**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

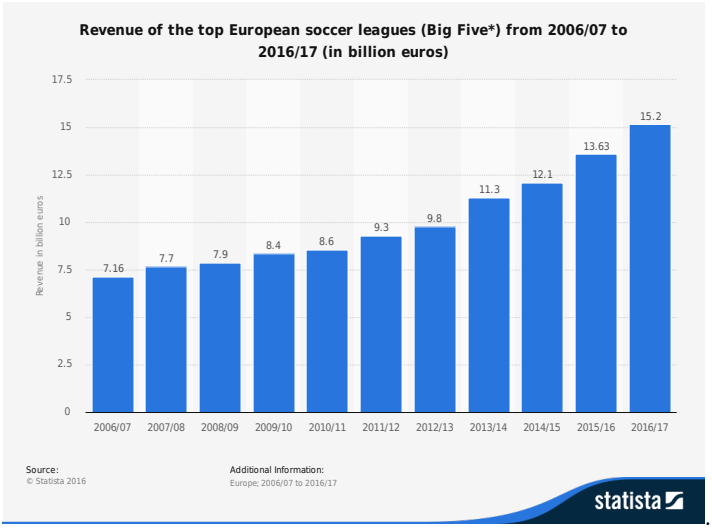
# 1 Introduction

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)*




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

# 2 Review of Literature

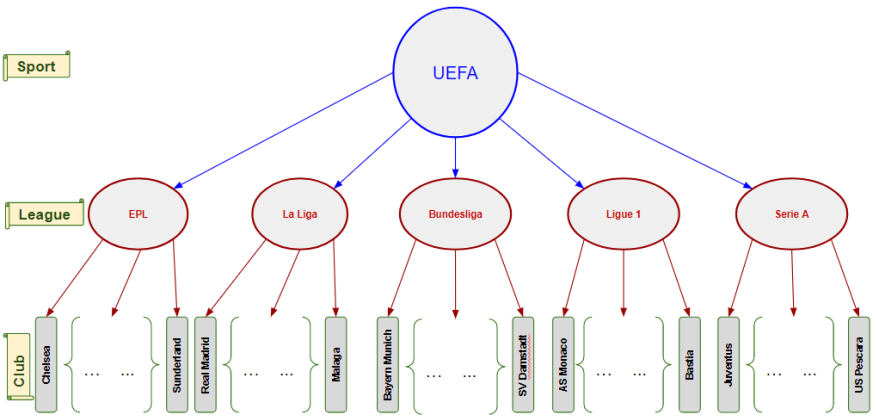
2

# 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). The process of Bayesian data analysis can be ideally divided into three steps:

1. Constructing a full project-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.

**Fig. 2:** The Hierarchical Structure of Professional Soccer



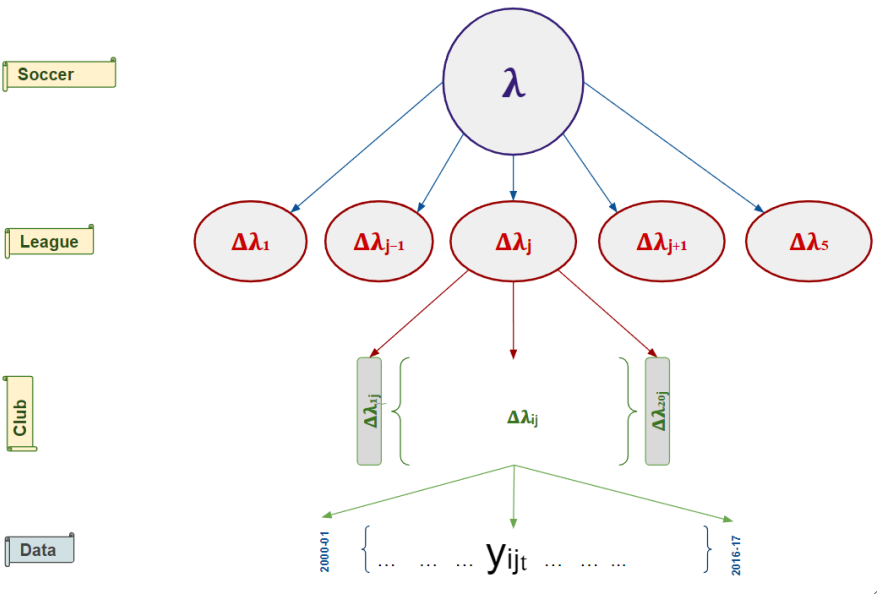

2. Computing and displaying the posterior distribution of the unobserved model parameters, given the observed data.
3. Evaluating the fit of the model and the implications of the resulting posterior distribution.

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 $\lambda^H_{ij}$ and $\lambda^A_{ij}$ 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 $\lambda_{ij}$ into $\Delta_{ij} + \lambda_j$ and thus acquire inference about league level latent scoring rate $\lambda_j$. By the same token, we can drill $\lambda_j$ down into $\Delta_j + \lambda$ and estimate sport level latent scoring rate of $\lambda$. 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).

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

**Fig. 3:** The Hierarchical Model of Home Field Advantage




$$\begin{cases} \delta_{ij} = \lambda_{ij}^H - \lambda_{ij}^A \\ \Delta_{ij}^H, \Delta_{ij}^A \sim N(0, \sigma_c^2) \\ \sigma_c \sim cauchy(0, 2) \end{cases} \tag{2}$$

$$\begin{cases} \delta_j = \lambda_j^H - \lambda_j^A \\ \Delta_j^H, \Delta_j^A \sim N(0, \sigma_l^2) \\ \sigma_l \sim cauchy(0, 2) \end{cases} \tag{3}$$

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

# 4 Data and Results

We run 4 chains using the default sampler in Stan, the HMC variant of No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014) and set

**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 |

The model estimates are shown in figure 4 as shift from the 0. The outer contour lines show the 99.5% credible intervals, while the shaded area underneath covers the corresponding 95% credible interval. The light bar in the middle represents the mean.

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

**Fig. 4:** Home Field Advantage Posterior Plot at Sport and League Levels

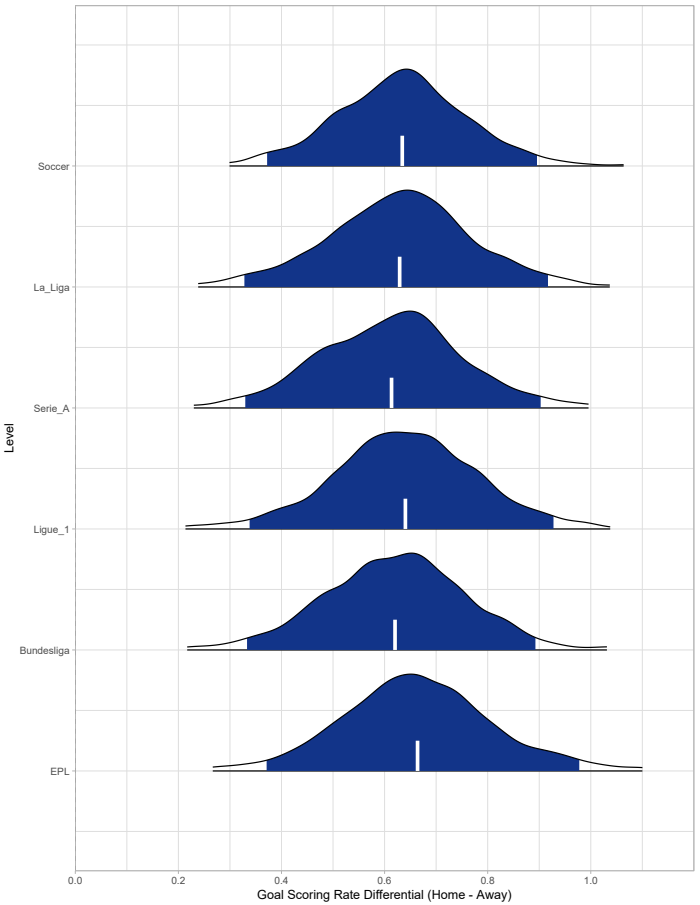


**Fig. 5:** Home Field Advantage Posterior Plot for La Liga Teams

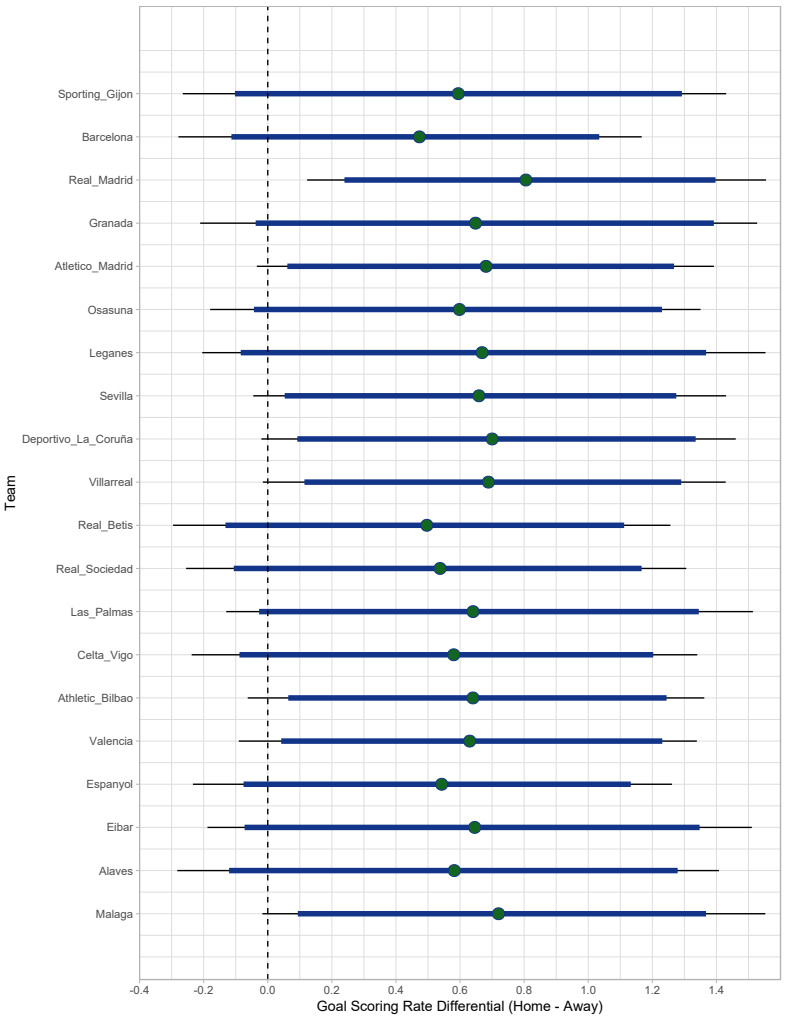


**Fig. 6:** Home Field Advantage Posterior Plot for Serie A Teams

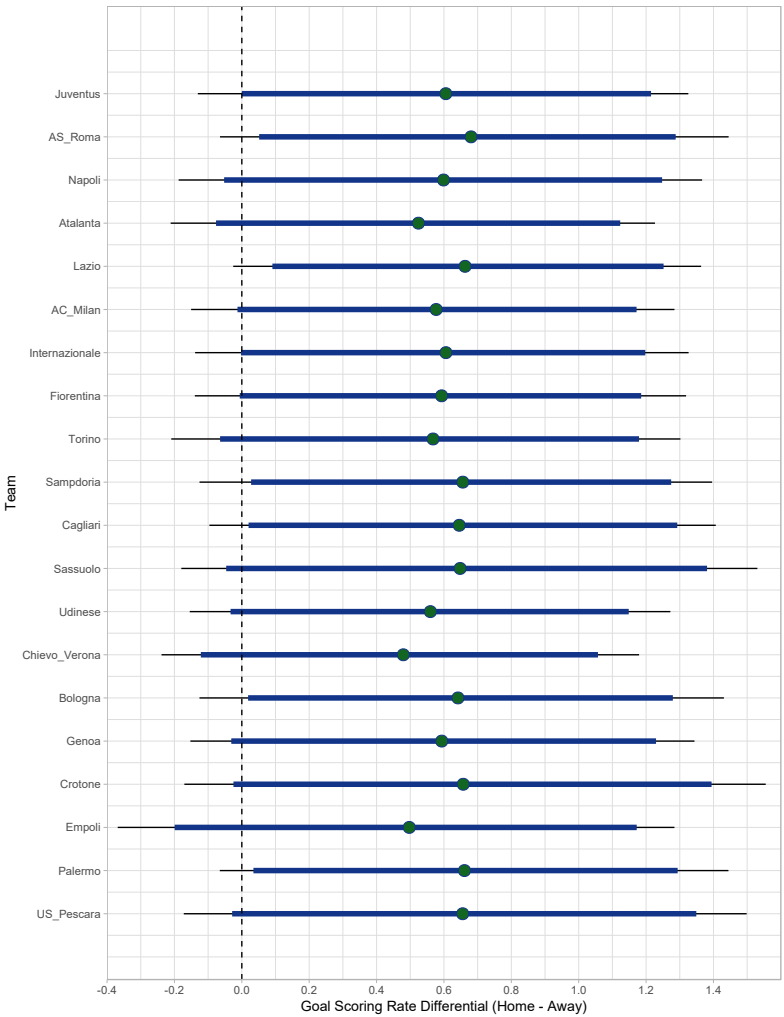


**Fig. 7:** Home Field Advantage Posterior Plot for Ligue 1 Teams

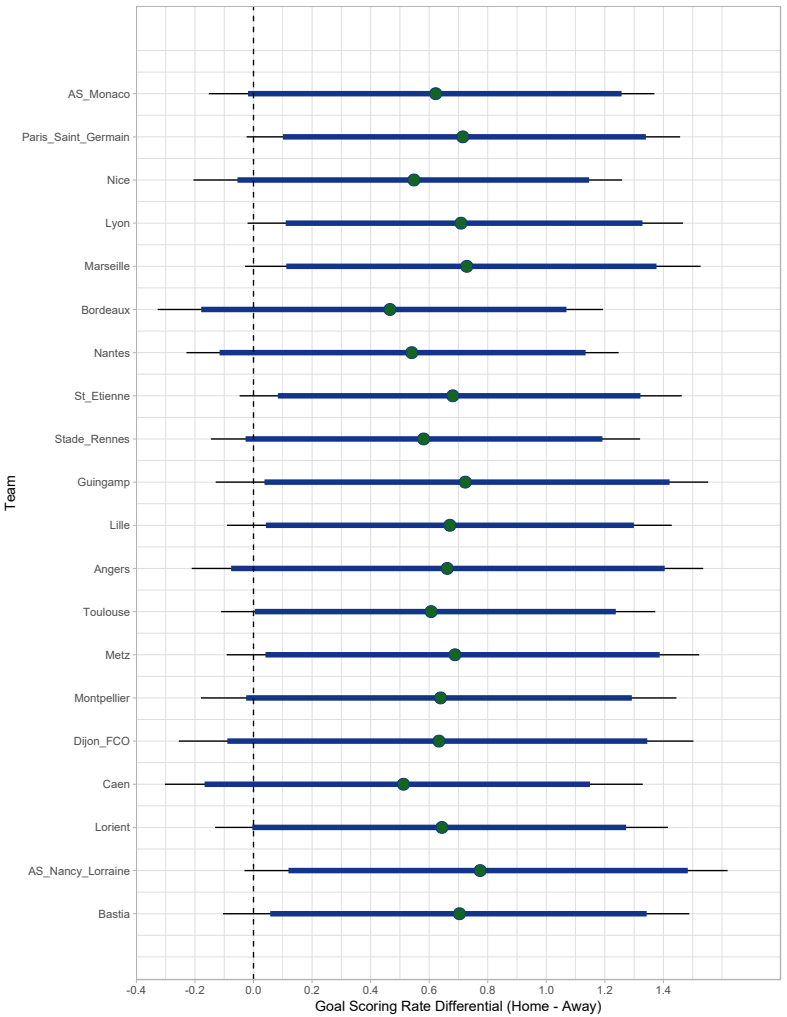


**Fig. 8:** Home Field Advantage Posterior Plot for Bundesliga Teams

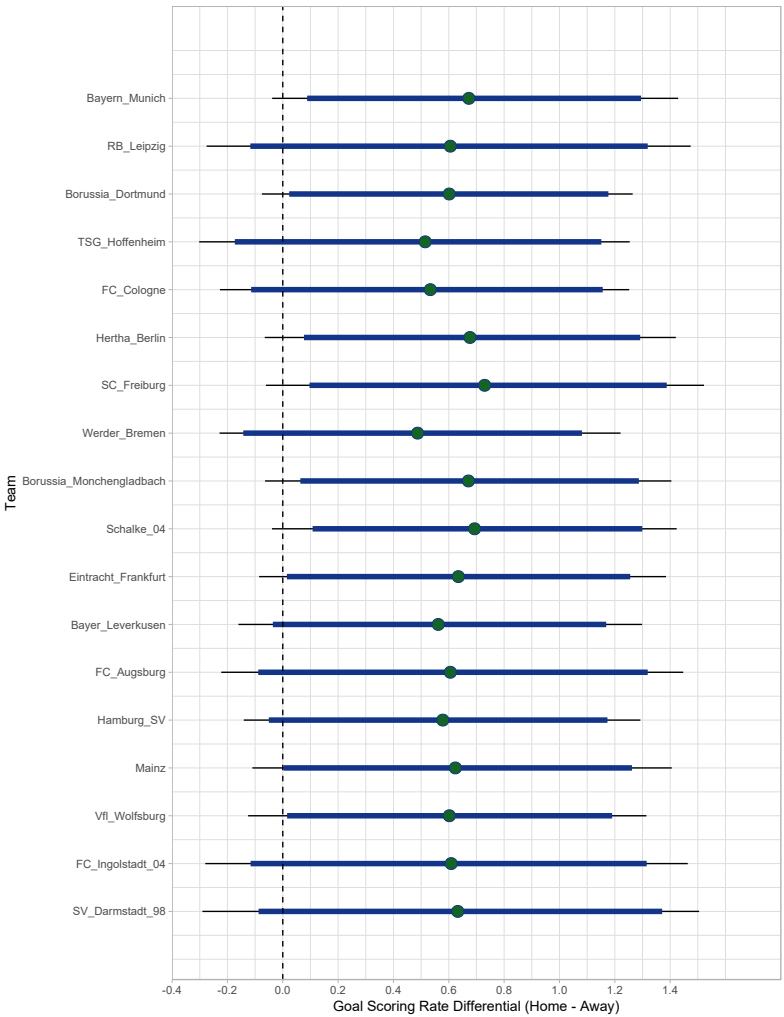
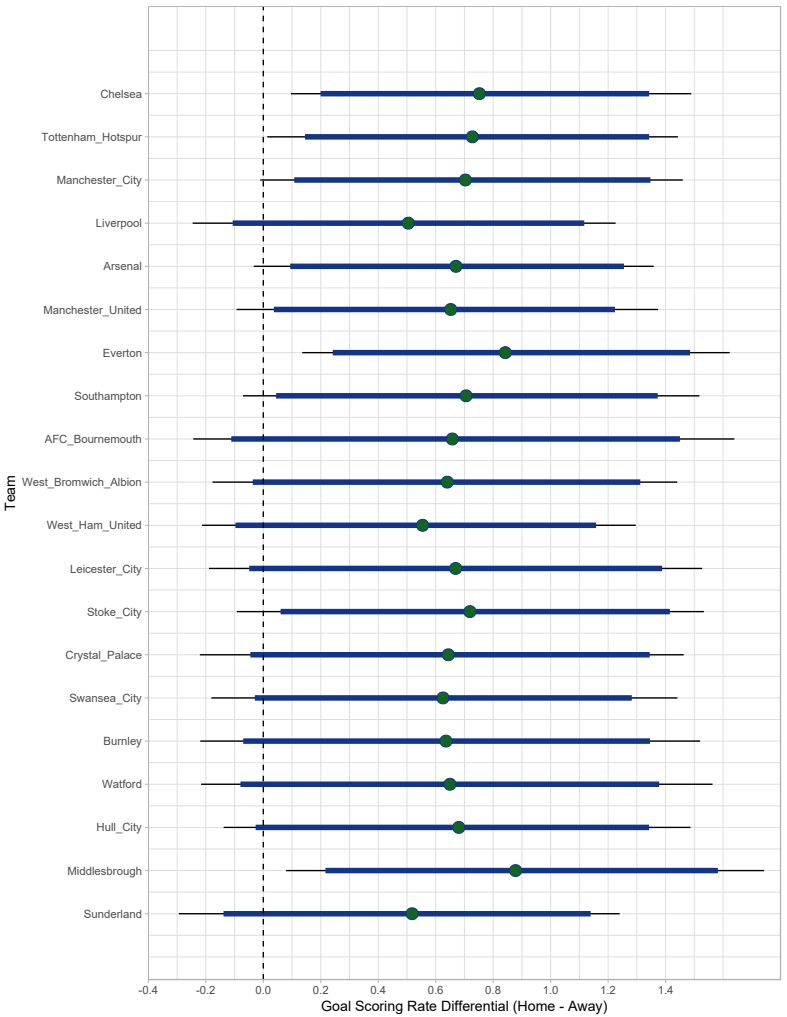


**Fig. 9:** Home Field Advantage Posterior Plot for English Premier League Teams



# References

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

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.

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).