RESEARCH ARTICLE

Team Contingent or Sport Native? A Bayesian Analysis of Home Field Advantage in Professional Soccer

ARTICLE HISTORY
Compiled February 17, 2020

ABSTRACT
We elevate the popular Log-Normal model of Home Field Advantage from the game level to season, and sport(professional soccer) levels. Using scoring performance data from ESPN FC, we fit a Bayesian multilevel nested model to the parameters in our proposed hierarchical model of HFA, allowing information obtained from the season level to inform the inferences about scoring capabilities at the upper team and sport levels. On the one hand, our analysis reveals that much of HFA is attributed to the nature of the sport of interest. Team level sources of HFA , on the other hand, can only be attributed to a handful of clubs among 98 studied.

Abbreviations: **HFA**-Home Field Advantage; **HWP**-Home Winning Percentage; **AWP**-Away Winning Percentage; **UEFA**-Union of European Football Associations; **MHG**-Most Home Goals; **MAG**-Most Away Goals **CI**: credible interval; **IFAB**-International Football Association Board; **VAR**-video assistant referee

KEYWORDS
European professional soccer leagues; Home Field Advantage; Poisson generative process; Stan and CRAN-R; Predictive accuracy and Credible Intervals (CI)

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 a balanced home and away schedule" (Courneya & 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 played at which place. 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).

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 the 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) Sort sources of HFA simultaneously into respective sport, team related contributions.
(4) Foster a refreshing perspective on exposition of HFA and advocate exploring inter-sport HFA as a potential venue for future research.

The remainder of the paper is structured 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 compute and display the posterior distributions of the unobserved model parameters, given the observed data collected from ESPN FC website. Also in the same section, we evaluate the fit of the hierarchical model in the context of model comparison and posterior predictive checking. We close our paper with limitations and directions for future HFA research.

2. Review of Literature

The speed and movement in top-level football competition, allied to the intense scrutiny of media on the actions on the field, means that officiating crews must be well-prepared, possess the tactical acumen, the mental strength to withstand pressure and the ability to take split-second decisions with confidence and consistency. Such split-second decisions made under severe pressure from home crowds have been proved to show systematic favoritism for the home squad both experimentally (A. Nevill,

Balmer, & Williams, 1999; A. M. Nevill, Balmer, & Williams, 2002) and in observational settings (Dohmen, 2008; A. M. Nevill, Newell, & Gale, 1996).

Following the episode of hooligan-induced riot on Feb. 2, 2007, the Italian authorities forced soccer clubs with deficient security standards at their home stadiums to play their home games with no spectators. The ruling inadvertently created a sizable and scarece sample of 21 professional soccer games played before empty bleachers. Pettersson-Lidbom and Priks (2010) seized on this historical opportunity and contrasted the performance metrics of both referees and players by looking at the matches played by the same team and officiated by the same referee crew. They found convincing evidence for the effect of spectators on referees, manifested in the 70% (26%) drop for red (yellow) cards issued favoring the home team. On the other hand, the players did not seem to play any differently whether the yelling crowds were present or absent.

With game as the anchoring unit of analysis, Courneya and Carron (1992) developed a conceptual framework along the timeline axis of a typical 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, Loughhead, and Bray (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 time-invariant tendency, coupled with the largest betrayed effect, makes professional soccer an excellent venue for studying HFA at a more aggregate level beyond individual matches and even seasons.

As Boyko, Boyko, and Boyko (2007) pointed out, traditional frequentist statistical approaches don't address whether referees or players alone or combined channel crowd effects to impart on final match outcome. Bayesian inferential approach separates itself from its frequentist counterpart due to its emphasis on modeling all forms of uncertainty rather than providing point estimates. Regardless of the inferential approaches taken, one major goal of statistical analysis is model selection among a set of competing models that were assumed to have generated the observed data. With the aid of posterior predictive checking (Gelman, Meng, & Stern, 1996), researchers can assess the fitness of competing models with realized discrepancies between the actual and replicated data points.

With the rare exceptions of Baio and Blangiardo (2010), Gajewski (2006) and Glickman and Stern (1998, 2005), Bayesian statistical approach has not been widely adopted in the analysis of HFA. One unique feature of the Gajewski (2006) study is that they model longitudinal data across seasons while utilizing a unique HFA parameter dedicated to each team involved in the investigation. One problem common to these Bayesian studies is that they directly model the match-based goal differentials between the hosting and visiting teams. Although Poisson-logNormal models (Baio & Blangiardo, 2010; Karlis & Ntzoufras, 2003) treat host goals and visitor goals separately, model parameters such as home advantage, offensive and defensive scoring intensities

still were estimated simultaneously using game-level observed pairs of goals. Such estimates based on score differentials or score-pairs of individual matches are effectually blending home team's HFA and visitors' guest field disadvantage. Thus, we sense an urgent need to break down HFA into sub-components, which we can pinpoint to their originating sources.

The subject and methodology-matter motivations for this research lie in the decomposition of home field advantage in a multilevel format that naturally reflects the structure of professional soccer competition. With the help of Bayesian nested modeling, we shall demonstrate next how easily we can alter the structural complexity of the main candidate model with just a few lines of code.

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 parameters encapsulated by the model (Gelman et al., 2014).

For our project, the data set contains the season-level best home and away scoring numbers (y_{is}^H and y_{is}^A respectively) of each club i in the Top 5 leagues. We treat the generative processes of y_{is}^H and y_{is}^A as similar but independently governed by their own respective parameters. At the measurement level, we encode y_{is}^H and y_{is}^A into corresponding latent scoring rate λ_i^H and λ_i^A with Poisson distribution, which is a commonly accepted distributional model for sports count data (Miller, 2015):

Instead of modeling directly HFA as constant for all teams across seasons,we acquire inference about the team level HFAs (δ_i) by assessing the difference between the latent scoring intensity θ_i^H and θ_i^A. By the same token, we can derive sport-level HFA (Δ) via assessing the differential between the hyperparameters μ^H and μ^A. We specify the hierarchical model of HFA formally as a set of equations consisting of (1), (2), (3).

$$\begin{cases} y_{is}^H \sim \text{Poisson}(\lambda_i^H) \\ y_{is}^A \sim \text{Poisson}(\lambda_i^A) \end{cases} \tag{1}$$

$$\begin{cases} \log \lambda_i^H = \theta_i^H \\ \log \lambda_i^A = \theta_i^A \\ \delta_i = \theta_i^H - \theta_i^A \end{cases} \tag{2}$$

$$\begin{cases} \theta_i^H \sim Normal(\mu^H, \tau^H) \\ \theta_i^A \sim Normal(\mu^A, \tau^A) \\ \Delta = \mu^H - \mu^A \end{cases} \tag{3}$$

4. Data and Results

For practical reasons, the Top 5 leagues serve as a convenient sample as performance data at season level are reliable and retrievable via internet. On ESPN FC website,

we find a pair of venue-delineating (home and away) goal scoring metrics used to characterize a professional soccer club's regular season. 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_{is}^{H}) = maximum goals scored in a single match played at home. For the season 2015/2016, Real Madrid's $y_{3,1,16}^{H}$ is 10. They beat Rayo Vallecano by 10-2 at Santiago Bernabéu Stadium on 12/20/2015.
- Most Away Goals (as y_{is}^{A}) = maximum goals scored in a single away match. For the season 2015/2016, Real Madrid's $y_{3,1,16}^{A}$ is 6. They defeated Espanyol 6-0 on 9/12/2015 at RCDE stadium.

Table 1 provides the summary statistics of y_{is}^{H} and y_{is}^{A}. Both averages and medians evince the existence of positive goal differential between maximum home and away goals. However, the MAG is more skewed than MHG in that the max. of MAG is actually greater than that of MHG. Deletion of such outliers is not an option in conducting sports analytics, because they are quintessential of the underlying exceptional performance by athletes. Fortunately, Bayesian statistics can accommodate such wide dispersion of data points with alternative distribution functions other than the commonly-applied Gaussian PDF (normal probability density function).

Table 1. Descriptive Statistics

	Mean	Median	Std. Dev.	Min.	Max.	Skewness	Kurtosis
MHG (y_{js}^{H})	3.634	4	1.676	0	9	0.246	0.034
MAG (y_{is}^{A})	2.884	3	1.676	0	10	0.627	0.786

Goal Scoring Metrics: Most Home & Away Goals at the Season Level

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 & Gelman, 2014).

The sport level estimates of goal-scoring rate differential are shown in figure 1 as shift from the 0. The outer contour line depicts the 95% uncertainty intervals, while the shaded area underneath covers the corresponding 90% uncertainty intervals. The light bar in the middle represents the mean. From figure 1, we observe absolutely strong (95%) manifestation of HFA for the sport of soccer (in the top panel). The goal-scoring differentials are centered around 0.16 goals with comparable lengths of uncertainty intervals.

To summarize team level estimates, we use the inner thick line to represent the 90% uncertainty interval and the outer thin line for 95% uncertainty level respectively. The dot in the middle still represents the mean as before. As shown in figure 2, Real Madrid is the only club among all 20 in La Liga enjoys strong HFA with 95% CI (credible Interval). Another five clubs exhibit only moderate HFA with 90% CI not encompassing the null 0 point.

Among 20 teams in Serie A (shown in figure 3), AS Roma is the only club possess strong HFA with 95% CI (outer thin line) not touching the null 0 line. Lazio is another team enjoys moderate HFA with 90% statistical significance.

Among the 20 teams in French Ligue 1 (figure 4), PSG shows strong HFA, while Lyon and Marseilles borderline the threshold of 95% statistical significance. Three other team enjoy HFA only to some moderate extent.

For teams in Bundesliga as shown in figure 5, Bayern Munich is the one in possession of dominant HFA with 95% statistical significance, while Hertha Berlin and Schalke

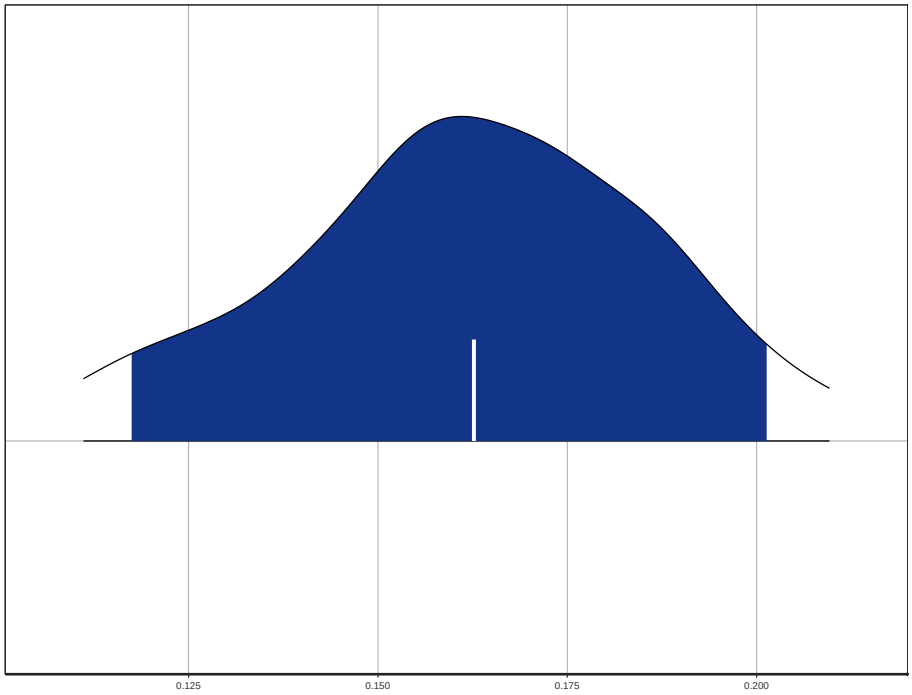

Figure 1. Posterior Density Plot (Δ) HFA Impact at Sport Level

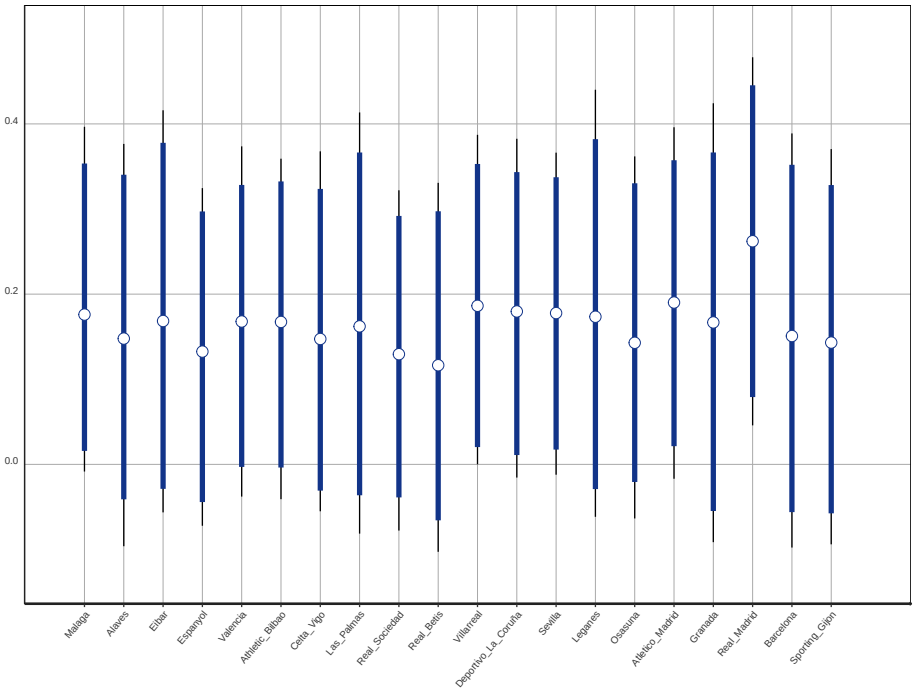

Figure 2. HFA (δ_i) Posterior Plot for Clubs in La Liga

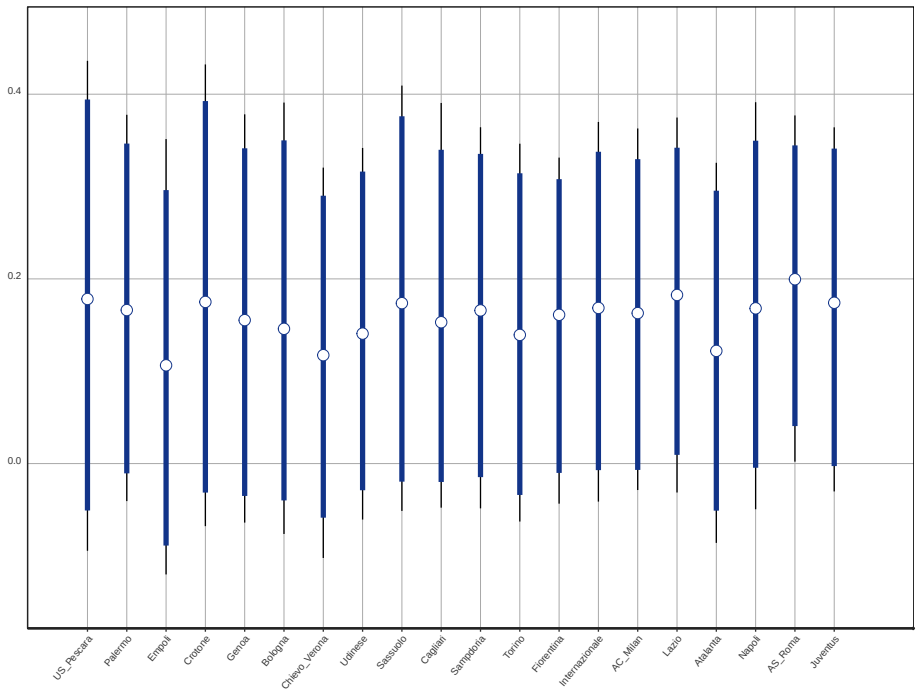

Figure 3. HFA (δ_i) Posterior Plot for Clubs in Serie A

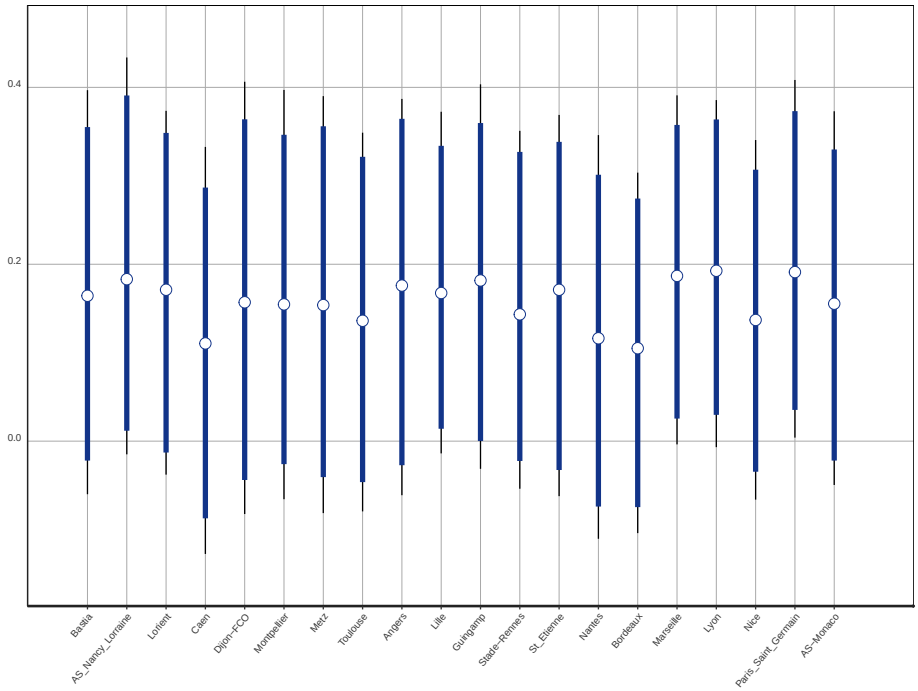

Figure 4. HFA (δ_i) Posterior Plot for Clubs in Ligue 1

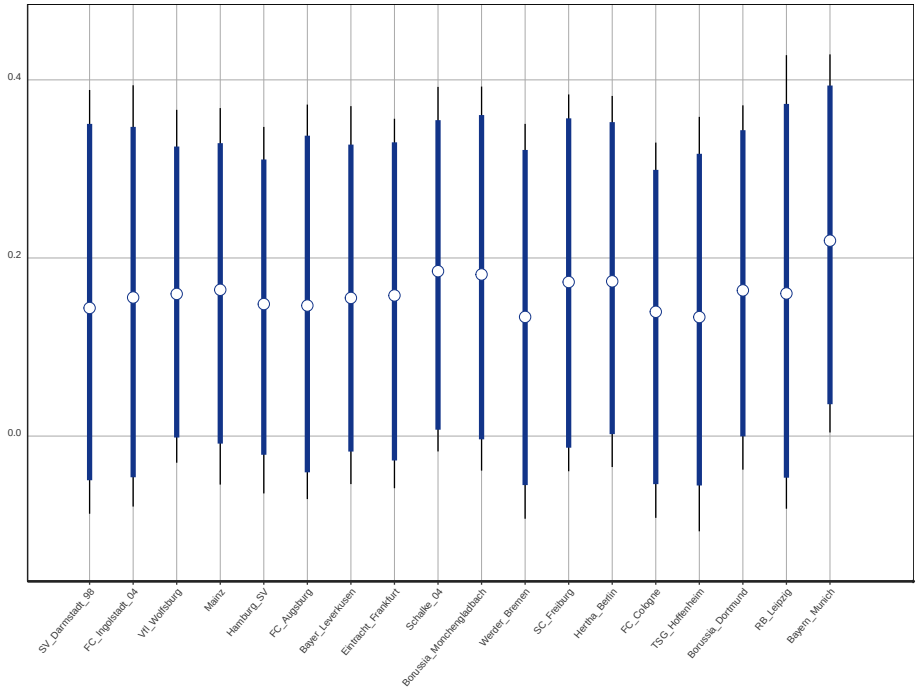

Figure 5. HFA (δ_i) Posterior Plot for Clubs in Bundesliga

04 enjoy moderate HFA.

Among teams in English Premier League (in Figure 6), Chelsea, Everton and Middlesbough are the three teams enjoy strong HFA, while Manchester City, Tottenham Hotspurs and Arsenal show signs of moderate HFA.

In summary, we observe, among 98 clubs included in this study, a total of 23 teams exhibit posterior 90% CIs not encompassing the null zero line. Further, 7 out of the 23 teams exhibit 95% CIs not encompassing the null zero line. It is worthy pointing out that the seven clubs cherishing strong HFA include usual league champions (Real Madrid, PSG, Bayern Munich, Chelsea), as well as non-champion caliber teams like Everton and Middlesbough.

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 50% and 25% uncertainty intervals (UI) based on the replicated results. We observe that all of the actual ydiffs are in the 50% UIs and 97% in the 25% UIs.

5. Discussion

In this work, we give home field advantage full Bayesian treatment with unique hierarchical structure. Departing from the tradition of treating HFA as a constant intercept for all teams and across seasons, we are able to model HFA indirectly and explore the locality of sources of HFA. We tested the proposed model with maximum home

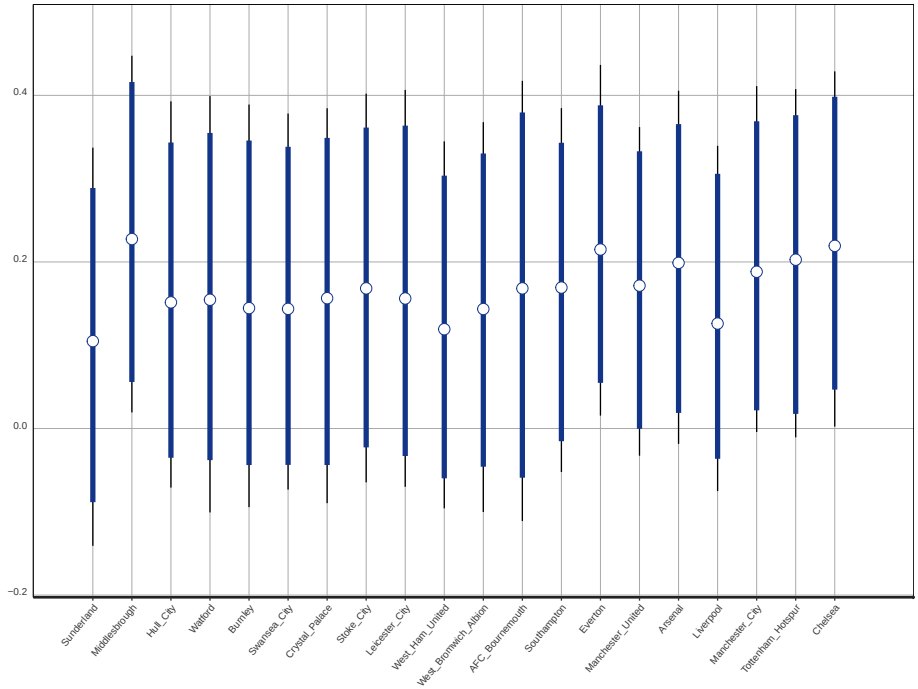

Figure 6. HFA (δ_i) Posterior Plot for Clubs in English Premier League

and away scoring data covering seasons of 2000/01 to 2016/17. The results reveal that HFA exists sport-wide while only a handful of teams command statistically significant effect.

We all know that 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 and quantifying the subjectivity of refereeing standards. Refereeing controversies regularly arise in professional competitions. Barcelona coach Ernesto Valverde complained on March 1, 2018 about an "invisible penalty" awarded to Las Palmas during the match that ended in 1-1 draw. The Spanish Football Federation (RFEF) has announced it is seeking approval from the International Football Association Board (IFAB) to roll out VAR (video assistant referee) in La Liga at the start of the 2018-19 season. VAR is already active in the top divisions in both Italy and Germany, while in England it has been tested on a trial basis in the 2017/18 season in selected domestic cup games.

Systems like VAR in soccer are purported to correct clear and obvious refereeing errors, regarding decisions on goals, red cards, penalties and cases of mistaken identity. If implementation of such machine-assisted officiating systems becomes widespread, we can expect a clear downward trend with regard to the effect size of home field advantage. When enough machine-generated officiating data are available, we should incorporate the technology related factors into our models accordingly and assess their contributions to our understanding of HFA at the sport level. Winning on home turf and picking up points away has long been cited as the path to success in soccer. We would argue that the same mentality is driving home field advantage, as long as the enthusiastic fans are in the stands and voicing out their enthusiasms to the right target

- the referees supposedly.

References

Atkins, C. (2013). How much does home-field advantage matter in soccer? *B/R*. Retrieved from http://bleacherreport.com/articles/1604854-how-much-does-home-field-advantage-matter-in-soccer

Baio, G., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, *37*(2), 253-264. Retrieved from https://doi.org/10.1080/02664760802684177

Boyko, R. H., Boyko, A. R., & Boyko, M. G. (2007). Referee bias contributes to home advantage in english premiership football. *Journal of Sports Sciences*, *25*(11), 1185–1194.

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

Dohmen, T. J. (2008). The influence of social forces: Evidence from the behavior of football referees. *Economic Inquiry*, *46*(3), 411–424.

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., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). CRC press Boca Raton, FL.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 733–760.

Glickman, M. E., & Stern, H. S. (1998). A state-space model for national football league scores. *Journal of the American Statistical Association*, *93*(441), 25–35.

Glickman, M. E., & Stern, H. S. (2005). A state-space model for national football league scores. In *Anthology of statistics in sports* (pp. 23–33). SIAM.

Hoffman, M. D., & 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.

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*(3), 381-393. Retrieved from https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9884.00366

Miller, T. W. (2015). *Sports analytics and data science: Winning the game with methods and models (ft press analytics)*. Pearson FT Press.

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

Nevill, A., Balmer, N., & Williams, M. (1999). Crowd influence on decisions in association football. *The Lancet*, *353*(9162), 1416.

Nevill, A. M., Balmer, N. J., & Williams, A. M. (2002). The influence of crowd noise and experience upon refereeing decisions in football. *Psychology of Sport and Exercise*, *3*(4), 261–272.

Nevill, A. M., Newell, S. M., & Gale, S. (1996). Factors associated with home advantage in english and scottish soccer matches. *Journal of Sports Sciences*, *14*(2), 181–186.

Pettersson-Lidbom, P., & Priks, M. (2010, aug). Behavior under social pressure: Empty italian stadiums and referee bias. *Economics Letters*, *108*(2), 212–214.

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*.