Euroleague occurred between the season 2020/2021.

Sport	League	Coefficient
Basketball	Slovenian league	0.287
	ABA	0.549
	Eurocup	0.602
	Euroleague	0.552
	NBA	0.677
Football	Premier League	0.807
	La Liga	0.742
	Serie A	0.822
	Bundesliga	0.828
	Ligue 1	0.804

Table 8.: **Pearson's correlation coefficients between HA_{SEASON} and $\overline{HA}_{GAME}$.** Coefficients for football leagues confirm a strong correlation between the metrics. The coefficients for the basketball are somewhat lower. However, the correlation is still clearly there.

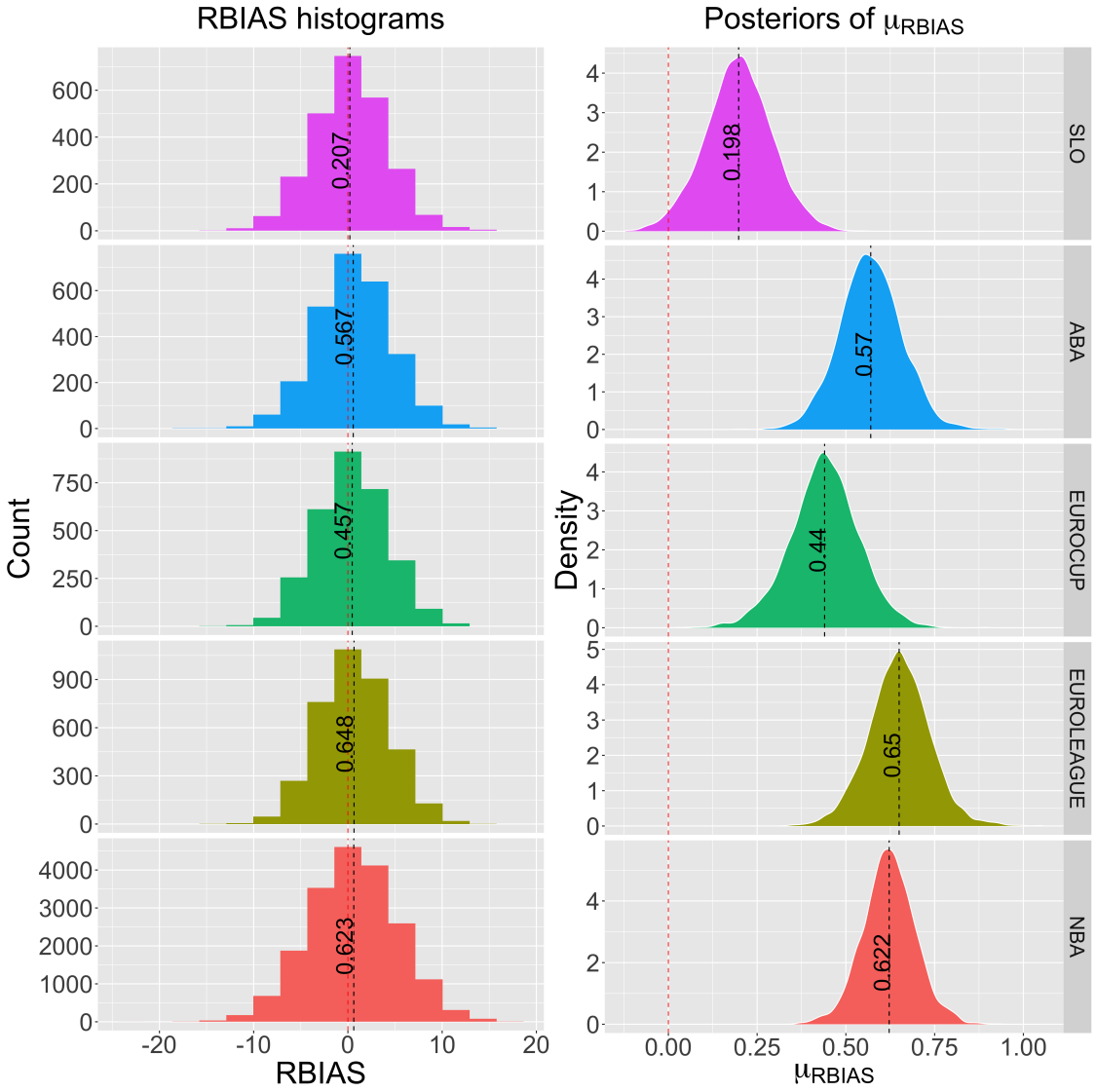

Figure 9.: **Histograms for $RBIAS$ and posterior distributions of μ_{RBIAS} for basketball leagues over the seasons.** The Slovenian league is very likely the one with the lowest μ_{RBIAS}, while Euroleague and NBA seem to have the highest values of referee bias.

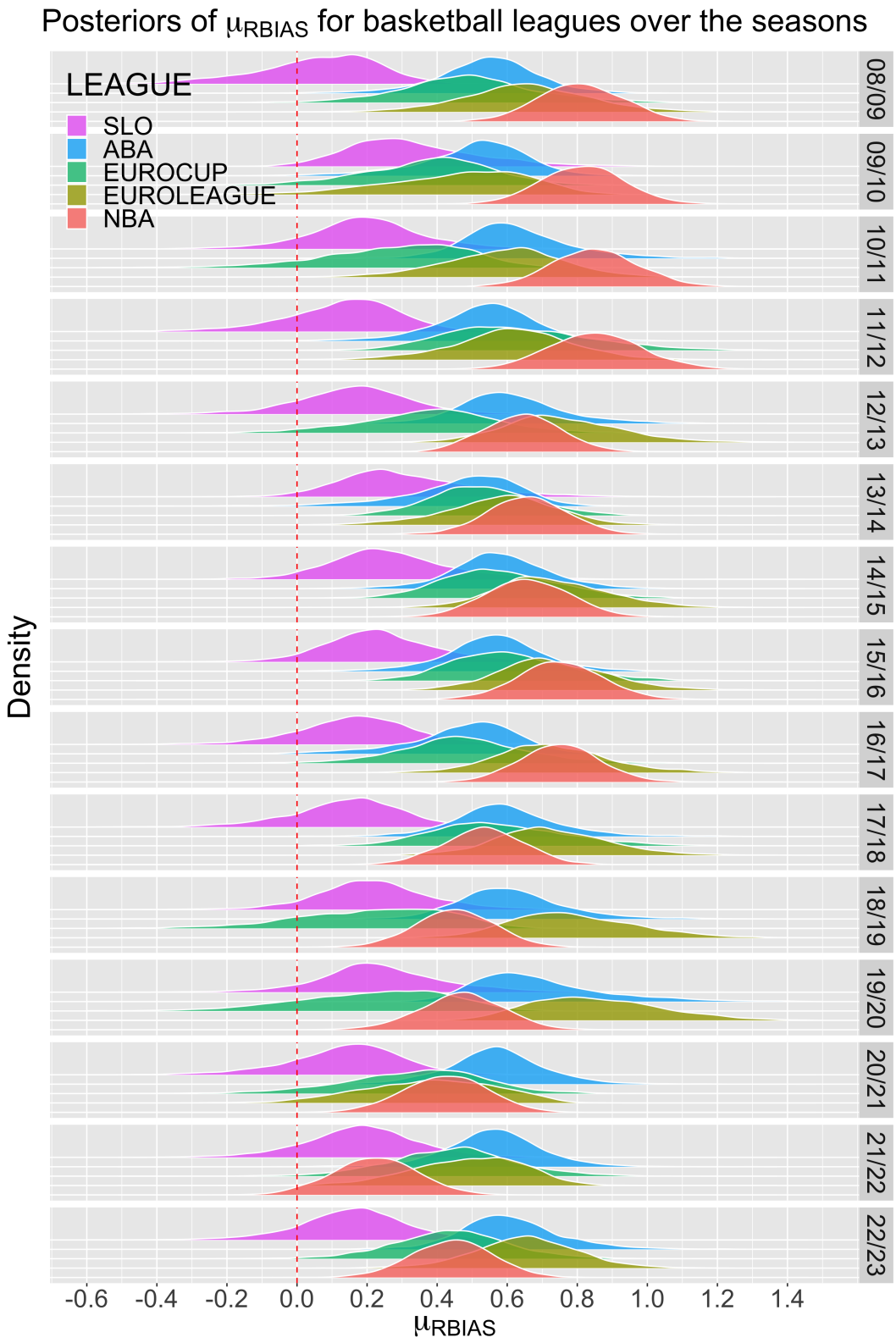

Figure 10.: **Posterior distributions of $\mu_{RBIAS_{season}}$ for basketball leagues over the seasons.** The order of teams by μ_{RBIAS} is consistent over the seasons and matches the order in Figure 9. The Slovenian league always had the lowest value, the Euroleague and NBA most often had the highest values, with NBA values declining in recent seasons, while the values for ABA and Eurocup are somewhere in between.

The purpose of Figure 11 is to have an insight into the distributions of ATT and $DIST$, which help us interpret the linear regression coefficients. In the left column, we observe that NBA games have an average the attendance of 84% of the capacity, while the games in Slovenian league have an attendance below 28% of the capacity.

In the right column, we check if any league has a $DIST$ distribution with a long tail. To some extent, this occurs in ABA because Maccabi Tel Aviv from Israel also competed for one season, while the other venues in ABA are mostly in the region consisting of West Balkan countries, Croatia and Slovenia. We cannot compare this variable between the leagues because it is normalised by the league median. Its purpose is to have a variable on the same scale across the leagues and not to directly compare the distances between the leagues.

We show β and intercept posterior distributions obtained by Model 9 in Figure 12 accompanied by probabilities of β being larger than zero in Table 9. The results differ from league to league. For none of the three variables we can say that the impact is the same in every league. In the Slovenian league, β_{RBIAS} is distributed around 0, therefore $RBIAS$ seems to neither have a negative nor a positive impact on HA_{GAME}. Interestingly, ATT has a negative correlation with HA_{GAME}, with β_{ATT} average of around -3, which means that on a typical Slovenian league game with $ATT = 0.275$ this factor adds -0.825 of HA_{GAME}, i.e. almost an extra point for the away team. This is something that we did not expect and the explanation for this is not straightforward. Considering that β_{ATT} is negative only for this league, the reason might lie within certain specifics of the league. $DIST$ variable seems to have positive correlation with HA_{GAME} in the Slovenian league, but this is not significant – $P(\beta_{DIST} > 0) = 0.787 \pm 0.009$. Due to the negative β_{ATT}, the intercept for the Slovenian league (3.865) is higher than the average of HA_{GAME} for this league (3.335).

The results for the ABA, Eurocup and Euroleague are somewhat similar. Negative values of β_{RBIAS} show that $RBIAS$ seems to have a negative impact – probabilities $P(\beta_{RBIAS}) > 0$ are 0, 0.012 and 0.006, respectively for ABA, the Eurocup and the Euroleague. On the contrary, the probabilities that ATT has a positive effect are very high, especially for ABA and the Eurocup, and less so for the Euroleague – probabilities $P(\beta_{RBIAS} > 0)$ are 0.961, 0.986 and 0.826, respectively for ABA, the Eurocup and the Euroleague. β_{DIST} for ABA is distributed around 0 – $P(\beta_{DIST} > 0 = 0.604)$, while the values are higher for the Eurocup and the Euroleague, hence $DIST$ might have a positive effect but we cannot claim this with very high confidence. We compare the intercepts with the averages of HA_{GAME}. The average of ABA is 4.3, while the intercept is 4.15. Hence, the difference is not remarkable. The negative impact of $RBIAS$ and positive impact of ATT in ABA cancel out, implicating the existence of some other impactful factors that we did not include in our analysis. Similar findings hold for the Eurocup and the Euroleague, with HA_{GAME} averages of 3.784 and 3.915. Their averages of the intercepts are 2.651 and 3.102, respectively. If we compare the averages for the Eurocup, the positive impact of ATT and $DIST$ (and negative impact of $RBIAS$) can be accounted for more than one point of HA_{GAME} on average (1.1), while in the Euroleague, the impact is a bit less than one point on average (0.8).

In the NBA, the impact of $RBIAS$ is low, but it is very likely positive – $P(\beta_{RBIAS} > 0 = 0.964)$. Compared to the other leagues, the value of β_{ATT} is quite high, which indicates a strong correlation between ATT and HA_{GAME} in the NBA. The $DIST$ factor does not seem to be of great importance in the NBA. With $P(\beta_{DIST} > 0) = 0.353$, it does not seem to have either a positive or a negative effect. Comparing the NBA average of HA_{GAME} (2.711) with the intercept (0.864), we infer that the included factors account (ATT mostly) for almost 2 points of HA_{GAME}, which is noticeably more than for the other leagues. Still, there is 0.864 of a point of HA_{GAME} that cannot be explained with included variables and is caused by some other factors.

We saw that for three leagues $P(\beta_{RBIAS} > 0)$ is close to 0. Hence, the results suggest $RBIAS$ has a negative correlation with HA_{GAME}. This is something that we did not expect and should be investigated more in-depth. However, one explanation could be that the referees are helping teams when they are having a bad game with negative HA_{GAME}.

3.5 Influence of various factors on home advantage in football

Before analysing the factors in football, we removed the data for some leagues before 2012/2013 since they were missing the attendance information. In Figure 13, we display the distribution of $RBIAS$, from the histograms, we see that $RBIAS$ is distributed normally around 0. On the right side, we observe the posterior distributions of μ_{RBIAS} obtained by applying the Model 8 on the football data.

We notice that the probability of La Liga having a negative μ_{RBIAS} is very high, while for the other four leagues, we can be confident that μ_{RBIAS} is positive. In these leagues, the referees seem biased in favor of the home team. Bundesliga is likely the league with the highest μ_{RBIAS}, where, on average, the referees seem to call 0.785 additional foul in favor of the home team. The distributions for Ligue 1 and Premier League have higher variance than the other three leagues. Since μ_{RBIAS} was modelled using the hierarchical model with the seasons on the second level, we visualize the μ_{RBIAS} over the seasons in Figure 14.

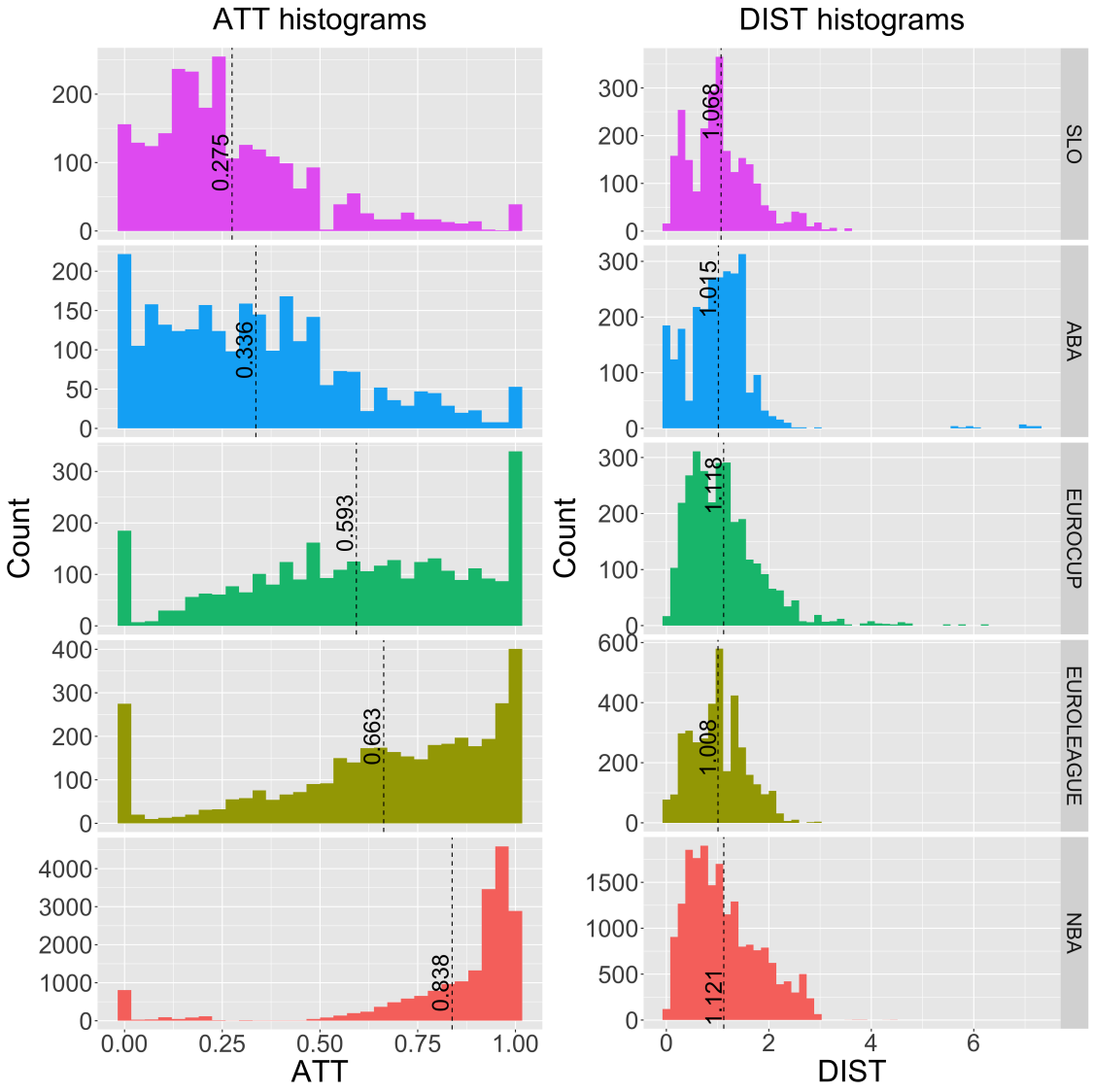

Figure 11.: **Histograms for ATT and $DIST$ for basketball leagues.** NBA seems to have the highest level of arena fullness, while the arenas in the Slovenian league have an average attendance below 28% of the arenas' capacities.

	$P(\beta_{RBIAS} > 0)$	$P(\beta_{ATT} > 0)$	$P(\beta_{DIST} > 0)$	α	$P(\alpha > 0)$
SLO	0.624 ± 0.009	0.005 ± 0.002	0.787 ± 0.009	3.872 ± 0.015	≈ 1
ABA	≈ 0	0.961 ± 0.004	0.604 ± 0.008	4.149 ± 0.01	≈ 1
EC	0.012 ± 0.002	0.986 ± 0.003	0.804 ± 0.008	2.651 ± 0.012	≈ 1
EL	0.006 ± 0.002	0.826 ± 0.008	0.886 ± 0.006	3.102 ± 0.014	≈ 1
NBA	0.964 ± 0.004	≈ 1	0.353 ± 0.009	0.864 ± 0.009	0.995 ± 0.002

Table 9.: **Probabilities of linear regression coefficients being positive for basketball leagues.** We also include the mean and MCSE for α to compare it to $\mu_{HA_{GAME}}$. The results suggest that the probabilities of $RBIAS$ and ATT having positive correlation with HA_{GAME} in NBA are high, while the $RBIAS$ in ABA, Eurocup, and Euroleague is likely to have a negative effect. It is also very likely that ATT has positive correlation with HA_{GAME} in ABA and Eurocup.

Indeed, the distributions over the seasons for Ligue 1 and Premier League are inconsistent, which explains the higher variance of μ_{RBIAS} in Figure 13. Similarly, the low variance in Serie A, La Liga, and Bundesliga results from the high consistency of their seasonal distributions. The distributions for Bundesliga are relatively

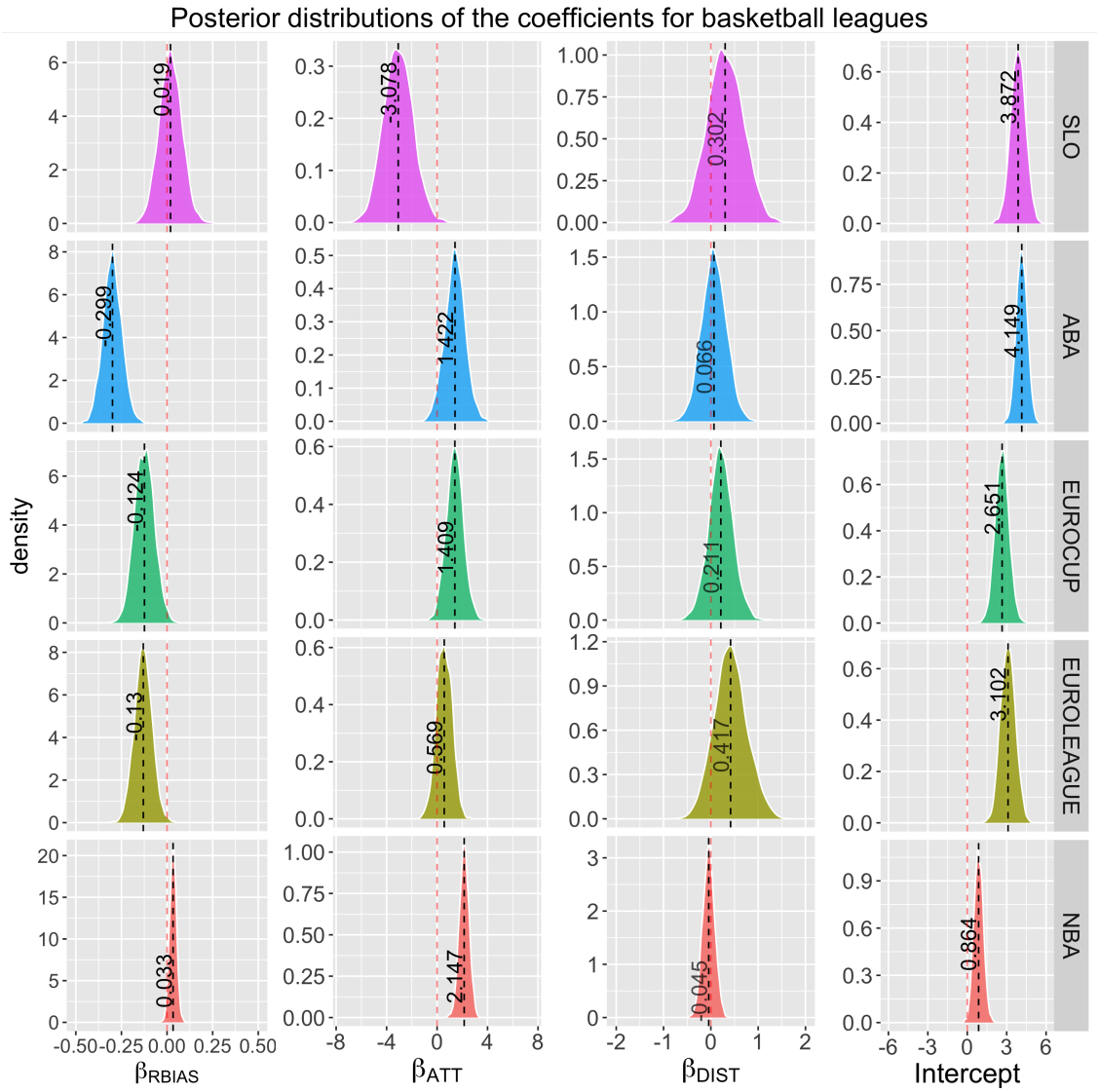

Figure 12.: **Posterior distributions of the β coefficients and the intercept for basketball leagues.** The red dashed line at 0 visualises whether a certain factor has a positive or a negative impact on HA_{GAME}. The results are inconsistent between the leagues, but we can make some general observations. Overall, ATT has a positive correlation with HA_{GAME}, while $RBIAS$ seems to have a negative correlation in three leagues.

consistent, but we observe a change when comparing the first few seasons (μ_{RBIAS} around 1) and the last few seasons (μ_{RBIAS} around 0.5). In 2020/2021, where most of the stadiums had no spectators, Ligue 1 and Premier League (along with the Bundesliga) very likely had negative μ_{RBIAS}, which indicates that spectators might influence referee bias. Overall, there are extra foul calls for the home team on average (with the exception of La Liga). However, this value is lower than 1.

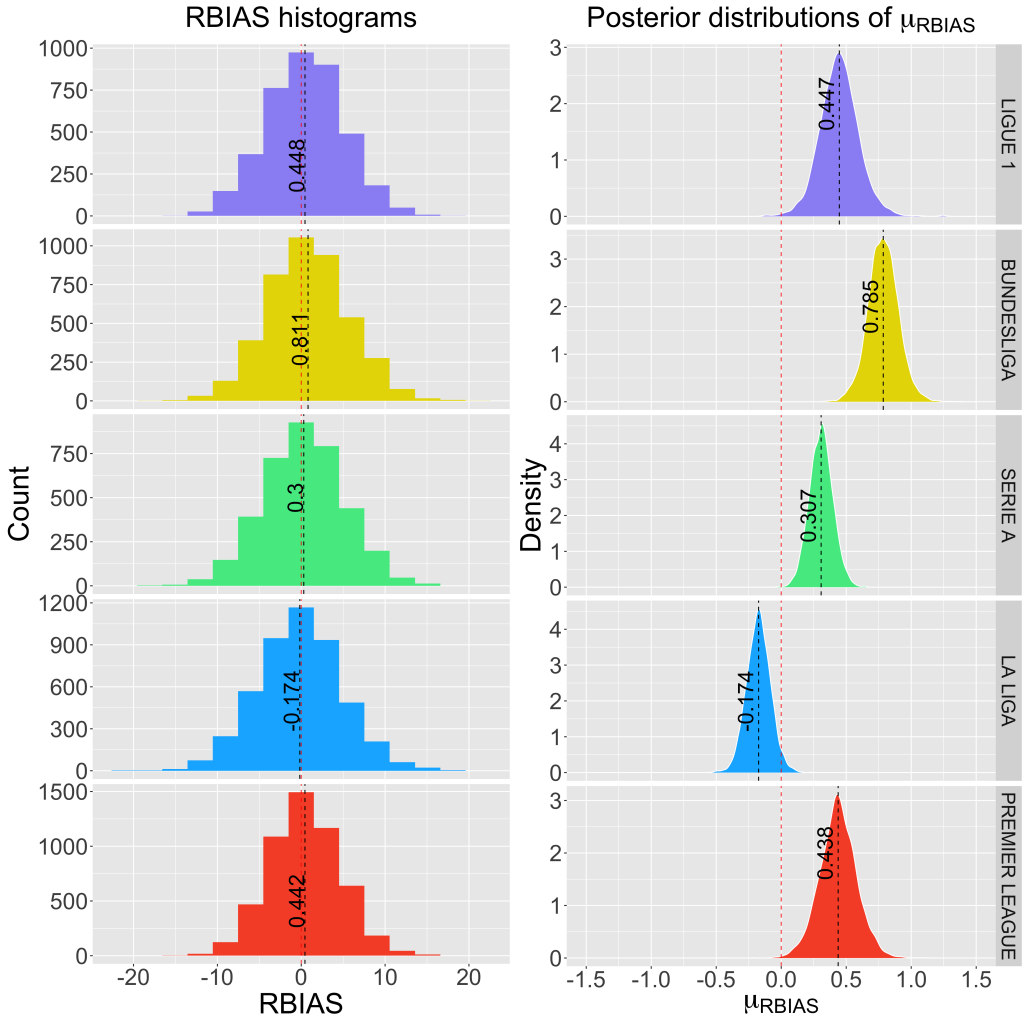

Figure 13.: **Histograms for $RBIAS$ and posterior distributions of μ_{RBIAS} for football leagues over the seasons.** In the histograms, we observe that $RBIAS$ is distributed approximately around 0. From the posterior distributions we see that μ_{RBIAS} is very likely negative for La Liga and positive for the other four leagues.

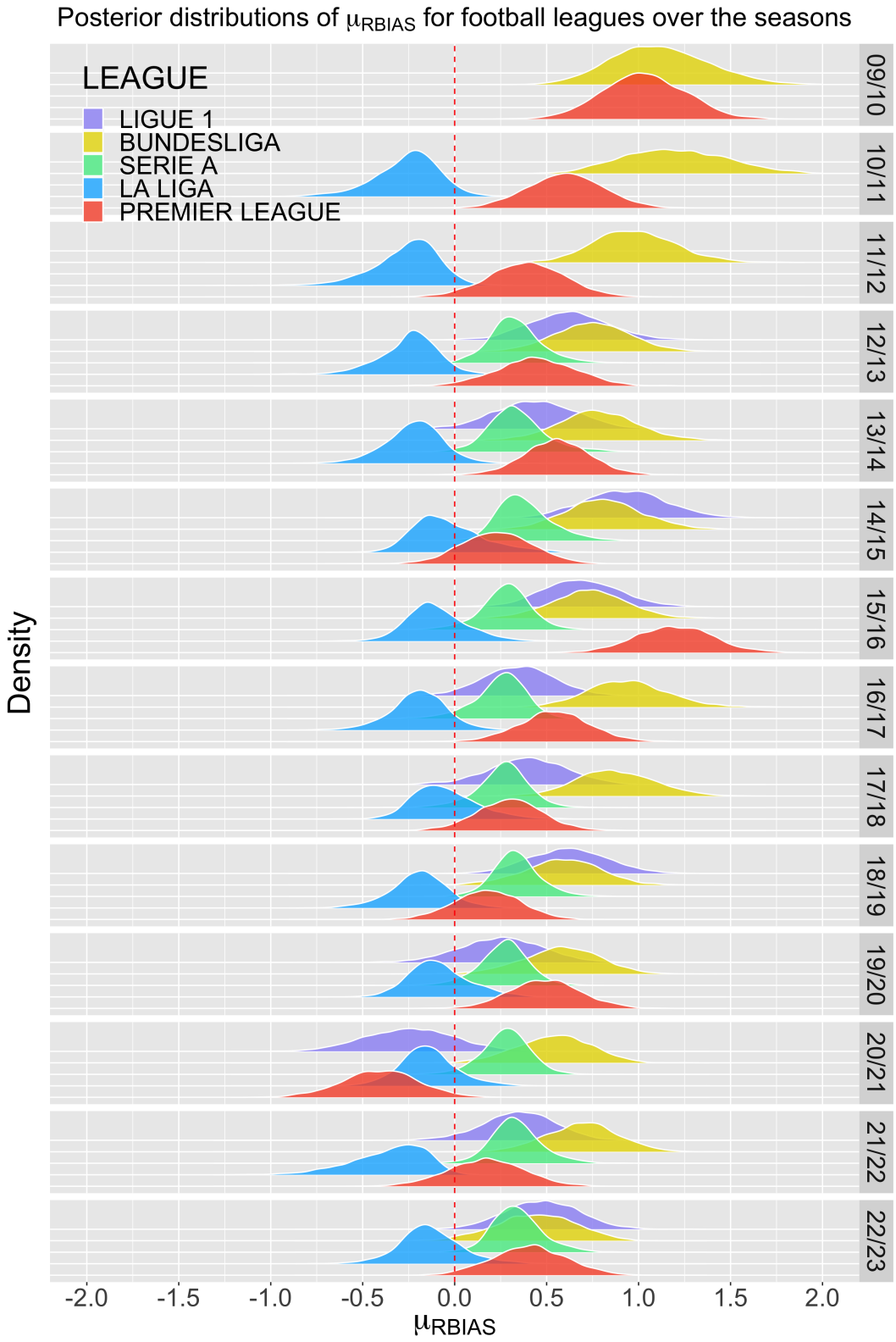

Figure 14.: **Posterior distributions of $\mu_{RBIAS_{season}}$ for football leagues over the seasons.** The distributions for La Liga and Serie A are quite consistent over the seasons, while those for Ligue 1 and Premier League are less consistent. In the 2020/2021 (the season with the lowest number of spectators) we see that three leagues were highly likely to have negative μ_{RBIAS}. Data for some leagues is missing in the first seasons.

Like in basketball, we show the distributions for ATT and $DIST$ for football leagues in Figure 15. We see that in all the leagues, there were approximately 500 games without attendance. The stadiums are often fully packed ($ATT = 1$) in the Bundesliga and the Premier League. These two leagues also have the highest ATT average. In the $DIST$ histograms, we notice some outliers in Spanish La Liga, which are due to 2 clubs from the Canary Islands (Las Palmas and Tenerife), that are quite far from Mainland Spain.

We apply a Bayesian linear regression Model 9 to the football data. The obtained posterior distributions of the coefficients are displayed in Figure 16, the probabilities of coefficients being larger than 0 are presented in Table 10. Similar to basketball, the distributions for β_{RBIAS} are heavily on the negative side, which means it is quite probable that $RBIAS$ has a negative effect on HA_{GAME}. This holds for all the leagues, but it is the most prominent in La Liga, which also has the lowest average of μ_{RBIAS}. Once again, a negative influence of $RBIAS$ is something we did not expect and should be investigated in the future.

We can be quite confident that β_{ATT} is positive and that ATT has a positive correlation with HA_{GAME}. In Serie A the probability of β_{ATT} being larger than 0 is the lowest – $P(\beta_{ATT} > 0) = 0.862 \pm 0.008$, while in other leagues, this probability is above 0.97.

There is no clear answer when deciding whether $DIST$ has a positive or negative effect on HA_{GAME}. The distributions of β_{DIST} for Bundesliga and La Liga are distributed around 0. Their respective probabilities $P(\beta_{DIST} > 0)$ are 0.345 and 0.314, suggesting that $DIST$ has likely no effect in these two leagues. In Ligue 1 the effect of $DIST$ is very unlikely to be positive – $P(\beta_{DIST} > 0) = 0.081$. In the Premier League and Serie A, $DIST$ seems to have a positive correlation with HA_{GAME} (with probabilities 0.93 and 0.992).

Finally, let's look at the posteriors of the intercepts and check how well the factors predict the HA_{GAME}. The results for the Premier League seem to be the most promising. The mean of intercept distribution is 0.046, suggesting that when setting the factors to 0, the predicted average of HA_{GAME} is 0.046. In Table 10, we also observe that $P(\alpha > 0)$ for Premier League is not significant – 0.722. Hence, only a small amount of home advantage in the Premier League can be attributed to the factors we did not consider in the analysis. The difference between the average of HA_{GAME} and the average of the intercept (0.346 vs. 0.054) is prominent in the Premier League. However, it is less prominent in the other leagues – Ligue 1 (0.344 vs. 0.241), Bundesliga (0.355 vs. 0.248), Serie A (0.311 vs. 0.153), La Liga (0.429 vs. 0.327). In these leagues, the models explain some part of the home advantage with the used factors, but a large part of it happens due to the other factors that were not included in our research.

4 DISCUSSION

In this work, we analysed the home advantage in basketball and football by picking five different professional leagues in each sport. Since we did not find any comprehensive datasets, we first had to collect the required data, which was done by using different web scraping libraries. In the process, we were unsure as to what information would be needed for further analysis. As a result, a large amount of obtained data ended up unused. The cleaned and consistently preprocessed datasets are now published in a public repository. As such, they are available to anyone who wants to use them for their research.

Once we collected all the necessary data, we analysed the level of the home advantage. The analysis was done with hierarchical Bayesian models that helped us to infer the parameters describing the distributions of used home advantage metrics. In both sports, we used two different metrics. The first one is often used in the related work, while the second one was proposed in this work. The newly proposed metric HA_{GAME} aims to quantify the home advantage on the level of a single game, while the other metric can only yield a score based on a group of games. The interpretation of the HA_{GAME} metric is such that it quantifies the number of goals (in football) or points (in basketball) that the home team scored above their expected number.

The results for basketball show that the ABA league has the highest average of $\mu_{HA_{GAME}}$ with 4.4, while the NBA league has the lowest average with 2.7. We also analysed the values of $\mu_{HA_{GAME}}$ over the seasons, where we found out that they are fairly consistent over time with a slight decrease in the 2020/2021 season, which was the season that was highly affected by the COVID-19 outbreak.

The football results show that Serie A has the lowest average $\mu_{HA_{GAME}}$ (0.311), not far behind are Ligue 1, Bundesliga, and Premier League, while La Liga has a significantly larger average of 0.429. Compared to basketball, the football distributions over the seasons are much more overlapping, which suggests that the differences between the leagues are lower. The decrease of $\mu_{HA_{GAME}}$ in the 2020/2021 season is noticeable in football as well.

In the second part of our analysis, we chose three factors (referee bias, crowd impact, and travel fatigue), which we hypothesised might have an impact on the home advantage. Since we could not measure the factors directly, we created proxy variables to serve in their place. For referee bias, we proposed the variable $RBIAS$ that uses the number of foul calls, for crowd impact, we used ATT, which is a number that tells how packed the venue was, and for travel fatigue we

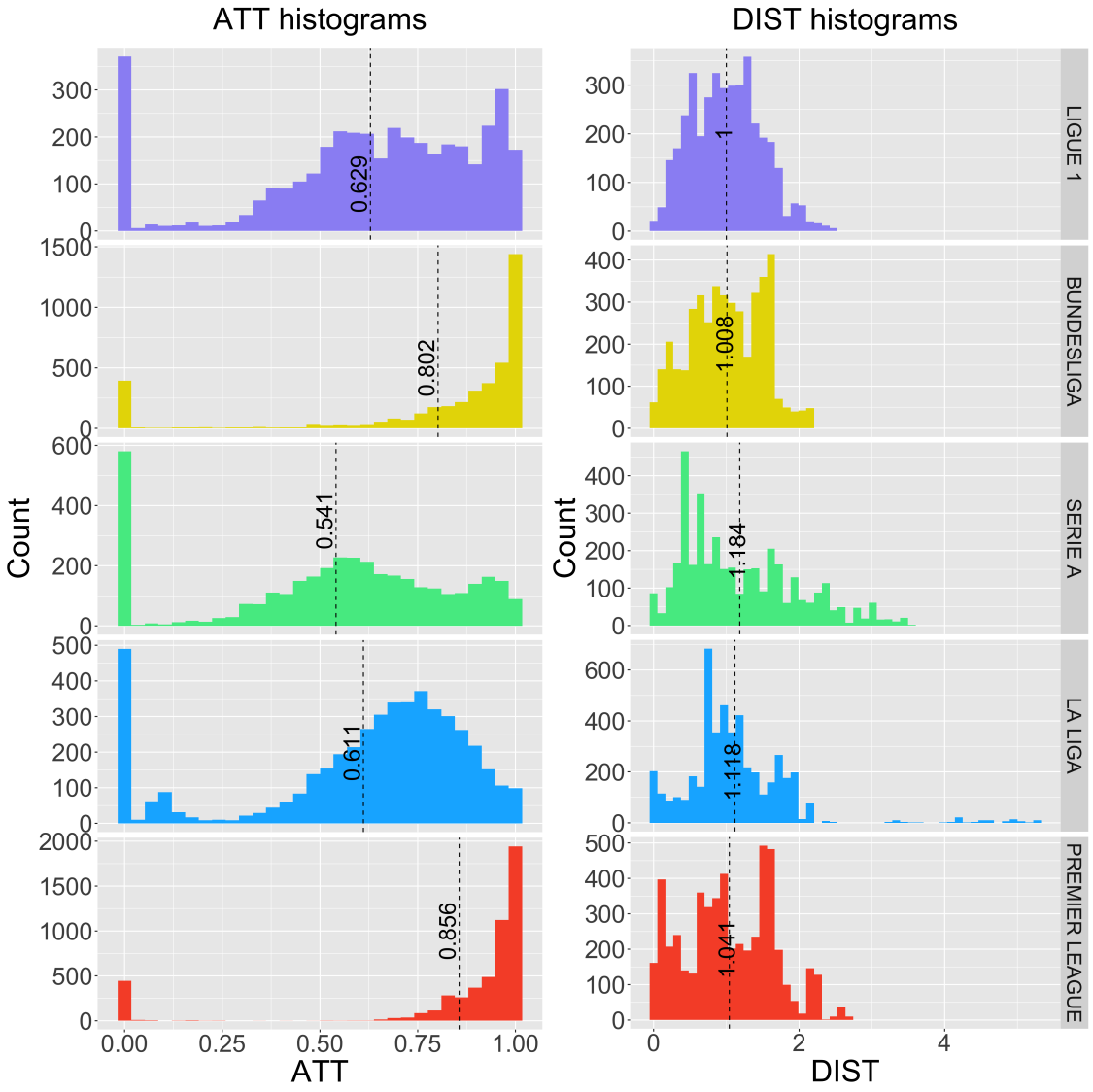

Figure 15.: **Histograms for ATT and $DIST$ for football leagues.** The black dashed lines represent the mean of the distributions. The Bundesliga and Premier League have the highest ATT values. In the right part, we notice the outliers in the $DIST$ histogram for La Liga, which are due to the clubs from the Canary Islands.

	$P(\beta_{RBIAS} > 0)$	$P(\beta_{ATT} > 0)$	$P(\beta_{DIST} > 0)$	α	$P(\alpha > 0)$
FRA	0.086 ± 0.006	0.999 ± 0.001	0.081 ± 0.005	0.241 ± 0.002	0.999 ± 0.001
GER	0.149 ± 0.007	0.972 ± 0.004	0.345 ± 0.01	0.248 ± 0.002	0.998 ± 0.001
ITA	0.018 ± 0.003	0.862 ± 0.008	0.992 ± 0.002	0.153 ± 0.001	0.994 ± 0.002
SPA	≈ 0	0.978 ± 0.003	0.314 ± 0.009	0.327 ± 0.002	≈ 1
ENG	0.01 ± 0.002	≈ 1	0.93 ± 0.005	0.046 ± 0.002	0.722 ± 0.011

Table 10.: **Probabilities that linear regression coefficients are positive for football leagues.** Low probabilities for $P(\beta_{RBIAS} > 0)$ suggest that $RBIAS$ is likely to have a negative effect on HA_{GAME}, while high probabilities for $P(\beta_{ATT} > 0)$ suggest a positive effect of ATT.

used $DIST$, the air distance between the home towns of the teams. To quantify the effect of these variables on HA_{GAME}, we used a Bayesian linear regression.

The results for basketball suggest that ATT has a positive effect on the home advantage in general, whereas for $DIST$ we cannot claim with high probability that it has a positive or a negative effect. In three basketball leagues, $RBIAS$ is likely to have a negative effect on home advantage, which is a bit surprising and motivates us to investigate this in future research. While the

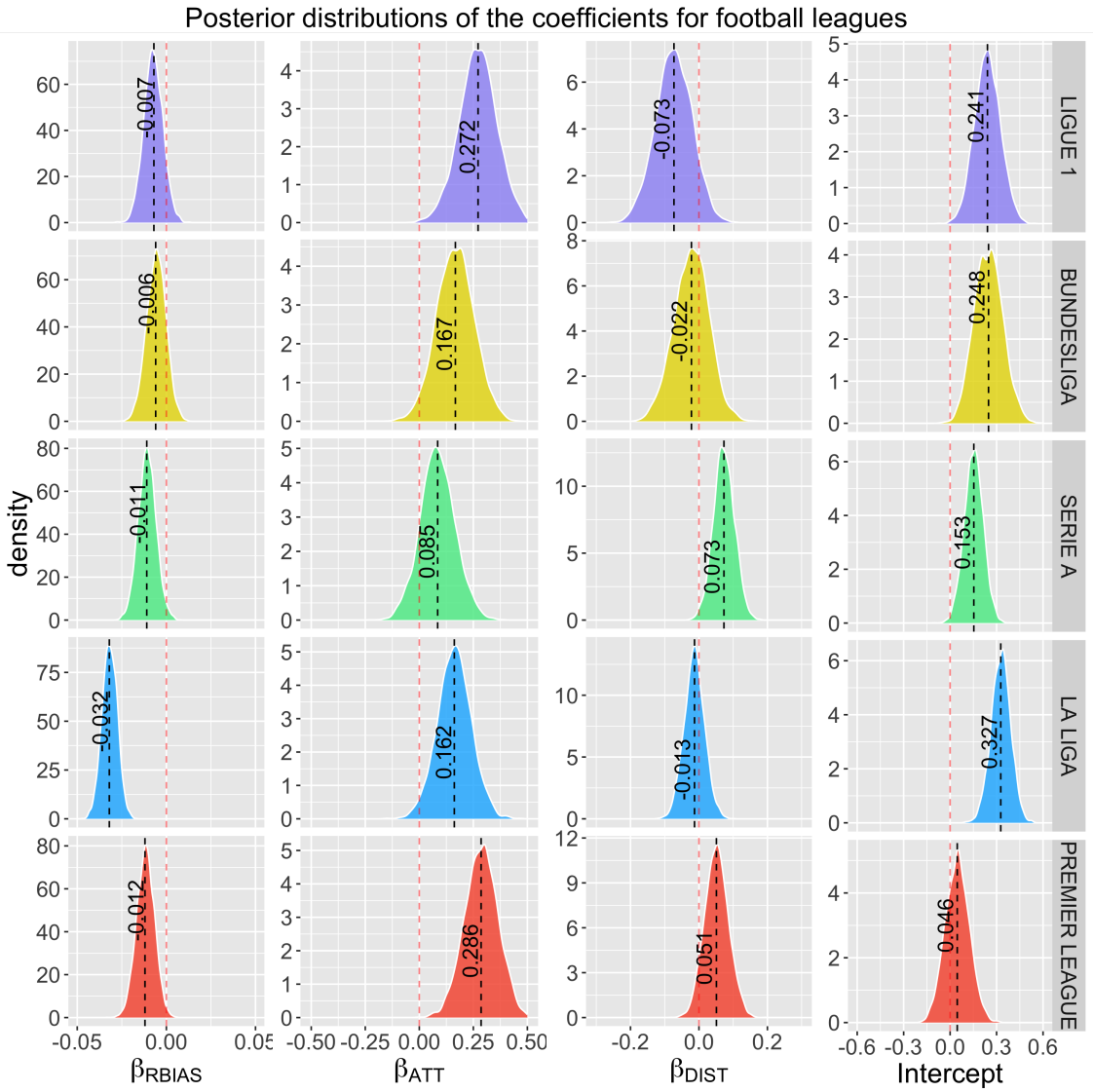

Figure 16.: **Posterior distributions for the β coefficients and the intercept for football leagues** The red dashed line is at 0 to see which variables have a positive or a negative impact on HA_{GAME}. In general $RBIAS$ has somewhat negative impact on HA_{GAME}, ATT has positive impact, while the $DIST$ results are mixed across the leagues.

variables have some correlation with home advantage, there certainly are other impactful factors that were not taken into account.

For football, we reached similar conclusions. We are fairly confident that ATT has a positive correlation with home advantage, whereas the results suggest that $RBIAS$ has a negative correlation. The effects of $DIST$ depend on the league. For the home advantage in the Premier League, the used factors seem fairly vital because they can explain most of HA_{GAME}. This is, however, not the case with the other four leagues, where some part of the home advantage seems to be influenced by other factors.

We found out that the chosen factors had some correlation with the home advantage but also that there exist other factors that seem to have a substaintail impact on the home advantage. Therefore, one thing that could be improved in future work would be to include other potential factors. We observed that $RBIAS$ often had a negative correlation with home advantage. In future work, we should investigate why this occurs and possibly update the methodology of obtaining the variable that serves in place of the referee bias.

FUNDING

This work was partially funded by the Cognitive control beyond executive functions research project (Slovenian Research and Innovation Agency, J5-4590) and the Physiological mechanisms of neurological disorers and diseases research programme (Slovenian Research and Innovation Agency, P3-0338).

REFERENCES

[1] R. Pollard, "Home advantage in football: A current review of an unsolved puzzle," *The open sports sciences journal*, vol. 1, no. 1, 2008.
[2] R. Pollard, "Home advantage in soccer: A retrospective analysis," *Journal of sports sciences*, vol. 4, no. 3, pp. 237–248, 1986.
[3] W. S. Leite, "Home advantage: Comparison between the major european football leagues," *Athens Journal of Sports*, vol. 4, no. 1, pp. 65–74, 2017.
[4] T. Peeters and J. C. van Ours, "Seasonal home advantage in english professional football; 1974–2018," *De Economist*, vol. 169, no. 1, pp. 107–126, 2021.
[5] P. Marek and F. Vávra, "Comparison of home advantage in european football leagues," *Risks*, vol. 8, no. 3, p. 87, 2020.
[6] R. Pollard and V. Armatas, "Factors affecting home advantage in football world cup qualification," *International Journal of Performance Analysis in Sport*, vol. 17, no. 1-2, pp. 121–135, 2017.
[7] N. Van Damme and S. Baert, "Home advantage in european international soccer: Which dimension of distance matters?," *Economics*, vol. 13, no. 1, 2019.
[8] M. Ponzo and V. Scoppa, "Does the home advantage depend on crowd support? evidence from same-stadium derbies," *Journal of Sports Economics*, vol. 19, no. 4, pp. 562–582, 2018.
[9] C. Goumas, "Home advantage and referee bias in european football," *European journal of sport science*, vol. 14, no. sup1, pp. S243–S249, 2014.
[10] R. H. Boyko, A. R. Boyko, and M. G. Boyko, "Referee bias contributes to home advantage in english premiership football," *Journal of sports sciences*, vol. 25, no. 11, pp. 1185–1194, 2007.
[11] C. J. Boudreaux, S. D. Sanders, and B. Walia, "A natural experiment to determine the crowd effect upon home court advantage," *Journal of Sports Economics*, vol. 18, no. 7, pp. 737–749, 2017.
[12] F. Sors, D. Tomé Lourido, V. Parisi, I. Santoro, A. Galmonte, T. Agostini, and M. Murgia, "Pressing crowd noise impairs the ability of anxious basketball referees to discriminate fouls," *Frontiers in psychology*, vol. 10, p. 2380, 2019.
[13] M. A. Gómez and R. Pollard, "Reduced home advantage for basketball teams from capital cities in europe," *european Journal of sport science*, vol. 11, no. 2, pp. 143–148, 2011.
[14] R. Pollard, J. Prieto, and M.-Á. Gómez, "Global differences in home advantage by country, sport and sex," *International Journal of Performance Analysis in Sport*, vol. 17, no. 4, pp. 586–599, 2017.
[15] M. Tilp and S. Thaller, "Covid-19 has turned home-advantage into home-disadvantage in the german soccer bundesliga," *Frontiers in sports and active living*, vol. 2, p. 165, 2020.
[16] M. C. Leitner and F. Richlan, "No fans–no pressure: Referees in professional football during the covid-19 pandemic," *Frontiers in Sports and Active Living*, p. 221, 2021.
[17] F. Sors, M. Grassi, T. Agostini, and M. Murgia, "The sound of silence in association football: Home advantage and referee bias decrease in matches played without spectators," *European journal of sport science*, pp. 1–9, 2020.
[18] K. Fischer and J. Haucap, "Does crowd support drive the home advantage in professional soccer? evidence from german ghost games during the covid-19 pandemic," *Journal of Sports Economics*, 2020.
[19] "Selenium." https://selenium-python.readthedocs.io/index.html, 2023.
[20] "Requests." https://requests.readthedocs.io/en/latest/, 2023.
[21] L. Richardson, "Beautiful soup documentation," *April*, 2007.
[22] "Basketball dataset." https://github.com/timurkulenovic/basketball-dataset, 2023.
[23] "Football dataset." https://github.com/timurkulenovic/football-dataset, 2023.
[24] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan : A probabilistic programming language," *Journal of Statistical Software*, vol. 76, 01 2017.

Jure Demšar received his PhD at the Faculty of Computer and Information Science, University of Ljubljana in 2017. He is currently an assistant professor at the same faculty. He is also partially employed at the Department of Psychology at the University of Ljubljana, where he serves as a senior researcher on neuroscience oriented projects. His research interests lie in neuroscience, Bayesian statistics and machine learning.

Timur Kulenović graduated from the University of Ljubljana, Faculty of Computer and Information Science and is currently a data scientist at Sportradar. His research interests lie in analyses of data in sport.