Research Article

C.J.(Chaojie) Duan*

Latent vs. Observable Home-Field Advantage in Professional Soccer

A Multilevel Bayesian Operationalization

Received 5/21/2018

Abstract: Home Field Advantage (HFA) was traditionally defined in terms of winning percentage of home games at the team level. In this article, we present a hierarchical model of HFA, spanning from the top sport level to the middle league level and all the way to lowest club level. Using scoring performance data from ESPN FC, we fit a Bayesian multilevel nested model to the parameters in the hierarchical model of HFA, allowing information obtained from the season level to inform the inferences about scoring rates at the upper team, league, and sport levels. On the one hand, our analysis reveals that much of HFA is attributed to the nature of the sport of interest. League level source of HFA , on the other hand, can be safely ignored. While only a handful of teams out of 98 in top 5 European leagues enjoy statistically significant HFA, we found absolutely no teams suffer from home disadvantage.

Keywords: European Professional Soccer Leagues, Latent Home Field Advantage, Poisson generative process, Stan

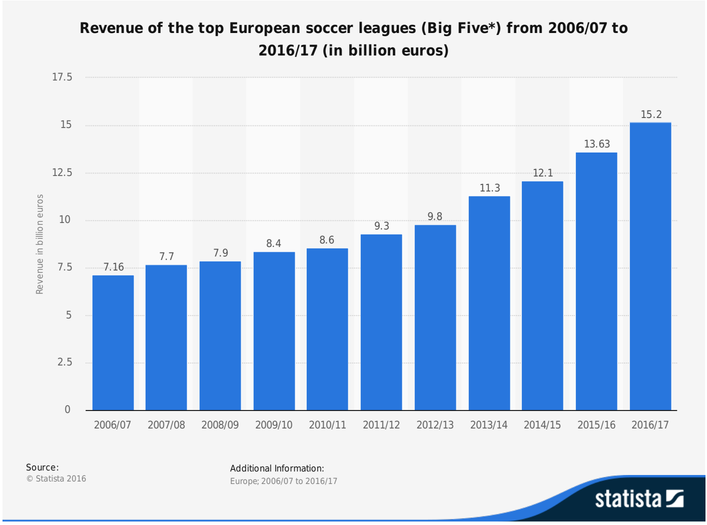
1 Introduction

In professional team sports, the term home field advantage (HFA) – also called home advantage, home ground or home court advantage, defender's advantage, home-ice advantage – describes the benefit that the home team is believed to gain over the visiting opponent. Its scientific definition is "the consistent finding that home teams in sport competition win over 50% of the games played under s balanced home and away schedule" (Courneya and Carron, 1992, p. 13). Due to the existence of HFA, many vital games, such as playoff or elimination matches, in major professional sports have special rules for determining which match is

***Corresponding author: C.J.(Chaojie) Duan,** Dulun Consulting Group, research@dulun.com

played at which place. As shown in Figure 1, the combined revenue of the Big Five European soccer leagues (English Premier League, Spanish La Liga, French Ligue 1, Bundesliga, Italian Serie A) more than doubled to 15 billion euros in 10 years from 2006/07 to 2016/17. The financial implications might partially explain UEFA's (the Union of European Football Associations) decision that a second leg of any Champions League knock-off series is favorable to playing away with the the scores still in balance after the first leg competition (Atkins, 2013).

Fig. 1: *Revenue of the top European soccer leagues (Big Five*) from 2006/07 to 2016/17 (in billion euros)*



Revenue of the top European soccer leagues (Big Five*) from 2006/07 to 2016/17 (in billion euros)

Source:
© Statista 2016

Additional Information:
Europe; 2006/07 to 2016/17

.

The existence of HWP (home winning percentage) -denominated HFA measure has been well documented for a variety of sports, even though the contributing factors are still being debated. In their book *Scorecasting*, Moskowitz and Wertheim (2012) compiled the HWPs in all the major sports with some datasets going back as further as 1903 for MLB and 1966 for NFL. MLS figures date back to only 2002, but show the strongest evidence of HWP of 69.1%. MLB figures, on the other hand, yield the lowest HWP of only 53.9%. This disparity raises an important high-profile question: "Are all sports created equal in terms of HFA?". A subsequent but related question is "Is HFA primarily determined by the sport being played or teams who play the sport?". Answering such questions demands a completely new way of conceptualizing HFA and signals a major departure from the reigning framework proposed by Courneya and Carron (1992), which hinges on game being the unit of analysis.

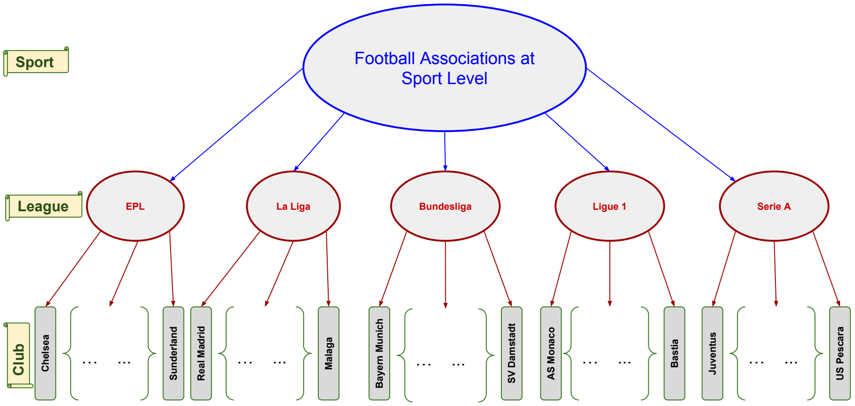
A second motivator for this study is related to the treatment of sports data in general, and scoring in soccer matches in particular. HWP based measures tend to upstage and upgrade the originally discrete count-based outcome to continuous type, while ignoring the underlying data generating process. To complicate matters further, consider the two extreme cases of all winning and losing regular season. The HWP and AWP(away winning percentage) are equal, taking values of either 1.o or 0.0. If we adopt HWP as the sole indicator of HFA, we go straightforward to absurd conclusions - the all winning club enjoys 100% HFA and the zero-win team suffers from 100% home field disadvantage.

The current conceptualization and operationalization of HFA prompt us to take an alternative route in search of true latent HFA underlying the numbers in record books. Specifically, we seek in this paper to achieve the following goals:

1. Propose a fresh new vertical hierarchical model of HFA, complementing the existing horizontal framework.
2. Highlight the different generative process underlying most sports performance metrics and suggest corresponding approaches for analysis.
3. Reveal sources of HFA simultaneously at sport, league, team levels.
4. Presenting a new way of measuring latent HFA via contrast the same performance metric at home and away venues.

The remainder of the paper is organized as follows: In the section immediately after this opening introduction, we review relevant literature and assemble existing knowledge for the development of our unique hierarchical view of HFA. In the section of *Definition of the Hierarchical Model*, we construct a full HFA-specific probabilistic model, which is mainly a joint probability distribution for all observed and latent quantities in a problem, consistent with domain knowledge and the data collection process. In the next section of *Data and Results*, we computing and display the posterior distribution of the unobserved model parameters, given the observed data collected from ESPN FC website. In the Discussions section, we evaluate the fit of the hierarchical model and discuss the implications of the resulting posterior distribution. We conclude our paper with limitations and directions for future HFA research.

Fig. 2: The Hierarchical Structure of Professional Soccer



2 Review of Literature

The UEFA oversees 55 European country-level member associations (such as the English Football Association- EFA), which in turn oversees all professional football leagues within their respective jurisdictions. Figure 2 provides a rough sketch of the organizational structure of professional football in Europe, with an emphasis on the Top 5.

With game as the anchoring unit of analysis, Courneya and Carron (1992) developed a conceptual framework along the timeline axis of a typical professional soccer game. For simplicity of reference and purpose of contrasting, we designate their framework as the horizontal view of HFA (HVHFA). From left to right along the axis, HVHFA incorporates five major components: game site location, game location factors, psychological states, behavioral sates, and the final performance outcomes. At the end of their review, they pointed out that future research should be directed at factors causing HFA rather than the verification of its existence. After taking stock of decades' HFA research findings suffused with equivocality, Carron et al. (2005) surprisingly revised the original HVHFA with the deletion of "officials" and the inclusion of "psychological states". The rationale behind their removal of officiating factors is rather methodological inconvenience. Unlike spectators, players and coaches, referees and umpires can't be easily assigned to either hosting or visiting status for each game they officiated.

Pollard (1986) discovered that the extent of HFA in English soccer has remained relatively consistent since the formation of the English Football League in 1888.

The main substantive goal of this paper is HFA decomposition in a multilevel format.

Frequentist approaches with the rare exception of Gajewski (2006).

3 Definition of the Hierarchical Model

The essence of Bayesian inference is fitting a probability model to a dataset and generating probability distributions on the parameter encapsulated by the model (Gelman et al., 2014).

For our project, the data set contains the season (s) -level best home and away scoring numbers (y^H_{ijs} and y^A_{ijs} respectively) of each club i in each j of the Top 5 leagues. As shown in figure 3, our hierarchical model reflects the organizational structure of professional soccer shown in figure 2. We treat the generative processes of y^H_{ijs} and y^A_{ijs} as similar but independently governed by their own respective parameters. At the measurement level, we encode y^H_{ijs} and y^A_{ijs} into corresponding latent scoring rate λ^H_{ij} and λ^A_{ij} with Poisson distribution, which is a commonly accepted distributional model for sports count data (Miller, 2015):

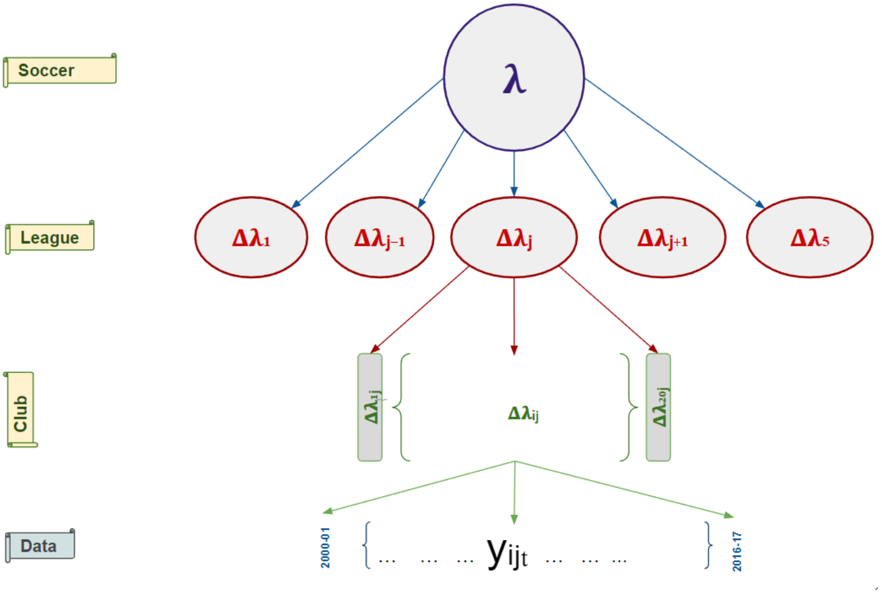
Because team i is nested within league j, we can decompose the latent λ_{ij} into $\Delta_{ij} + \lambda_j$ and thus acquire inference about league level latent scoring rate λ_j. By the same token, we can drill λ_j down into $\Delta_j + \lambda$ and estimate sport level latent scoring rate of λ. In the final step, we take the differentials between the three matched pairs of home-away scoring rate and express the hierarchical model of HFA formally as the set of equations consisting of (1), (2),(3), and (4).

$$\begin{cases} y^H_{ijs} \sim Poisson(\lambda^H_{ij}) \\ y^A_{ijs} \sim Poisson(\lambda^A_{ij}) \end{cases} \tag{1}$$

$$\begin{cases} \delta_{ij} = \lambda^H_{ij} - \lambda^A_{ij} \\ \Delta^H_{ij}, \Delta^A_{ij} \sim N(0, \sigma^2_c) \\ \sigma_c \sim cauchy(0,2) \end{cases} \tag{2}$$

$$\begin{cases} \delta_j = \lambda^H_j - \lambda^A_j \\ \Delta^H_j, \Delta^A_j \sim N(0, \sigma^2_l) \\ \sigma_l \sim cauchy(0,2) \end{cases} \tag{3}$$

Fig. 3: The Hierarchical Model of Home Field Advantage



$$\begin{cases} \delta = \lambda^H - \lambda^A \\ \lambda^H, \lambda^A \sim cauchy(0, 10) \end{cases} \tag{4}$$

4 Data and Results

On ESPN FC website, we find only a pair of venue-delineating (home and away) goal scoring metrics used to characterize a professional soccer team's regular league season in search of possible HFA. Below, we define those statistics using the 2015/16 La Liga season of Real Madrid C.F. as an example.

- Most Home Goals (as y^H_{ijs}) = maximum goals scored in a single match played at home. For the season 2015/2016, Real Madrid's $y^H_{3,1,16}$ is 10. They beat Rayo Vallecano by 10-2 at Santiago Bernabéu Stadium on 12/20/2015.
- Most Away Goals (as y^A_{ijs}) = maximum goals scored in a single away match. For the season 2015/2016, Real Madrid's $y^A_{3,1,16}$ is 6. They defeated Espanyol 6-0 on 9/12/2015 at RCDE stadium.

Table 1 provides the summary statistics of y_{ijs}^H and y_{ijs}^A. Both averages and medians suggests the existence of positive goal differential between home and away scoring metrics.

Tab. 1: Descriptive Statistics

	Mean	Median	Std. Dev.	Min.	Max.	Skewness	Kurtosis
MHG	3.634	4	1.676	0	9	0.246	0.034
MAG	2.884	3	1.676	0	10	0.627	0.786

We fit our model with 4 chains of length 999 (with the first 1/3 for warmup) using the default sampler in Stan, the HMC variant of No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014).

The sport and league level estimates of goal-scoring rate differential are shown in figure 4 as shift from the 0. The outer contour lines show the 99.5% uncertainty intervals, while the shaded area underneath covers the corresponding 95% uncertainty intervals. The light bar in the middle represents the mean.

In Figure 3, we observe universal and absolutely strong (99.5%) manifestation of HFA for the sport of soccer. The goal-scoring differentials are centered around 0.65 goals with comparable lengths of uncertainty intervals. It is also clear that the Top 5 leagues as a whole did not assert much influence on either location or shape of the parameters of interest. The English Premier League was able to slightly push the center close to 0.7 goals, which indicates EPL teams enjoy relatively stronger HFA.

To summarize team level estimates, we use the outer thin line and the inner thick line to represent the 95% and 90% uncertainty intervals respectively. The light bar in the middle still represents the mean as before.

As shown in Figure 5, Real Madrid is the only club in La Liga enjoys strong HFA with the left tip of its 95% uncertainty interval not crossing the dashed line of zero. Another set of 7 teams enjoys marginal HFA with the left tips of their 90% uncertainty interval not crossing the dashed line of zero. It is noteworthy that the worst performer of the current 2016/17 season - Malaga - enjoys almost strong HFA.

For Serie A, Figure 6 paints a different picture. Only a total of 6 out of 20 clubs exhibit marginal HFA. As displyed in Figure 7, the number of teams enjoying marginal HFA improved to 10 out of 20 for French Ligue 1. Still, no teams in Ligue 1 enjoy statistically significant HFA. In Figure 8, Bundesliga demonstrates a similar pattern with only 9 teams possesses marginal HFA.

Fig. 4: HFA Posterior Plot at Sport and League Levels

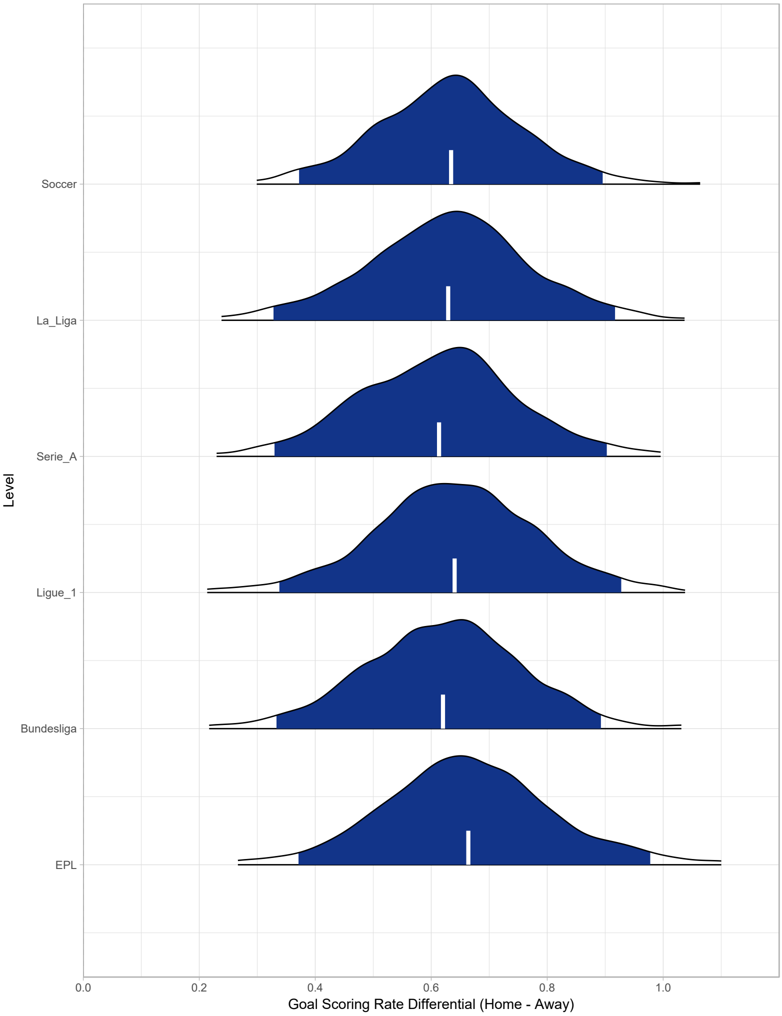


Fig. 5: Home Field Advantage Posterior Plot for La Liga Teams

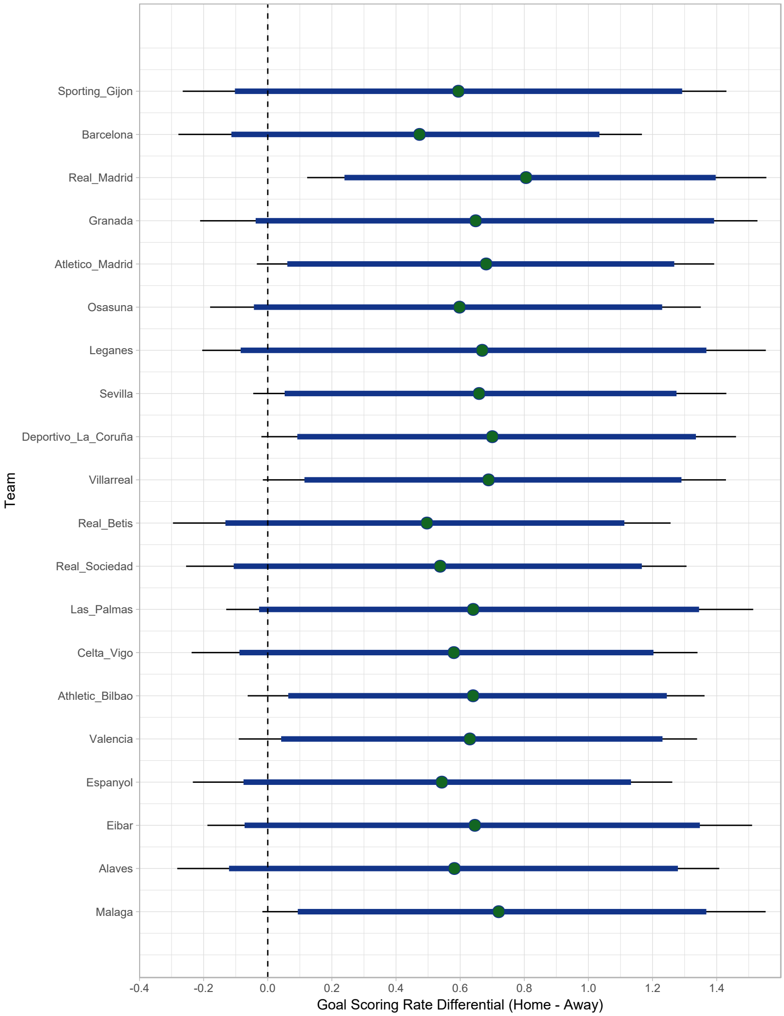


Fig. 6: Home Field Advantage Posterior Plot for Serie A Teams

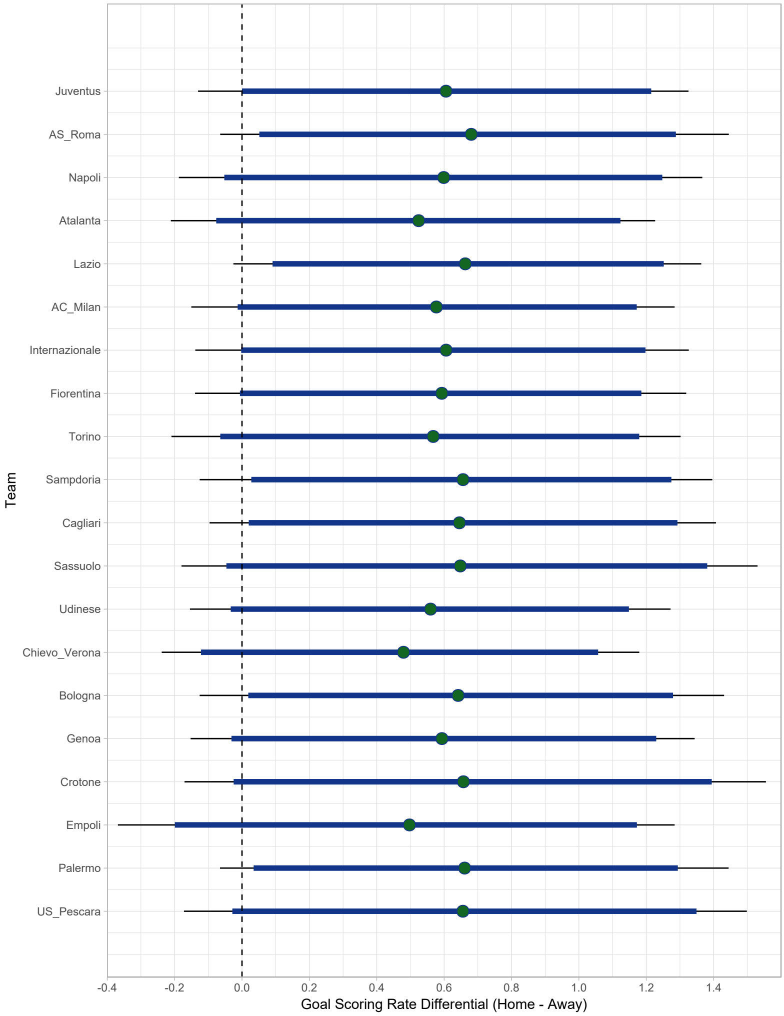


Fig. 7: Home Field Advantage Posterior Plot for Ligue 1 Teams

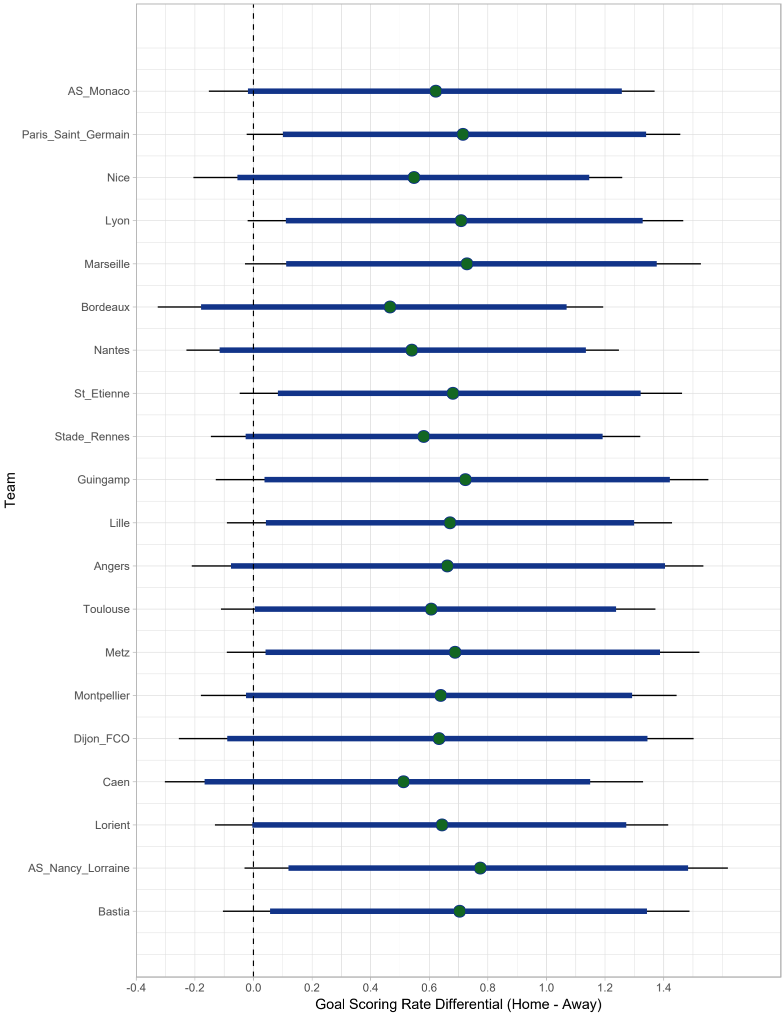


Fig. 8: Home Field Advantage Posterior Plot for Bundesliga Teams

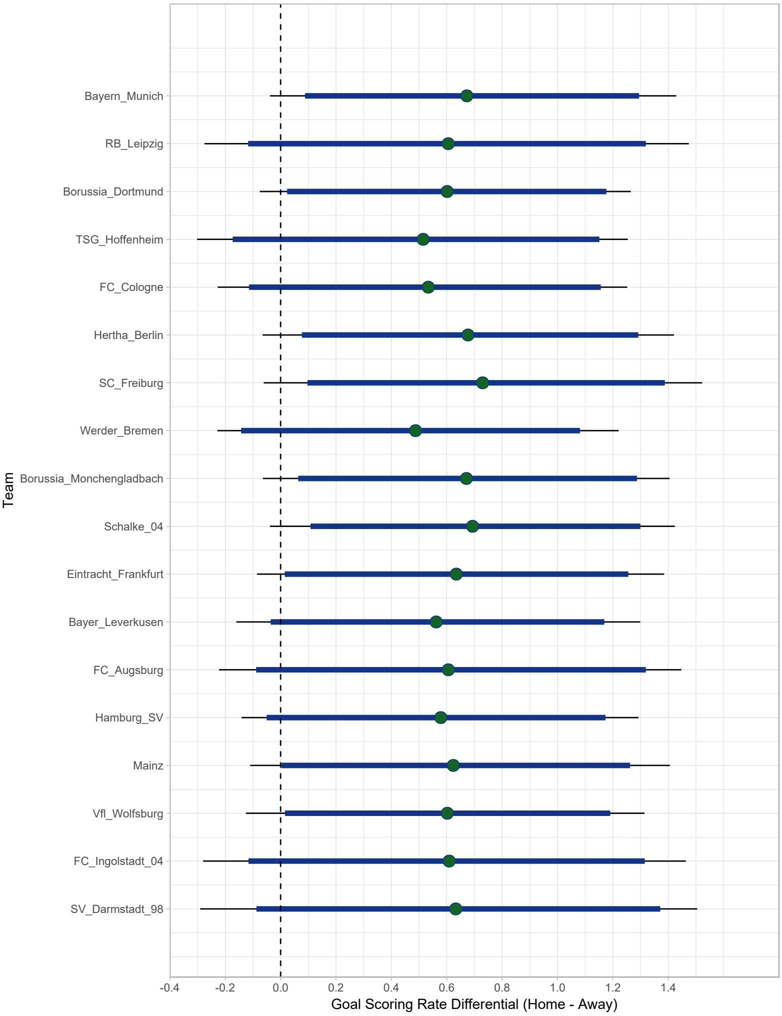
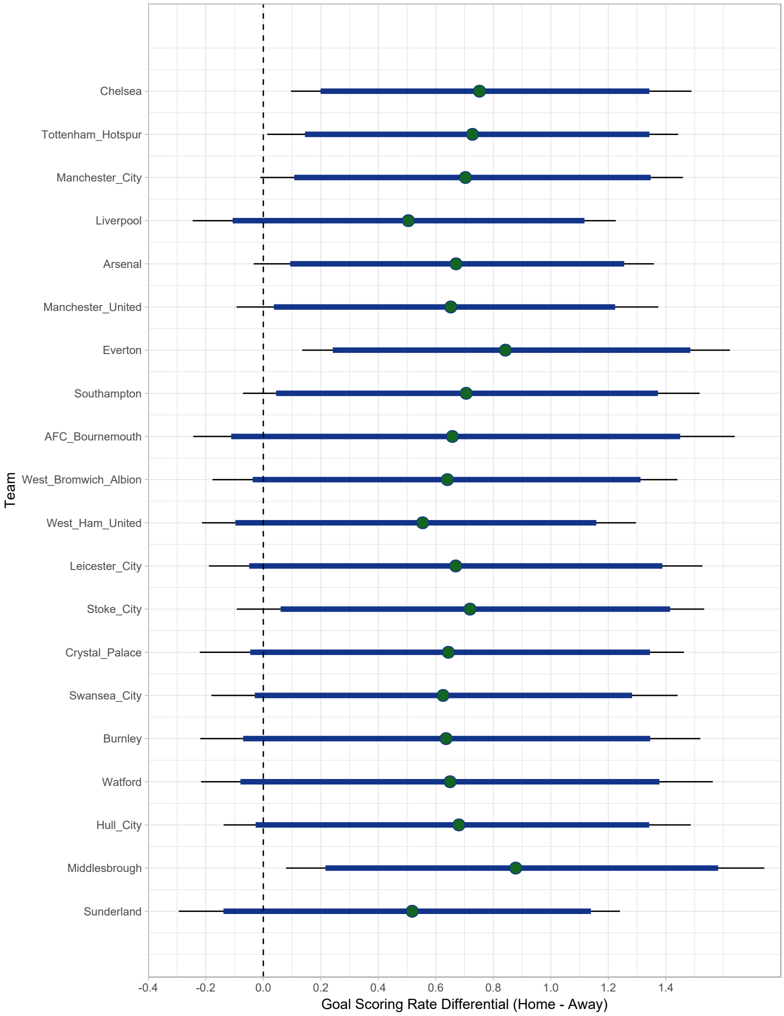


Fig. 9: Home Field Advantage Posterior Plot for English Premier League Teams



Compared to the rest of Top 5 leagues, the English Premiere League shines with 4 out of 20 teams enjoy strong HFA and another 5 teams show signs of marginal HFA. Again, we witness one bottom team enjoys the strongest HFA among all Top 5 clubs. Altogether, we observe with confidence that no teams among the Top 5 leagues suffer home field disadvantage with its corresponding uncertainty intervals lying completely to the right of the neutral line.

As part of the Stan model (Team, 2015), we sample replicated data for the best scoring differential - ydiff - in the *generated quantities* block. We can then check whether the actual score differences are consistent with the distribution of replicated data. For each of the 1122 seasons, we compute the 95% and 50% uncertainty intervals (UI) based on the replicated results. We observe that all of the actual ydiffs are in the 95% UIs and 84.1% in the 50% UIs.

As the last step of model checking, we adopt a one sport-level common parameter for all leagues and teams. After fitting the uni-parameter model to the data, we notice a 3% drop in the 50%-UI containment rate from 84.1% to 81%. In addition, we eliminate the middle layer of leagues from the original model and test the simplified two-layer sport-team model. The reduced model complexity is actually compensated by a minor 0.5% increase in 50%-UI containment rate. The relatively small incremental effect of team-level parametrization seems confirm our earlier observation that only a few teams were able to show statistically strong HFA and leagues show no palpable impact on HFA distribution.

5 Discussion

With the unique hierarchical view and modeling flexibility of Bayesian inferential analysis, we were able to explore the locality of sources of home field advantage. Mirroring the organization structure of professional soccer, we originally proposed a three-level (sport-league-team) model of HFA and tested it with maximum home and away scoring data from 2000/01 to 2016/17. Alternatively, we tested other configurations of the multilevel modeling structure, namely one-level model with sport only and two-level model without the middle layer of leagues.

The home filed advantage exists in all sports with varying degrees. A great deal of future research efforts should be devoted to the inter-sport investigation of HFA.

Acknowledgment: We would like to thank ESPN FC for compiling the season-level club performance data and allow public access.

References

Atkins, C. (2013). How much does home-field advantage matter in soccer? *B/R*.

Carron, A. V., Loughhead, T. M., and Bray, S. R. (2005). The home advantage in sport competitions: Courneya and carron's (1992) conceptual framework a decade later. *Journal of Sports Sciences*, 23(4):395–407.

Courneya, K. S. and Carron, A. V. (1992). The home advantage in sport competitions: A literature review. *Journal of Sport and Exercise Psychology*, 14(1):13–27.

Gajewski, B. J. (2006). There's no place like home: Estimating intra-conference home field advantage in college football using a bayesian piecewise linear model. *Journal of Quantitative Analysis in Sports*, 2(1).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Miller, T. W. (2015). *Sports Analytics and Data Science: Winning the Game with Methods and Models (FT Press Analytics)*. Pearson FT Press.

Moskowitz, T. and Wertheim, L. J. (2012). *Scorecasting: The hidden influences behind how sports are played and games are won*. Three Rivers Press (CA).

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

Team, S. D. (2015). Stan modeling language: User's guide and reference manual. *Version 2.12*.