# SAGE Research Methods Cases: Business & Management

**Authors: Please complete only the white fields below.**

| | | |
|---|---|---|
| Case Study Title<br>**There are no sources in the current document.**e | | [Using Bayesian Inferential Approach to Analyse Home Field Advantage of 2020 Bundesliga Teams] |
| Authors. **Please add additional rows for co-authors if necessary.** | | |
| 1 | Name | C.J. (Chaojie) Duan |
| | Author email | [Data.Scientist@dulun.com] |
| | Affiliation, country | [DRC Lab of Dulun Research & Consulting, U.S.A.] |
| | SAGE Author ID | [office use only] |
| Author bio.<br>*Please include a separate biography for each author. Maximum of 200 words per author. Bios will not be copy-edited; please ensure they are correct.* | | [Dr. Duan is the founder and chief research data scientist of DRC Lab of Dulun Research & Consulting (DRC). He earned his Ph.D in Management from Clemson University in 2007. His research interests include BPO (business process outsourcing), Bayesian Inference, Sports Analytics, Media Mix Model, and State Space Model. He has published articles in journals like Journal of Business Analytics, International Journal of Production Research, European Journal of Information Systems, International Journal of Management Research and Reviews, Studies in Business and Economics, Operations Management Education Review, South Asian Journal of Management Sciences.] |
| 2 | Name | |
| | Author email | [Insert contact email] |
| | Affiliation, country | [Insert institution and country] |
| | SAGE Author ID | [office use only] |
| Author bio.<br>*Please include a separate biography for each author. Maximum of 200 words per author. Bios will not be copy-edited; please ensure they are correct.* | | [Maximum of 200 words] |
| 3 | Name | |
| | Author email | [Insert contact email] |

<table>
<tr><td></td><td>Affiliation, country</td><td>[Insert institution and country]</td></tr>
<tr><td></td><td>SAGE Author ID</td><td>[office use only]</td></tr>
<tr><td colspan="2">Author bio.<br>*Please include a separate biography for each author. Maximum of 200 words per author. Bios will not be copy-edited; please ensure they are correct.*</td><td>[Maximum of 200 words]</td></tr>
<tr><td colspan="2">Discipline</td><td>Business & Management [D12]</td></tr>
<tr><td colspan="2">***Sub-discipline within Business & Management***</td><td>[Click here to select discipline]</td></tr>
<tr><td colspan="2">**Academic Level of intended readership.** Select the level best suited to the case study content.</td><td>[Click here to select level]</td></tr>
<tr><td colspan="2">Please include the reference for any **published articles based on the research project** this case study discusses.</td><td>[Insert an APA-style reference, 7th edition, for any publications based on the research this case study reflect upon.]</td></tr>
<tr><td colspan="2">*For office use only:*</td><td></td></tr>
<tr><td colspan="2">Title/Spin ID</td><td></td></tr>
<tr><td colspan="2">Access/Product Code</td><td></td></tr>
<tr><td colspan="2">ISBN</td><td></td></tr>
<tr><td colspan="2">DOI</td><td></td></tr>
<tr><td colspan="2">URI</td><td></td></tr>
<tr><td colspan="2">URL</td><td></td></tr>
<tr><td colspan="2">Copyright year</td><td>2023</td></tr>
<tr><td colspan="2">Copyright statement</td><td>© SAGE Publications Ltd 2023</td></tr>
</table>

**Your case study must not exceed 5000 words. Discussion Questions, MCQs, and References do not count towards this limit.**

**Please ensure you have read through this template and the manuscript guidelines before you begin writing your case study.**

# Abstract

*The abstract should be a concise summary of this case study. What original research is this case study based on? What aspect of the research process, or specific methodological and practical challenges, will your case study address? Emphasize what the reader will learn from reading this case study, and how they might apply it in their own research practice. Please do* not *cite references within the abstract.*

[Using the 2020 Bundesliga match results, this case study demonstrates how to build and fit a Poisson-logNormal model in Stan. Also, we model the impact of the outbreak of COVID-19 pandemic on home field advantage. By contrasting the 82 matches played with empty stands with the other 224 regular games, our fitted model reveals that home field advantage vanishes with the removal of fans from the playing venue. In addition to model formulation and fitting, readers will also learn how to diagnose a model in Stan and perform posterior predictive checking.]

---

# Learning Outcomes

*Learning outcomes must explain what the reader will learn from reading your case study. Readers should be learning about research methodology, methods/analytics, and practicalities. How will the reader be able to apply what they have learned to their own research practice?*

*Please refer to these learning outcomes when writing your case study. Your case study must satisfy each proposed outcome. It is vital that you provide achievable and measurable learning outcomes. Please start each learning outcome with an action verb.*

*See the links below for guidance on writing effective learning outcomes:*

- *Writing learning outcomes*
- *Bloom's Taxonomy Action Verbs*

[Insert 3–5 learning outcomes under the statement provided below: “By the end of this case, readers should be able to . . .”].

By the end of this case study, readers should be able to . . .

- [Formulate a Poisson-logNormal model in Stan]
- [Interpret outputs from Stan model fitting]
- [Perform posterior predictive model checking]

---

# Case Study

Insert your case study below. The main body of the text should be between 2,000 and 5,000 words.

*We encourage the use of headings and sub-headings add structure to the body of your case, enhance online discoverability and make your case easier to read on screen.*

*Suggested top-level headings (H1s) are included in the template. If you are using subheadings in a section, please apply the appropriate Word style tags (H2 or H3) so that the desired nesting structure is clear.*

***Every section with a heading must be followed by a Section Summary.*** *Each Section Summary should consist of 2-3 bullet points, written out as full sentences, which summarize the key information in the section.*

# Project Overview and Context

*Here you can include information about the focus of your research project. Why were you interested in studying this topic? In what context was this research undertaken? If you are reflecting on industry research, please describe how your analysis was, or could be, used by the relevant business.*

*This section should not read as a literature review but should explain the rationale behind your research project. In the following sections you will be concentrating on your research methodology, which is the primary focus of your case study.*

[On June 11, 2021, the UEFA (the Union of European Football Associations) EURO 2020 finally kicked off in front of 16000 fans at a 1/4-full *Stadio Olimpico* between Italy and Turkey. The tournament was originally scheduled from June 12 to July 12, 2020. Due to the outbreak of COVID-19 pandemic, it was rescheduled from June 11 to July 11,2021, with 11 host cities staging 51 matches. The 11 game sites - Amsterdam, Baku, Bucharest, Budapest, Copenhagen, Glasgow, London, Munich, Rome, Saint-Petersburg and Seville - have all approved fans will be allowed to fill between 25% and 100% of full stadium capacity. Wembley is the stadium for both semi-finals as well as the final on Sunday, July 11.

Euro 2020, in the calendar year of 2021, had a minimum of 24 games with one side playing in their home country. Italy, Denmark, Russia, Netherlands, England, Scotland, Spain, Hungary and Germany were guaranteed a minimum of two or three home group matches, after which the knockout matchups could deliver more home comforts. According to an article in *Sporting Life* (Taylor, 2021), England potentially had most to gain, with three home group matches ensured, a possible last-16 game, and the semi-final and final earmarked for Wembley Stadium. Before the kickoff, England was No. 4 in the FIFA world soccer ranking behind Belgium and France at No.1 and 2 respectively. Without their presumed home field advantage (HFA) in the group stage, England's likelihood of topping Group D could fall from 70% to 60%. It was largely believed that HFA propelled the England squad all the way into the European Championship final.

Ubiquitous across all sports, the hosting teams tend to consistently perform to a higher level than their visiting opponents in familiar home surroundings and in front of a largely supportive fan base. HFA in terms of HWP (home winning percentage) has been well documented for a variety of sports, even though the contributing factors are still being debated. The scientific definition of HFA 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 & Carrón, 1992). 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. A second leg of any UEFA's Champions League knock-off series is favorable to playing away with the aggregate scores still in balance after the first leg competition.

In the tradition of modeling sports game outcomes, HFA has always been treated as invariant across all participating teams (see Lopez, Matthews & Baumer, 2017). Even though empirical results indicted average HFA values varied significantly from team to team (Glickman & Stern, 1998), it was the author's work (Duan & Chakravarty, 2021) that, for the first time, addressed HFA as variant across teams. In this case, we will demonstrate how to model the varying HFA effect using the scores from the 2020 Bundesliga matches.]

# Section Summary

*What are the key points the reader should take from this section?*

- *Euro 2020 took place in 2021 due to COVID-19 Pandemic*

- *Euro 2020 brought back live audiences back to stadiums*
- *Modeling of HFA has always treated it as static.*

# Research Design

*Describe how you designed your study, and why you designed it that way. Explain the rationale behind any fundamental decisions you made. In later sections you can describe any changes that were made to your original design.*

*Ensure that you define and explain any key terms for the reader.*

[The COVID-19 pandemic has paralyzed the world and could change sports forever. The public health crisis has turned soccer matches into *a puerta cerrada*. The sudden and swift removal of spectators from stands would result in the evanescence of any crowd effects. The pandemic-split 2019-2020 season would provide a once-in-history opportunity to corroborate Schwartz & Barsky (1977)'s assertion that HFA is mainly product of social congregation. For this project, the data set contains the game-level scoring numbers $y_{ik}$ and $y_{jk}$ of each game k played between the hosting team i and the visiting team j. We culled the data from the official website of Bundesliga (www.bundesliga.com). For the 2019-2020 season, the 18 teams, on a typical match day, play 9 games. Each pair of two teams play twice (home and away) during the season. The outbreak of COVID-19 separates the season into 224 regular games played before/on March 11 of 2020 and the remaining 82 games played without spectators after/on May 16 of 2020. A sample of game results are shown in Table 1. The COVID column assigns values of 0 to the 224 games packed with live audiences and 1 to the 82 games played in empty venues.

Table 1: 2020 Bundesliga Sample Game Results

| Time | Host | Visitor | $y_{ik}$ | $y_{jk}$ | COVID |
|---|---|---|---|---|---|
| 2019/8/16 | Bayern Munich | Hertha Berlin | 2 | 2 | 0 |
| … | … | … | … | … | 0 |
| 2020/3/8 | Mainz | Düsseldorf | 1 | 1 | 0 |
| 2020/5/16 | Dortmund | Schalke 04 | 4 | 0 | 1 |
| … | … | … | … | … | 1 |
| 2020/6/27 | Union Berlin | Düsseldorf | 3 | 0 | 1 |

Equations (1) to (9) represent the full Poisson-logNormal Bayesian model used in Baio & Blangiardo (2010) and Duan & Chakravarty (2021) to model the typical outcomes from a soccer match. The measurement part of the model in (1) and (2) translates the two integer-valued $y_{ik}$ and $y_{jk}$ into team i and j's scoring intensities respectively. In the Bayesian modeling context, equation (1) should read as "the goal total of $y_{ik}$ is drawn from a Poisson distribution governed by the parameter of $\theta_{ik}$".

$$y_{ik} \sim Poisson(\theta_i) \quad (1)$$
$$y_{jk} \sim Poisson(\theta_j) \quad (2)$$

The relationship layer of the model, as presented in (3) to (5), decomposes the scoring intensities (θs) further into quantities of our research interest, such as HFA ($\Delta_{COVID}$), offensive capabilities ($\lambda_i^O$ and $\lambda_j^O$) and defensive capabilities ($\lambda_i^D$ and $\lambda_j^D$). Please note that the site-specific HFA only enters and exists in (3), which decomposes the host's scoring intensity into three parts, including HFA. For the visitor's scoring intensity, it is only broken into visitor's offensive capability and the host's defensive capabilities. The binary COVID variable ensures that Δ alternates between regular games and pandemic-stricken games.

$$log\theta_i = \Delta_{COVID} + \lambda_i^O + \lambda_j^D \quad (3)$$
$$log\theta_j = \lambda_j^O + \lambda_i^D \quad (4)$$
$$COVID = \begin{cases} 1, \forall\ k \leq 224 \\ 0, \forall\ k > 224 \end{cases} \quad (5)$$

Lastly, we present the priors layer of the model in (6) - (9). The purpose of those hyper-parameters (μ, σ) is to regularize and govern the generative process for those regular parameters (λ, Δ).

$$\lambda_i^O, \lambda_j^O \sim Normal(\mu^O, \sigma^O) \quad (6)$$
$$\lambda_i^D, \lambda_j^D \sim Normal(\mu^D, \sigma^D) \quad (7)$$
$$\mu^O, \mu^D \sim Normal(0, 3) \quad (8)$$
$$\Delta \sim Nomal(0, 0.1) \quad (9)$$

The nine equations of the above model specification form the three parts of the Poisson-logNormal model.]

# Section Summary

*What are the key points the reader should take from this section?*

- *The 2019-20 Bundesliga season is split into two parts by COVID-19.*
- *The Poisson-LogNormal model translates scores into team capabilities.*
- *The binary COVID variable is introduced to mark the absence of spectators.*
- *Poisson and Normal refer to distributions, while log is transformation.*

# Research Practicalities

*Includes a discussion of practical and ethical considerations you had to navigate when conducting your research. Were there challenges that had to be overcome to access participants or data? Were your personal skills compatible with the research you were intending to carry out? What of time constraints, costs, and resources? What ethical considerations were essential?*

[Stan (Stan Development Team, 2022) is not only an expressive probabilistic programming language for coding probability models, but also provides the tools needed to fit models to existing data. It abstracts the technicalities of inference and allows users to focus on modeling and performing posterior data analysis for evaluating the results from MCMC (Markov Chain Monte Carlo) simulations.

A full `stan` file contains five coding blocks: `data`, `transformed data`, `parameters`, `transformed parameters`, `model`, and `generated quantities`. In the `data` block, we declare variables that take their values from the computing environment (RStan or PyStan). In our case, the `data` block contains declarations of five seven variables: two integer score variables ($y_h$ and $y_v$), two integer team identity variables ($x_h$ and $x_v$), the binary COVID variable, and two auxiliary variables (T for the total number of teams and G for total number of games). The complete `data` block in our model is as follows:

```
data {
      int<lower = 2> T; // total number of teams
      int<lower = 1> G; // total number of games
      int<lower=1, upper=T> xh[G]; // hosting team identities
      int<lower=1, upper=T> xv[G]; // visiting team identities
      int<lower=0> yh[G]; // host scores
      int<lower=0> yv[G]; // visitor scores
      int<lower=0, upper=1> COV[G]; // the binary COVID variable
}
```

The `<lower=?, upper=?>` statement inside declarations simply provides the upper and lower bounds, between which the variable can take legitimate values.

The `transformed data` block declares additional fixed variables and accommodates transformation of original data variables in the `data` block. In our case, we simply add 1 to the original COVID variable due to the fact index in `stan` always starts at 1.

```
transformed data{
      array[G] int<lower=1, upper=2> COV_t;
      COV_t = COV + 1;
}
parameters {
      vector[2] delta; // HFA alternating between regular and pandemic
      conditions
      vector[T] lambda_O_t; // Team offensive capabilities
      vector[T] lambda_D_t; // team defensive capabilities
      real mu_O; // hyper-team offensive capability for all teams
      real mu_D; // hyper-team defensive capability for all teams
      real<lower = 0> sigma_O; // hyper team offensive capability std. dev
      real<lower = 0> sigma_D; // hyper team defensive capability std. dev
}
```

In parameters block, we declare those free parameters (quantities) in our model as defined in equation (3) through (9) in last section. The `vector` and `real` identifier by default specify the parameter space is real-valued. Those dependent (unfree) parameters, like $\theta_i$, $\theta_j$ in equation (1) – (4), are handled in the transformed parameters block shown as below:

```
transformed parameters {
      vector[G] theta_H; // hosting team scoring intensities
      vector[G] theta_V; // visitor scoring intensities
      for (g in 1:G){
            theta_H[g] = exp(delta[COV_t[g] + lambda_O_t[xh[g] +
            lambda_D_t[xv[g]]);
            theta_V[g] = exp(delta[lambda_O_t[xv[g] + lambda_D_t[xh[g]]);
      }
}
```

In the above `for` loop, we use the `exp()` function instead of `log()` to establish the relationship as defined in equations (3) and (4.).

```
model {
      // priors
      mu_O ~ normal(0, 3);
      mu_D ~ normal(0, 3);
      delta ~ normal(0, 0.1);
      // logNormal part of the model
      lambda_O_t ~ normal(mu_O, sigma_O);
      lambda_D_t ~ normal(mu_D, sigma_D);
      // the Poisson measurement part of the model
      yh ~ poisson(theta_H);
      yv ~ poisson(theta_V);
}
```

The `model` block provides line-to-line correspondence to the equations (1), (2), (6) - (9) in reverse order. The final block in a standard `stan` file is the `generated quantities`, which is often used to compute predicted quantities based on parameters and variables declared and inferred in previous blocks.

```
generated quantities {
      int yp_H; // predicted Host goals
      int yp_V; // predicted Visitor goals
      yp_H = poisson_rng(theta_H);
      yp_V = poisson_rng(theta_V);
}
```

On top of the above six blocks, we could also add the `functions` block, which defines additional functions (as compared to those native or built-in functions, such as `abs() or log()`) to be called in any of those operative blocks.]

## Section Summary

*What are the key points the reader should take from this section?*

- *Coding in* `stan` *is like writing equations mathematically.*
- *Free parameters are governed by various distributions.*
- *Modeler can restrict parameters using* `lower` *and* `upper` *bounds.*

# Method in Action

*How did your research project play out in reality? Did it go according to plan, or did you need to adapt parts of the process? This should be a "warts and all" description and evaluation of how your chosen research method/approach actually worked in practice. What went well? What did not go to plan? What challenges did you face? How did you respond? Remember that cases should explore both the successes of your methodology and the challenges and problems. Both can provide rich learning opportunities.*

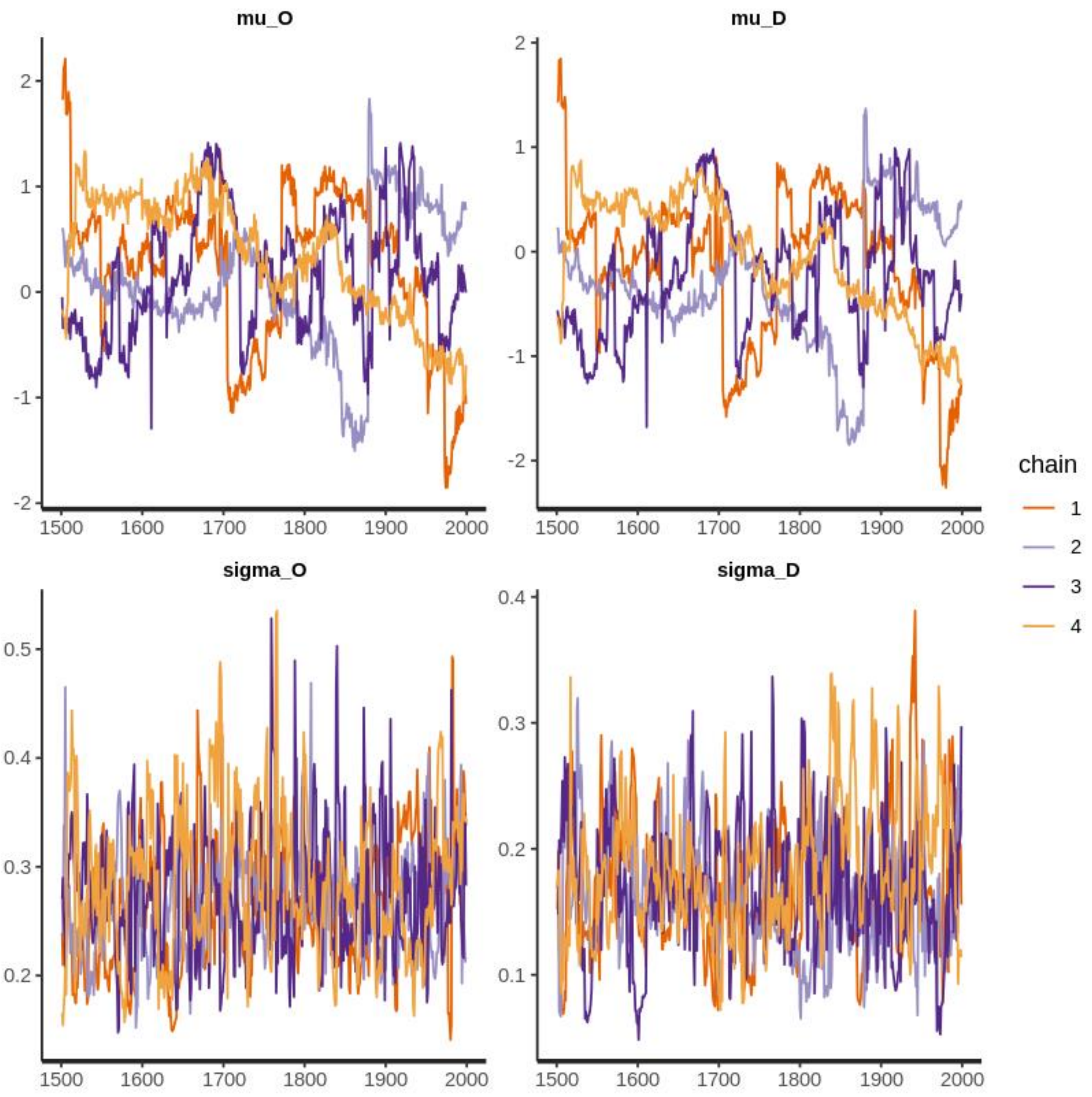
[The R (R Core Team, 2022) package `rstan` (Stan Development Team, 2020) is the computing environment and interface for modelers to feed data into the model and fetch the MCMC object `fit`. Using the default settings of Stan sampler, we run four separate Markov chains in parallel, with 1500 initial iterations as warmup and the last 500 iterations for sampling. During the sampling process, the following warning messages were issued by the sampler.

```
Warning message:
"There were 3 divergent transitions after warmup. See
https://mc-stan.org/misc/warnings.html#divergent-transitions-
after-warmup
to find out why this is a problem and how to eliminate them."
Warning message:
"There were 1 chains where the estimated Bayesian Fraction of
Missing Information was low. See
https://mc-stan.org/misc/warnings.html#bfmi-low"
Warning message:
"Examine the pairs() plot to diagnose sampling problems
"
Warning message:
"The largest R-hat is 1.13, indicating chains have not mixed.
Running the chains for more iterations may help. See
https://mc-stan.org/misc/warnings.html#r-hat"
Warning message:
"Bulk Effective Samples Size (ESS) is too low, indicating
posterior means and medians may be unreliable.
Running the chains for more iterations may help. See
https://mc-stan.org/misc/warnings.html#bulk-ess"
```

```
Warning message:
"Tail Effective Samples Size (ESS) is too low, indicating
posterior variances and tail quantiles may be unreliable.
Running the chains for more iterations may help. See
https://mc-stan.org/misc/warnings.html#tail-ess"
```

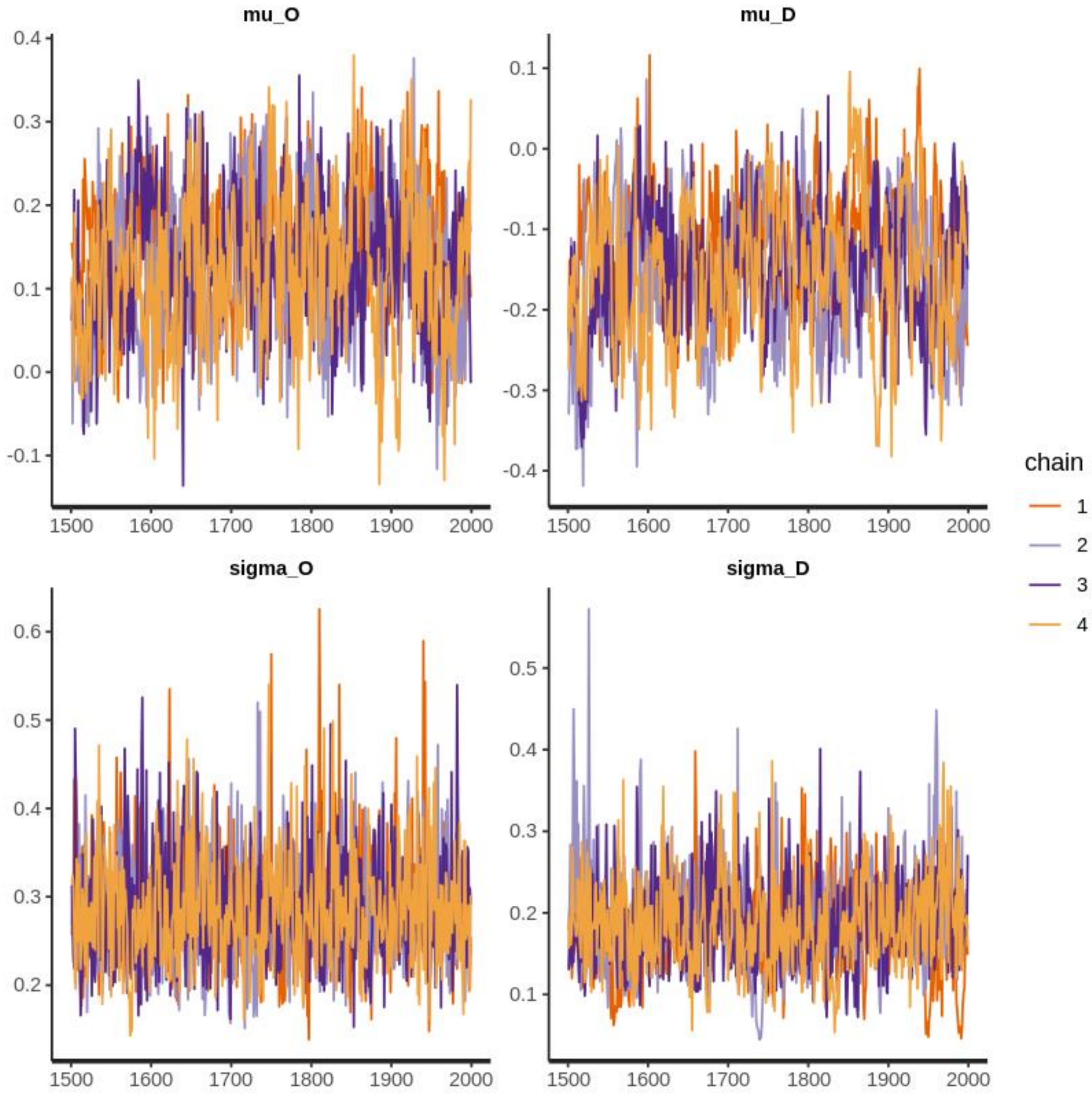
Next, we examine the traceplots for the four hyperparameters (`mu_O, mu_D, sigma_O, sigma_D`). Cleary, we can observe severe divergences in the traceplots for  the two mean hyperparameters (`mu_O` and `mu_D`). The presence of divergent transitions, as manifested in figure 1, poses the severe threat to the validity of the inference to be drawn.

Figure 1: Traceplots for Hyper-parameters with Original Values



To fix the problem of divergent transition, we need to revisit our model specification and tweak the hyperparameter values. After many iterations of experiments, we finally settled on the more informative (more sharply peaked) `normal(0, 0.1)` in place of the original `normal(0, 1)`. The newly tweaked hyperparameters yielded the following traceplots.

Figure 2: Traceplots for Hyper-parameters with Updated Values



The disappearance of divergent transition, as shown in figure 2, assured us that the inferential results we derive from the sampled object are credible and valid.

Table 2: Summary of Diagnostic Information – Team Defensive Capabilities

| Parameters | mean | sd | 5% | 50% | 95% | n_eff | Rhat |
|---|---|---|---|---|---|---|---|
| lambda_D_t[1] | -0.27 | 0.14 | -0.49 | -0.26 | -0.05 | 292.48 | 1.02 |
| lambda_D_t[2] | -0.33 | 0.14 | -0.57 | -0.33 | -0.10 | 269.34 | 1.01 |
| lambda_D_t[3] | -0.23 | 0.13 | -0.45 | -0.23 | -0.01 | 298.15 | 1.01 |
| lambda_D_t[4] | -0.03 | 0.14 | -0.26 | -0.03 | 0.20 | 458.70 | 1.01 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| lambda_D_t[5] | -0.30 | 0.14 | -0.53 | -0.30 | -0.07 | 285.70 | 1.01 |
| lambda_D_t[6] | -0.25 | 0.14 | -0.47 | -0.25 | -0.03 | 361.06 | 1.01 |
| lambda_D_t[7] | -0.09 | 0.14 | -0.31 | -0.10 | 0.14 | 445.98 | 1.01 |
| lambda_D_t[8] | -0.16 | 0.14 | -0.40 | -0.17 | 0.07 | 309.60 | 1.01 |
| lambda_D_t[9] | -0.34 | 0.14 | -0.55 | -0.33 | -0.12 | 241.49 | 1.01 |
| lambda_D_t[10] | 0.03 | 0.15 | -0.21 | 0.02 | 0.27 | 450.82 | 1.01 |
| lambda_D_t[11] | -0.06 | 0.14 | -0.28 | -0.06 | 0.18 | 524.20 | 1.01 |
| lambda_D_t[12] | -0.30 | 0.14 | -0.53 | -0.29 | -0.08 | 253.92 | 1.02 |
| lambda_D_t[13] | -0.01 | 0.15 | -0.25 | -0.02 | 0.23 | 466.13 | 1.01 |
| lambda_D_t[14] | 0.08 | 0.16 | -0.17 | 0.07 | 0.35 | 572.54 | 1.00 |
| lambda_D_t[15] | -0.38 | 0.14 | -0.62 | -0.38 | -0.15 | 280.79 | 1.01 |
| lambda_D_t[16] | -0.21 | 0.14 | -0.43 | -0.22 | 0.01 | 332.07 | 1.01 |
| lambda_D_t[17] | -0.21 | 0.13 | -0.44 | -0.21 | 0.01 | 310.27 | 1.01 |
| lambda_D_t[18] | -0.08 | 0.14 | -0.31 | -0.08 | 0.16 | 398.27 | 1.01 |

Table 3: Summary of Diagnostic Information – Team Offensive Capabilities

| Parameters | mean | sd | 5% | 50% | 95% | n_eff | Rhat |
|---|---|---|---|---|---|---|---|
| lambda_O_t[1] | 0.07 | 0.15 | -0.18 | 0.08 | 0.31 | 422.06 | 1.01 |
| lambda_O_t[2] | 0.03 | 0.15 | -0.23 | 0.03 | 0.27 | 430.25 | 1.01 |
| lambda_O_t[3] | 0.12 | 0.15 | -0.14 | 0.12 | 0.37 | 473.90 | 1.01 |
| lambda_O_t[4] | 0.59 | 0.14 | 0.36 | 0.59 | 0.81 | 300.10 | 1.02 |
| lambda_O_t[5] | -0.08 | 0.16 | -0.35 | -0.08 | 0.18 | 500.92 | 1.01 |
| lambda_O_t[6] | 0.29 | 0.15 | 0.05 | 0.29 | 0.53 | 361.17 | 1.01 |
| lambda_O_t[7] | 0.11 | 0.15 | -0.15 | 0.12 | 0.35 | 360.19 | 1.01 |
| lambda_O_t[8] | 0.19 | 0.15 | -0.05 | 0.20 | 0.43 | 440.42 | 1.01 |
| lambda_O_t[9] | 0.17 | 0.15 | -0.07 | 0.18 | 0.42 | 416.47 | 1.01 |
| lambda_O_t[10] | 0.56 | 0.14 | 0.33 | 0.55 | 0.79 | 349.26 | 1.01 |
| lambda_O_t[11] | 0.30 | 0.15 | 0.06 | 0.30 | 0.55 | 372.36 | 1.01 |
| lambda_O_t[12] | 0.05 | 0.15 | -0.21 | 0.06 | 0.29 | 519.89 | 1.01 |
| lambda_O_t[13] | 0.36 | 0.14 | 0.12 | 0.35 | 0.60 | 375.80 | 1.01 |
| lambda_O_t[14] | 0.75 | 0.13 | 0.53 | 0.75 | 0.96 | 300.59 | 1.01 |
| lambda_O_t[15] | -0.07 | 0.16 | -0.34 | -0.06 | 0.19 | 496.67 | 1.01 |
| lambda_O_t[16] | -0.06 | 0.17 | -0.34 | -0.06 | 0.20 | 458.31 | 1.01 |
| lambda_O_t[17] | 0.00 | 0.16 | -0.25 | 0.00 | 0.26 | 414.79 | 1.01 |
| lambda_O_t[18] | 0.10 | 0.15 | -0.16 | 0.10 | 0.35 | 445.25 | 1.01 |

The Rhat ($\hat{R}$) statistics estimates the ratio between across-chain and within chain variances. $\hat{R} \approx 1$ indicates the four Markov Chains mixed well and the simulation reached convergence. The effective sample size $n_{eff}$ measures how "informative" the obtained samples. The closer to the sampling iterations, the more representative those samples are. Conversely, large Rhat and small $n_{eff}$ combined indicates symptoms of inefficient and ineffective exploration of parameter space. From table 2 and 3, we note that all Rhats are close to 1 and $n_{eff}$s are large relative to the number of sampling iterations (500).

]

## Section Summary

*What are the key points the reader should take from this section?*

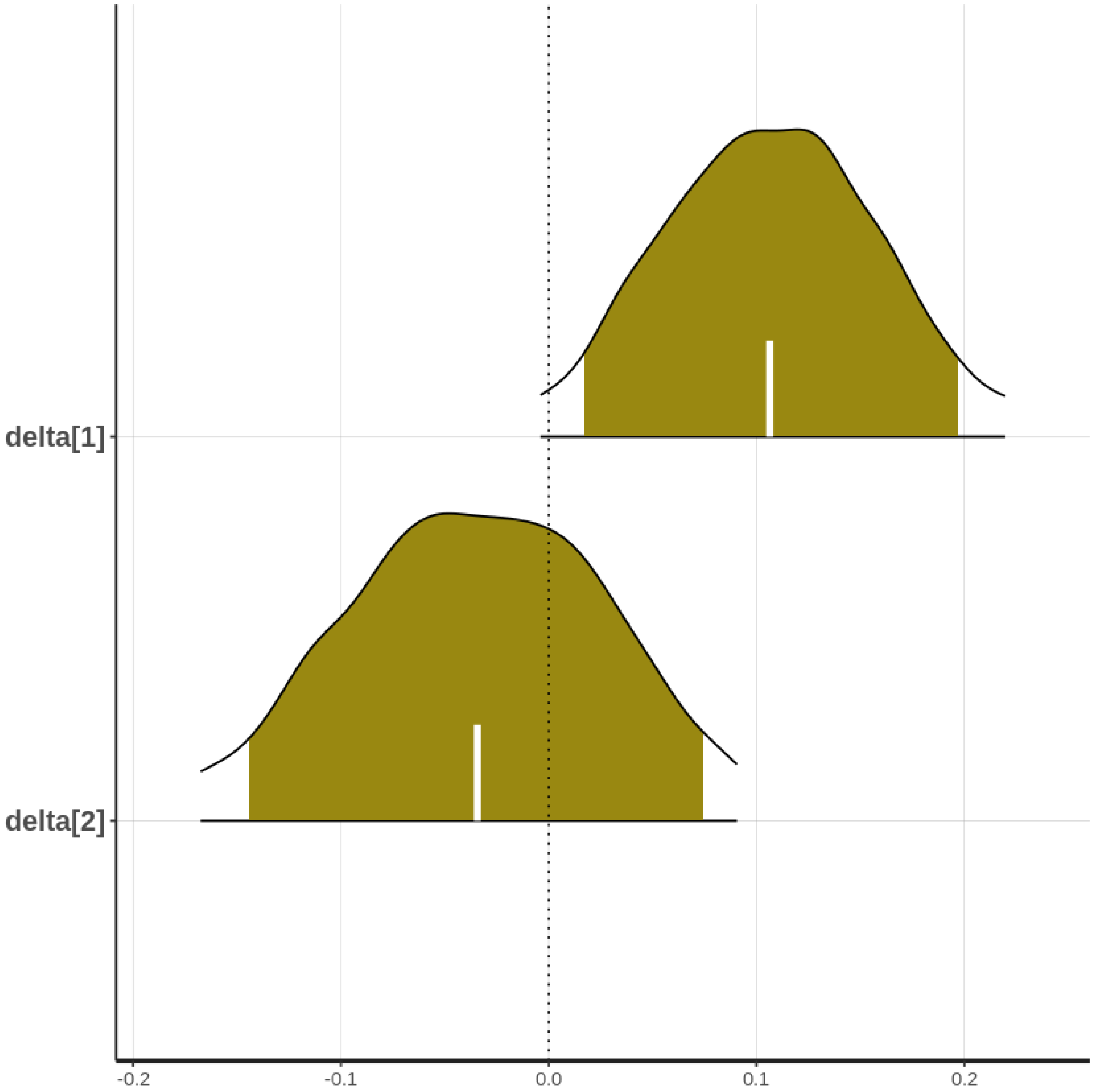
- *Warnings of divergent transitions after warmup iterations are serious threats to validity of inference.*
- *Rhat ≈ 1 and large* $n_{eff}$ *are indicators of convergence of Markov Chains.*
- *Revisiting model specification and tweaking parameters are normal practices.*

# **Practical Lessons Learned**

*This is perhaps the most important section of your research methods case study. Looking back, reflect on which aspects of your methodology went well, and which aspects did not go well. What would you do differently? What did you learn from the experience, and what advice do you have for readers planning their own research projects?*

[Now that we know our Poisson-logNormal model is trustworthy and can proceed to the next stage of Bayesian inference – checking utility of our model and evaluating posterior estimates. In figure 3, we plot the posterior density plot for HFA (delta). The parameter – delta[1] - represents the HFA for those regular games played before the outbreak. The parameter – delta[2] – represents HFA for those games played without spectators. The contrast is significant. When spectators were cheering in stadium, they overall contributed 0.1 goal per game. When they were absent, HFA disappeared and dropped to a statistical 0. At this point, we can safely conclude that social aggregation generates home field advantage.

Figure 3: Posterior Density Plot – HFA



]

## Section Summary

*What are the key points the reader should take from this section?*

- *Inference is drawn from posterior estimates of parameters*
- *Congregation of fans in stands gives rise to home field advantage*

# Conclusion

*Includes a round-up of the issues discussed in your case study. This should* not *be a discussion of conclusions drawn from the research findings, but should focus reflectively on the research methodology. Include just enough detail of your findings to enable the reader to understand how the method/approach you used could be utilized by others. Would you recommend using this method/approach or, on reflection, would you make difference choices in the future?* ***What can readers learn from your experience and apply to their own research?***

[In this case, we went through the typical "Box's loop" in performing probabilistic data analysis (Blei, 2014). First, we build our model by revising the existing *Poisson-LogNormal* model of Baio & Blangiardo (2010) to allow variable HFA – intercepts. Second, we use `rstan` and the core `stan` inference engine to fit our formulated model to the 2019-20 Bundesliga match data we collected. Third, we encountered the not so uncommon divergent transitions problem and had to revisit our model specification. When the convergence issues were addressed using modified hyperparameters, we proceed to posterior estimation of quantities of interest – HFA under different conditions. Finally, we close the loop and reached the conclusion that social congregation is the main factor giving rise to HFA. Even though our model can be easily extended to other soccer leagues and other sports, the most important lesson we learned is that modeling is an iterative process. As we continuously uncover shortcomings in our model specification, we build our way to a better representation of the data we collected and a better understanding of the phenomenon we investigate.]

---

# Discussion Questions

[Insert three to five discussion questions related to the methodology and practical considerations described in your case study]

*Discussion questions should be suitable for eliciting debate and critical thinking. The questions should encourage the reader to apply what they have learned beyond the context of the research project discussed. They should not test the reader's memory of specifics about the discussed project. Avoid questions which require only a single-word answer such as "yes" or "no."*

1. Can we use normal distribution instead of *Poisson* distribution in the model?
2. How can we model transitioning team capabilities due to injury and coaching change?

3. What is the differences between prior and posterior predictive checking in Bayesian inference?

---

# Multiple Choice Quiz Questions

*Multiple Choice Quiz Questions should:*

- *Test readers' understanding of your case study*
- *Not require any information that is not included in this case study*
- *Relate to research methodology, not the substantive research topic*
- *Not include 'all of the above', 'none of the above' or implausible distractors*
- *Cause the reader to identify the rationale behind the answer. For example:*

  *What was the method used to increase the reliability of this field observation study?*

  *A. Inter-coder reliability was calculated to ensure an acceptable Krippendorff's alpha.*

  *B. Constant comparison was used, whereby two coders visiting the same site simultaneously would conduct independent coding and reconvene to resolve any discrepant codes to produce a single set of codes for the observation. - CORRECT*

  *C. Researchers were asked to write about how their personal idiosyncrasies might have shaped the coding process, so these reflexive accounts can be used by the reader in assessing the study's reliability*

*Guidance for writing MCQs can be found here:*

- [Tips for Writing Effective multiple-choice questions]
- [The process of writing a multiple-choice question]

[Insert three to five multiple choice quiz questions below. Each question should have three possible answers (A, B, or C), with one correct answer. Please indicate the correct answer by writing CORRECT after the relevant answer.]

1. When a model fails, the modeler faces three options to address the failure

   i. Checking coding errors

   ii. Checking model specification

   iii. Changing tuning parameters

   The right choice of action(s) is:

A. iii only

B. i and ii only

C. ii and iii only

**D. i, ii, and iii**

2. What is the maximum number of code blocks in a *Stan* model?

**A. 5**

B. 7

C. 9

3. Which of the following values can’t be drawn from *Poissson* distribution?

A. 0

B. 1

**C. 2.2**

4. *Rhat* is the ratio between ___ and ____

A. Number of sampling iterations and warmup iterations

**B. Across chain and within chain variances**

C. Number of rows in data table and number of columns

5. Effective sample size $n_{eff}$ is relative to which of the following?

A. Number of data points in the dataset.

**B. Number of sampling iterations.**

C. Number of parameters to be estimated in the model.

D. Number of warmup iteration.

# Further Reading

Please ensure content is inclusive and represents diverse voices. In your references, further readings, and web resources you should aim to represent a diversity of people. We have a global readership, and we want readers of a wide range of perspectives to see themselves reflected in our pedagogical materials.

[Insert list of up to six further readings here, in APA Style.]

Jayal, A., McRobert, A., Oatley, G., & O’Donoghue, P. (2018). Sports Analytics: Analysis, Visualisation and Decision Making in Sports Performance (1st ed.). Routledge. https://doi.org/10.4324/9781315222783

Kim, C. J., & Nelson, C. R. (2017). *State-space models with REGIME Switching: Classical and gibbs-sampling approaches with applications*. MIT press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

# Web Resources

[Insert links to up to six relevant web resources here, in APA style]

Duan. C. J. (2022), Companion Jupyter Notebook to the Case of “Using Bayesian Inferential Approach to Analyse Home Field Advantage of 2020 Bundesliga Teams”, https://github.com/hublun/Stan/blob/main/Stan/Jupyter/Bundesliga_2020.ipynb

Katoch, A. (2018),  Hierarchical Bayesian Modeling of the English Premier League, https://vonarchimboldi.github.io/hierarchical-bayesian-model-epl/ .

Kharratzadeh, M. (2017), Hierarchical Bayesian Modeling of the English premier League, https://mc-stan.org/events/stancon2017-notebooks/stancon2017-kharratzadeh-epl.pdf

# References

[Insert bibliography of references cited in text here]

*References should conform to American Psychological Association (APA) style, 7th edition, and should contain the digital object identifier (DOI) where available. SAGE will not accept cases that are incorrectly referenced. Please ensure accuracy before submission. For help on reference styling see* https://apastyle.apa.org/style-grammar-guidelines.

Baio, Gianluca & Blangiardo, Marta. (2010). Bayesian hierarchical model for the prediction of football results. Journal of Applied Statistics. 37. 253-264. 10.1080/02664760802684177.

Blei, D. M. (2014). Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models. *Annual Review of Statistics and Its Application*, 1(1), 203 - 232. doi:10.1146/annurev-statistics-022513-115657

Courneya, K.S., & Carrón, A.V. (1992). The Home Advantage In Sport Competitions: A Literature Review. *Journal of Sport & Exercise Psychology*, *14*, 13-27.

Duan C. J. (Chaojie) & Ananyo Chakravarty (2021) Team Contingent or Sport Native? A Bayesian Analysis of Home Field Advantage in Professional Soccer, Journal of Business Analytics, 4:1, 67-75, DOI: 10.1080/2573234X.2020.1854625

Glickman E. Mark & Hal S. Stern (1998) A State-Space Model for National Football League Scores, Journal of the American Statistical Association, 93:441, 25-35, DOI: 10.1080/01621459.1998.10474084

Lopez, Michael & Matthews, Gregory & Baumer, Ben. (2017). How often does the best team win? A unified approach to understanding randomness in North American sport. The Annals of Applied Statistics. 12. 10.1214/18-AOAS1165.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna,

Austria. URL https://www.R-project.org/.Schwartz, B., & Barsky, S.F. (1977). The Home Advantage. *Social Forces*, *55*, 641-661.

Stan Development Team. (2022). Stan Modeling Language Users Guide and Reference Manual, 2.29. https://mc-stan.org

Stan Development Team. (2022). *RStan: the R interface to Stan*. http://mc-stan.org/

Taylor, M. (2021). Euro 2020: Will home field advantage matter? Retrieved 2021-06-12, from https://www.sportinglife.com/football/news/will-home-advantage-matter-at-euro-2020/191851