Bachelor thesis

Department of Statistics

Kandidatuppsats, Statistiska institutionen

Nr 2020:2

Assessing “big match” ability in professional football using Bayesian linear modelling

Bayesiansk modellering för att bedöma fotbollsspelares prestationsförmågor i matcher mot svårt motstånd

Pär Hammarström & Markus Bölske

Självständigt arbete 15 högskolepoäng inom Statistik III, VT2020

Supervisor: Parfait Munezero

Abstract

The term “big match player” is a well-established and often used term among both football professionals and fans, yet it is rather vaguely defined. It is a quality that we attribute to players that seem to inhabit a special ability to step up when their teams need it the most and succeeding against all qualities of opposition and levels of pressure.

In this thesis, we present a method and a metric based on Bayesian linear modelling for assessing football players and their abilities to perform under pressure posed by strong opposition.

The average performance score per minute played is selected as the response variable, and the ability to perform well against strong opposition is estimated using a covariate that describes the expected difficulty of the opposition in each match played.

We use a simple non-parametric frequentist method along with hierarchical and non-hierarchical Bayesian linear modelling to explore the problem.

A program based on the models was developed in the R programming language. The program uses a Gibbs sampler (a Markov Chain Monte Carlo algorithm) to estimate the parameters of the Bayesian models by sampling from their conditional distributions. Using the program, the models are applied to observed data from the 2019 season of Allsvenskan (the Swedish men’s premier league), but the models and the program are applicable for any equivalently structured dataset.

Based on the models, we introduce the PPS score as a novel metric for the “big match ability” by computing the probability space of the positive regression coefficient values. This PPS score is interpreted as the estimated probability of a player playing better against a "big" team than against a "small" team and is the main focus of our study.

Preface

This bachelor's thesis project was carried out at the department of Statistics at Stockholm University.

Special thanks go to our supervisor Parfait Munezero, doctoral student in Statistics at the department of Statistics at Stockholm University, for his patience, opinions and guidance throughout the project.

We would also like to thank David Sumpter, Professor at the department of Mathematics at Uppsala University and data scientist at Hammarby IF, for the idea for the thesis and his support along the way.

Stockholm, May 2020.

Pär Hammarström, Markus Bölske.

Table of contents

1 Introduction

1.1 Background

Whether we like it or not, modern professional football is a business in which money plays a bigger and bigger role, and transfer sums and salaries represent big parts of clubs' expenditures. This makes refined and accurate scouting tools increasingly important key elements of making informed recruitment decisions in order to find players with the maximum positive impact on club success.

Physical and technical abilities of football players can be estimated using models that assess player performances. One of those models is the Twelve[1] points system (see Appendix A), which is based on a machine learning algorithm that determines how much player actions during a match impact the probabilities of goals being scored and conceded. This algorithm was applied using data from several seasons of the top European football leagues.
Using this algorithm, the performance of players during individual games or over entire seasons can be ranked and compared based on different aspects of in-game actions to obtain quantitative evidence of strengths and weaknesses.

However, there are other more latent abilities that cannot be estimated merely by looking at key performance indicators, but rather how those indicators are affected by certain match contexts. These latent abilities can give additional clues of the true value of a player from a club's perspective.

We may for example want to assess and compare players based on their ability to perform under different types of pressure:

- Pre-match pressure defined by how much the outcome of a match will change the probabilities of seasonal success. A match during the later stages of a season may determine the home team's chances of avoiding relegation (high pre-game pressure), while the away team is stuck in the middle of the table with no incentive to win (low pre-game pressure.)
- In-game pressure defined by how much player actions will affect the probabilities of different match outcomes given the current match minute and goal difference. A 3-0 lead in the 75th minute is almost guaranteed to result in a win no matter how bad the team plays, while a score of 0-1 in the 75th minute gives the players more incentive to play well in order to avoid a defeat.
- Pressure defined by the level of opposition. Players may not have to play at their full potential in order to win against lesser teams. But their abilities to step up and perform when it matters the most are tested when playing against greater teams.
- Other factors, such as rivalry with an opposing team, big crowds, or pressure to break a recent trend of underperformances.

[1] https://twelve.football/analytics

This thesis focuses on modelling the performance ability of a player in matches against different levels of opposition.

1.2 Objective

The objective of this thesis is to design a model which provides a metric for assessing and comparing players based on their ability to perform against strong opposition. This is accomplished by creating and evaluating three different Bayesian linear models that describe players' expected performance scores per minute played in a match using the level of opposition as the regressor.

These models allow us to evaluate the probability that the effect of strong opposition is positive given historic observations from the players actions in different matches. We use this probability to describe how likely it is that a player performs better in matches against "big teams" than in matches against "small teams".

We use the Twelve points system and analytics tool for gathering observational data. The models are implemented in an R program, but will also be translated into Python and integrated with the Twelve system in order to be used in practical applications.

1.3 Possible applications

The model can improve the analysis toolkit used by football organizations and be used together with already established metrics and key performance indicators to make even more informed decisions:

- It can be used to assess, rank, and compare players by the ability of performing well against difficult opposition, based on which decisions regarding player recruitment can be made. The probability that the effect of strong opposition for a given player is positive can be translated into a performance indicator which may serve as a tool for ranking, comparing and assessing the player.
- It could be used to predict a player's performance in a match given a level of opposition. This type of prediction could be relevant when making decisions regarding the starting line-up and choosing substitutes.

Furthermore, the models built here and insights we obtain can be used as a start-off point for future research into a more general "big match player" ability.

1.4 Previous research

The topic of football analytics is a constantly growing field of study. Spatio-temporal event data is recorded in an increasing level of detail, and new tracking technology offer more data regarding off-the-ball aspects of the game that can be analysed and used for inference. In Beetz *et al.* (2009), a hierarchical model for automatic gathering and analysing of such data is presented, and the authors underline the importance of context-sensitive concepts.

Peralta (2019) analysed and presented a model for optimal positions during attacking situations using collective motion simulations in his dissertation, and was able to quantify pass success probabilities, pitch control and pass impact. This study lead to improvements of player assessments and was incorporated into the Twelve Player assessment model that we use for observational data in this thesis.

Bransen *et al.* (2019) analysed how mental pressure affects different aspects of players' performances. In this paper, mental pressure metrics are defined in terms of pre-game pressure and the in-game pressure. Pre-game mental pressure is defined by the team's ambition, match importance (implicit gain or loss of a potential match outcome), recent performance, and match context. Match context is defined by variables such as location (home/away), rivalrousness and match attendance. The model is then used to rank each match that a team plays.
In-game mental pressure is based on the probabilities of the different match outcomes, modelled by the time remaining and the probabilities of each team scoring. Player performances are measured much like using the Twelve Player Assessment model, and different actions are rated categorically in terms of contribution, decision and execution. All player actions were then ranked according to the combined pressure model, and metrics were aggregated per player. Using this model, the authors were able to show how high-mental pressure situations affect performances and how these insights can be used in strategic decision making.

In an article from Link and de Lorenzo (2018), an analysis of the effect of match importance on high intensity player activity was presented. Match importance is defined as the difference between the probability of a team achieving a season outcome given that they win their next match and the probability of the team achieving the same outcome given they lose their next match, a model of measurement that was first conceived by Bedford and Schembri (2006). The analysis found a significant effect of importance on player activity, but no significant difference in player activity was found between matches at the beginning and in the end of a season.

Egidi and Gabry (2018) analysed individual performances of football players using Bayesian hierarchical modelling. Three models were tested for predicting the performances of players in the Italian top-division and compared to the scores from a popular Italian football fantasy game called Fantacalcio[2]. The two best performing models incorporated missingness, where a variable denotes if a player is participating in a match or not. From these models four of the eleven top players in the league were successfully predicted.

We have found no previous studies that use Bayesian hierarchical modelling for the explicit purpose of assessing and presenting a metric for athletes' abilities of performing well against difficult opposition.

1.5 Delimitations

We are using an already established model for assessing player performances. We assume that this is the best assessment model available. If a better model for player

[2] https://www.fantacalcio.it/

performances is developed though, we can apply it in our models and use it to further improve our assessments.

The data upon which the study relies is from the 2019 season of Allsvenskan, but the models are generic and can be used with different datasets to analyse players in other leagues.

We will not address further analytical topics, such as the impact of having “big match” players in a team.

We focus on the match context defined by the level of opposition; no other aspects of mental pressure such as rivalry or match importance are included in our analysis.

The assessment of the level of opposition used in this thesis is based on overall team performance over an entire season. No other aspects that may be considered relevant to define the level of match difficulty such as current form or missing key players will be addressed in our model.

1.6 Structure of the thesis

In section 2 we define the dependent variable that represents player performance and the main covariate “level of opposition”. Here we also perform a preliminary analysis using a frequentist non-parametric test. In the next section (3) we describe the statistical methods used in the report and present some theory about Bayesian inference and how we will apply it to our models. These models are evaluated in section 4, where two different samples of players are compared using the models. In the last two sections the findings from the analysis are discussed (section 5) and what conclusions that could be drawn from those (section 6).

2 Data

The observational data used for this thesis consists of event scores from Twelve. The event scores are based on raw event data sourced from Opta[3] and covers every game in the Swedish top-division Allsvenskan in the season of 2019. The event data is processed by Twelve Player Performance Assessment system and each event is assigned a performance score (see Appendix A). The event scores are categorized and aggregated by player and match. Since the assessment model is designed for outfield players, goalkeepers have been removed from the data. Also, player performances based on less playtime than 15 minutes in a match have been excluded. This decision was motivated by the assumption that players need to play a certain number of minutes in order for the average points per minute played to be representative enough and to reduce the effects of physical fatigue, while keeping the sample size as large as possible. The cut-off point of 15 minutes can be adjusted in our code, but its relevancy was not further analysed in this thesis.

A full season of Allsvenskan consists of 240 matches. There were 16 teams playing in Allsvenskan 2019 with 373 unique outfield players (excluding players with playtime less than 15 minutes in all participating matches). Players can be substituted during the match, and therefore more than 10 outfield players can participate in each match for a team. The data contains 5760 event scores where each event represents a sum of scores for each participating player in every match.

2.1 Variables

2.1.1 Player score by match

The data set contains all the Twelve player assessment scores per match and player in different action categories. The sum of the categories “Defence”, “Attack” and “Shot” per minute played is used as an indicator of overall performance. The total score for each player and match is summarized to a new variable; the TotalScore.

Since the play time varies between matches and players, we calculate an average score per minute played (TotalScorePM) to represent the performance. This variable is more representative for the performance of a player in a specific match he played.

[3] https://www.optasports.com/sports/football/

Table 2.1 Description of total scores variables.

	TotalScorePM	*TotalScore*
Count	5760	5760
Mean	10.40	813.64
Standard deviation	7.63	582.76
25%	5.97	392.49
50%	9.02	726.35
75%	13.05	1112.05
Minimum	-54.86	-1408.01
Maximum	94.53	5360.59

Table 2.1 displays the statistical summary of the variable TotalScorePM. The mean value for the total score of a player in a match is 813.64, and the mean value for the total score per minute played in a match is 10.40.

Exploring the distribution of total scores per minute played in Figure 2.1, it is observed that the probability distribution tends to be normally distributed, but it is right skewed.

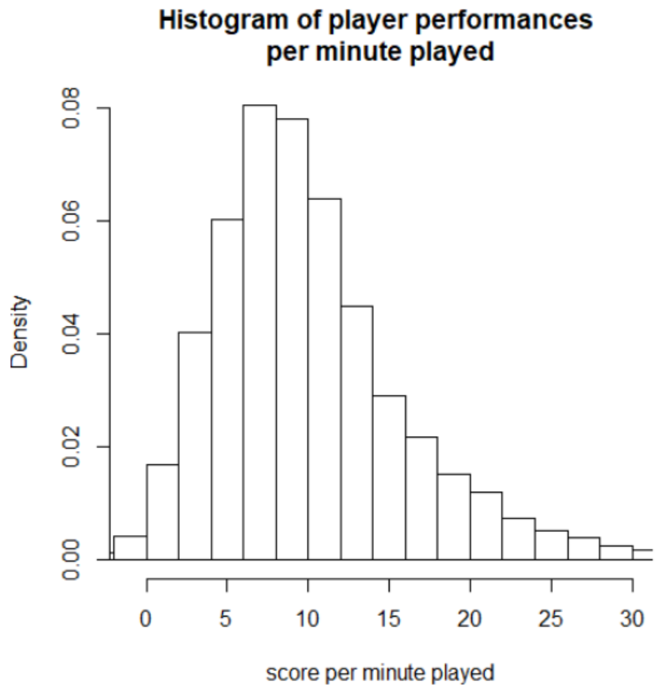

Figure 2.1 Histogram of player performances.

Figure 2.2 displays the distribution of the total score in the entire season per total time played. The symmetry of the distribution is more even as can be observed in the figure.

In Figure 2.1 the right tail of the distribution is longer than the left, but as can be observed in Figure 2.2, the tails could be considered more evenly distributed.

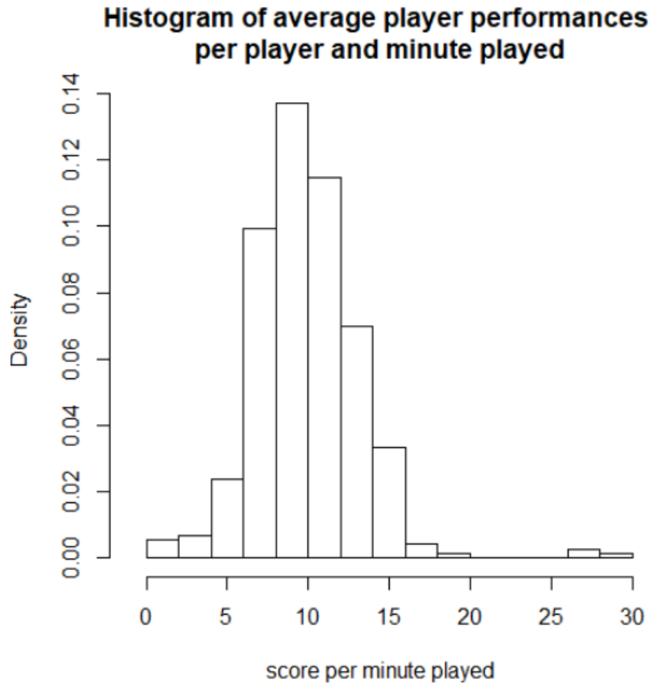

Figure 2.2 Histogram of average player performances.

2.1.2 Variables in the model

In Table 2.2 other variables used in the data analysis are listed. The variable BigMatch is generated for the analysis and more information can be seen in section 3.1.1.

Table 2.2 Variables used in the model.

Variable name	*Description*	*Type*
PlayerId	The ID for the player.	Integer
BigTeam	Indicator of whether the player's team is categorised as a "big team".	Binary
MatchId	The ID for the match.	Integer
BigMatch	Indicator of whether the opponent team is categorised as a "big team".	Binary/Non-binary Integer

The BigMatch variable defines the level of opposition by indicating whether the opponent team is defined as a "big team". We defined it based on the final standing of a team in the league. A team is categorised as a "big team" if it reached a resulted in the top 7 teams (based on K-means clustering, see 3.1.1).

2.2 Preliminary analysis

As a first preliminary analysis, we perform a non-parametric test to examine the occurrence of "big game" players and "small game" players. In this analysis we use the Mann-Whitney U test (or Wilcoxon rank-sum test), which is a statistical non-parametric test for comparing two independent samples.

The statistic U is calculated by ordering all the observations and assigning each a rank according to its magnitude. Counting the number of observations in one sample that tops observations in the other sample, the sum is the statistic U. The null hypothesis is that the populations have identical distributions and the possible alternative hypotheses are that the distributions of the populations have different locations in general, or that that one specific population distribution is shifted to the right or to the left.

For every player, we are comparing the distributions of performance scores against big teams with the distributions of performance scores against small teams and look for significant location differences.
The tests return p-values that represent the probability of the difference in the samples given that the null hypothesis (that there is no difference in distribution between the samples) is true.
The tests also return pseudo median differences between the observations in the two samples (Hodges–Lehmann estimators), which can be used as measures of how much the performances against the two categories of opposition differ. (Hodges & Lehmann, 1963)

For each player, we ran two one-sided rank-sum tests to test for positive location difference and negative location difference at 10% significance level. We then identified and recorded the number of significant subjects: the number of players with a significant positive location difference were counted as "big game players" and the players with a significant negative location difference were counted as "small game players".

Table 2.3 Players with at least six matches played whose rank-sum test result was significant for either positive location difference ("Big match players") or negative location difference ("Small game players").

	"Big game players"	*"Small game players"*	*Undefined players*	*Total*
Count	9	89	208	*306*
Percent	3%	29%	68%	*100%*

The result of the analysis shows that there indeed are players that perform significantly better and players that perform significantly worse against difficult opposition, which indicates that there is value in continuing the study and approach the topic using a more sophisticated model. The lowest number of matches played in any of the groups of players that tested significantly (either with a positive location difference or a negative location difference) was six. Anecdotally, this is a number of matches that often seem to be used as a minimum for evaluating average performances of players to be considered representative enough, and our test procedure seems to be corroborating its relevancy.

3 Methodology

3.1 The model for the total score per minute

In our analysis we can think of the player performance to be nested with a factor indicating the level of opposition. We, therefore, define the variable Y to represent the "score per minute played" and we assume that it is related to the covariate X (which in this thesis represents the level of opposition) as

$$y_{i,p} = \alpha_{0p} + \alpha_{1p} \cdot x_i + e_{i,p}, \quad e_{i,p} \sim N(0, \sigma_p) \tag{3.3}$$

for $i = 1, \cdots n$, where n is the total number of matches played by the player p and $p = 1, \ldots,$ denotes the player ID.

In the model (3.3), α_{0p} is the player intercept, which is interpreted as the score for each player without the effect of other factors. The parameter α_{1p} is the effect of the covariate X; that is, the effect of playing against a "big team" on the total score of the player. Finally, $e_{i,p}$ is the error term.

We want to measure the effect of playing in big matches on the total score, and one way of accomplishing this is to create a model that models each player's performance (in performance score per minutes played) as a function of his own intercept and slope. Also, we want to add an effect of the players' own team level, which is represented by the covariate Z. Therefore, the regression parameters in (3.3) are modelled as

$$\alpha_{0p} = \beta_{00} + \beta_{01} \cdot z_p + u_p, u_p \sim N(0, \tau^2_{\alpha_0}) \tag{3.4}$$

$$\alpha_{1p} = \beta_{10} + \beta_{11} \cdot z_p + v_p, v_p \sim N(0, \tau^2_{\alpha_1}) \tag{3.5}$$

β_{00} and β_{01} denote the grand mean and the team level effect of the player performance score per minute and u_p is the individual adjustments to the grand mean. β_{10} and β_{11} denote the main effect of level of opposition and team level, with v_p as an individual adjustment to that effect.

The model proposed in (3.3), (3.4) and (3.5) is a hierarchical model. Hierarchical models include parameters that depend on values of other unobserved parameters commonly known as hyperparameters such as the ones defined in (3.4) and (3.5). Examples of hierarchical models can be found in Gelman *et al*. (2013).

The hierarchical model described by (3.3), (3.4) and (3.5) is very flexible and builds on the notion that players in the same league and in the same team share some baseline level of the ability we are analysing, thus brining the posteriors closer to the "true" values. One potential downside of this hierarchical model is overshrinkage, which makes it interesting to compare it to simpler models that could let the data speak more for itself and provide posteriors that better reflect the actual observations. One way to simplify this model is to remove the team level effect Z or even remove that level altogether making the regression model non-hierarchical. The latter model allows us to assess and compare players from the same league (or from leagues at equal levels) with no dependence to the players included in the dataset, which is a big advantage for practical applications. We will however fit and evaluate the other models (see section 4) to see if they would have better explanatory or predictive properties.

We will henceforth refer to these three models as Model 1 (the full hierarchical model), Model 2 (the reduced hierarchical model, i.e. the model without Z) and Model 3 (the non-hierarchical model).

3.1.1 Level of opposition variable

We fit the three models using two different definitions of the covariate X (level of opposition). We have chosen to evaluate two approaches to define the covariate, either as a binary variable or as a discrete variable.

To categorize the teams to define the binary variable, the clustering algorithm called k-means is used. Clustering can be used when observations with a similar set of characteristics can be combined into groups or clusters. The observations within a cluster should also be homogenous, but different from observations in other clusters.

K-means is a partitional clustering algorithm that belongs to the non-hierarchical clustering methods group. Initially, K representative points are chosen for the starting centroids of the clusters. The starting points are chosen at random within the observations. Every observation is then assigned to the closest centroid. In general, Euclidean distance is used as measurement for the distance from the centroids to each observation. When all observations have been assigned to a cluster, the centroid for each cluster is updated. The algorithm then repeats these two steps and stops when the centroids do not change or if another pre-defined convergence criterion is met. See Aggarwal and Reddy (2014) for more details.

In our study, we are using k-means clustering ($K = 2$) with the points in the final table to cluster the teams into groups of "big teams" ($n = 7$) and "small teams" ($n = 9$) and use these groups to define the level of opposition. This binary categorization could be made using other methods, but this approach was deemed suitable for the purposes of this study and divides the matches into 105 "big matches" and 135 "small matches". A binary definition of the level of opposition makes the covariate easily interpretable, as we can formulate it as being the effect on performance of playing against a "big team" compared to playing against a "small team".

We also want to evaluate how well a non-binary definition of the level of opposition fits in the model. For this purpose, we use a discrete variable defined by the opposing teams' number of points in the end-of-season table. This definition was chosen because it should better capture the range of opposition and accommodates the arguable difference between playing against mid-table teams and teams in the two extremes of the table.

However, interpreting the non-binary definition of the level of opposition covariate is not as intuitive as its binary counterpart, but could roughly be formulated as the effect on performance per one-point increase of the opposing teams' end-of-season total.

3.2 Bayesian Inference

Assuming that θ represents all parameters in the model, for instance the regression parameters in (3.3), then in Bayesian inference θ is treated an unknown random variable which is given a distribution $p(\theta|\phi)$ known as the prior distribution. The parameter ϕ is known as the hyperparameter which is usually predefined. A prior distribution $p(\theta|\phi)$ represents the distributional beliefs about a variable before new evidence from the data is considered.

In the Bayesian framework, the parameter inference is done on the posterior distribution, which as described by Gelman *et al.* (2013) is the probabilities of θ given the data y. When applying Bayes' rule (see Gelman *et al.* 2013), the posterior distribution is computed by multiplying the prior density of θ with the likelihood function of y given θ and dividing by the density of y

$$p(\theta|y) = \frac{p(y|\theta)p(\theta|\phi)}{p(y)} \tag{3.1}$$

However, the density of the formula above is always a constant which allows us to write the posterior as

$$p(\theta|y) \propto p(y|\theta)p(\theta|\phi) \tag{3.2}$$

As a football-related example, we might imagine that we know that a goal was scored in a match and we know that it was either scored from a penalty or a corner. We know that the mean probability of scoring from a penalty is 75% and the mean probability of scoring from a corner is 5%. With no additional information, we might be tempted to assume it was scored from a penalty since corners rarely result in goals. But if we add prior information which states that there is an average of 20 corners per match and an average of 0.5 penalties per match, we would have to change our opinion as it would now seem much more likely that the goal was scored from a corner given the probability distributions of the event itself being either a penalty or a corner.

For the model without hierarchies (model 3), the posterior is specified according to (3.2) and the value used for the hyperparameters are specified in the following subsection. However, in the hierarchical models (model 1 and 2) the hyperparameters ϕ are also treated as random variable following a distribution $p(\phi)$. Therefore, the posterior for the joint distribution of the parameters in the hierarchical model, is expressed as

$$p(\theta, \phi|y) \propto p(y|\theta)p(\theta|\phi)p(\phi) \tag{3.3}$$

More details about the prior distributions are provided in the following subsection.

3.2.1 Priors

The prior distribution should include all plausible values of θ, but it does not have to be a perfect representation of the true value. With enough observational data, the prior probability distribution specification will be outweighed by the information in the data. If there are few observations however, the posterior tends to be more influenced by the prior.
This is called shrinkage and means that the high-level parameters can affect the low-level parameters by pulling the estimates closer together and therefore reducing the variance (Kruschke, 2015). But it can also lead to the phenomena of overshrinkage, where extreme values are drawn towards the grand mean and cause problems for the estimation of players with few observations/matches played (Baio & Blangiardo, 2010).

There are different types of prior distributions. One class of priors commonly used is the conjugate prior, which is a prior that follows the same parametric form as the posterior distribution. This makes it simple computationally and easy to interpret.
A prior can be informative, which means that it aims to fully capture previous scientific knowledge about the parameter, for example if we use the posterior results from all previous experiments as a prior in a new repetition of the experiment.
In some cases, we may want to use a small amount of real-world information, enough to simplify computations and to keep the posterior within reasonable bounds, but not restraining it too tight around uncertain prior information. In this case we say that the prior is weakly informative.
If we have no knowledge at all about a prior distribution, or if we want to let the data "speak for itself" with no prior information, we can construct a "non-informative" prior that will put minimal influence on the posterior distribution. (Gelman *et al.*, 2013)

Our dataset consists of observations of player performances. Different players played different numbers of matches (1-30), which means that the posteriors of some players will be more influenced by a prior than others. We can however counter this by considering the data as hierarchically structured: each player belongs to a team, and the teams can be grouped by their general level of performance. The priors on the "team level" effect are viewed as a sample from a common population distribution (the entire league).
In our study, we will only analyse one league at a time, but we could include data covering multiple leagues and add the league as a level in the hierarchy.

In both the hierarchical models and the normal regression model we use individual Gaussian priors for the regression coefficient and Gamma priors for the precision ($1/\sigma^2$). We use non-informative priors since we have no previous information about these parameters. The prior for the intercept in the non-hierarchical model is a Gaussian distribution as well.
In the hierarchical model we use a Gaussian prior distribution for the intercept hyperprior. This prior is conjugate, since the posterior distribution is also Gaussian (Murphy, 2017). More specifically we assume the following priors

$$\beta_{00} \sim N(0, 1/t_1)$$

$\beta_{01} \sim N(0, 1/t_2)$

$\beta_{10} \sim N(0, 1/t_3)$

$\beta_{11} \sim N(0, 1/t_4)$

$\tau_{\alpha_0}^{-2} \sim Gamma(a, b)$

$\tau_{\alpha_1}^{-2} \sim Gamma(a, b)$

${\sigma^{-2}}_p \sim Gamma(c, d)$

However, in the non-hierarchical model, where the grand mean for the intercept also has been omitted, we assume the following priors

$\alpha_{0p} \sim N(0, 1/t_1)$

$\alpha_{1p} \sim N(0, 1/t_2)$

${\sigma^{-2}}_p \sim Gamma(a, b)$

To make these priors noninformative we set the precisions $t_k = 0.001$ for $k = 1, \dots, 4$. The same value for the Gamma distributions is used; i.e. a, b, c, d is set to 0.001.

3.2.2 Model fitting with Jags

JAGS[4] is a program for analysing Bayesian models using Markov Chain Monte Carlo (MCMC) and is designed to work with the R language environment. It uses a Gibbs sampling algorithm (Gelman *et al.* 2013, chapter 11.1) to draw samples from the joint posterior distribution by cycling randomly through its conditional distributions. We define the priors of the model parameters, and when we run the model we generate samples from the posterior distribution of the model parameters. This generates a sample from the posterior distribution of the variable of interest (Plummer, 2003).

The model definition of the complex hierarchical model can be seen in Figure 3.1, and the others in Appendix C.1.

We use the "autorun.jags" function from the "runJags" package[5] for running the JAGS model (with the following settings: n.chains=3, thin=2, method="rjags"). We make sure all parameters converge appropriately by controlling the Gelman Rubin statistic (Gelman & Rubin, 1992), which analyses the difference between multiple Markov chains. We also use the Raftery and Lewis (Raftery *et al.* 1992) diagnostic to ensure that we run enough iterations for the desired level of precision in the posterior samples.

[4] https://sourceforge.net/projects/mcmc-jags/

[5] https://cran.r-project.org/web/packages/runjags/index.html

The models are evaluated and compared using Deviance Information Criteria (Spiegelhalter *et al.* 2002).

We use the "coda" package[6] for analysing the output, and modified functions from the "DBDA2E-utilities" package by John Kruschke (Kruschke, 2015) to visualize the posteriors.

The full R-script along with model diagnostics can be seen in **Appendix C: Code**.

```
model {
    # Likelihood:
    for ( i in 1:N ) {
      y[i] ~ dnorm(b0[player[i]] + b1[player[i]] * x[i], tau.p[player[i]])
    }
    # Player priors:
    for ( p in 1:Nplayers ) {
      b0[p] ~ dnorm(b0.tl[teamLevel[p]], tau.b0.tl[teamLevel[p]])
      b1[p] ~ dnorm(b1.tl[teamLevel[p]], tau.b1.tl[teamLevel[p]])
      tau.p[p] ~ dgamma(0.001, 0.001)
    }
    # Team level priors:
    for ( tl in 1:2) {
      b0.tl[tl] ~ dnorm(b0.l, tau.b0.l)
      b1.tl[tl] ~ dnorm(0, 0.001)
      tau.b0.tl[tl] ~ dgamma(0.001, 0.001)
      tau.b1.tl[tl] ~ dgamma(0.001, 0.001)
    }
    # League priors:
    tau.b0.l ~ dgamma(0.001, 0.001)
    b0.l ~ dnorm(0, 0.001)
}
```

Figure 3.1 JAGS model definition for model 1. Note that JAGS use precision ($1/\sigma^2$) instead of variance.

3.3 A metric for comparing players

The goal of this study is to find a way to assess and compare the ability to play against difficult opposition between players. The posterior distribution of the intercept indicates the level of performance with no covariate, i.e. in matches against "small teams". While this is highly important when assessing the general level of a player and to predict the actual number of points scored against a certain level of opposition, we are focusing on the relative effect of playing against difficult opposition indicated by the slope coefficient of each player.

We are interested in three elements of the posterior distribution of the slope coefficient:

- Central tendency. The mean of the coefficient indicates the average effect of the covariate. This can be interpreted as the general effect of playing against difficult opposition. It will be used to compare the results of the binary covariate

[6] https://cran.r-project.org/web/packages/coda/index.html

model with the results of the two-group non-parametric tests performed in the preliminary analysis using the pseudo means of the players that showed significant location differences.

- Credible interval. A 90% credible interval indicates the uncertainty of the average effect. The range of the 90% credible interval can be interpreted as the level of consistency in performance against different levels of opposition and the level of uncertainty of the mean effect.
- Positive probability space (henceforth referred to as "PPS"). By measuring the probability space of all positive values in the posterior distribution (the area above zero in the posterior probability space) of the X covariate regression coefficient (the level of opposition), we get an indication of how likely it is that a player will play better against "big teams" than against "small teams". An example of the PPS is provided in Figure 3.2**.**

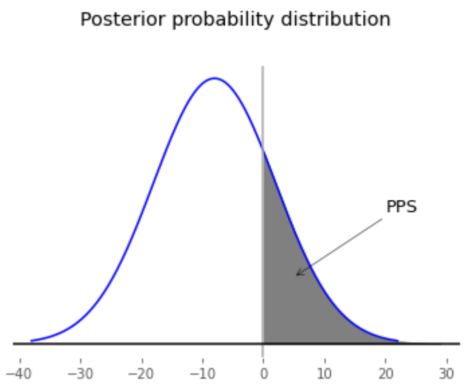

Figure 3.2 Plot of the posterior distribution with PPS as the shaded area. The PPS is the probability space above zero in the posterior probability distribution.

These three elements of the posterior distribution all add to the assessment of the "big match ability" of a player, but the one we find most interesting is the PPS. Not only is it suitable given the Bayesian approach, but it also gives a simple answer to a simple question: how probable is it that a certain player performs better against a greater team? The answer to this question can be used as a simple and understandable indication of the nature of the player's abilities, and to refine player assessments based on other expected performance indicators. We will therefore focus our study on this metric and use it to define the "big match effect".

The PPS metric will be evaluated to see if it is a suitable metric to use for assessing the "big match" ability of players. We will also be evaluating the posterior distributions as means for comparing pairs of players' abilities to perform against "big teams".

4 Results

The players we will use as examples in this section are the top ten outfield players in Allsvenskan 2019 as voted by the players themselves[7]. This list consists of three defenders, four midfielders and three attackers, representing the top four teams in the final league table. We chose these players instead of a random sample in order to make the results more relevant and interesting for readers with an interest in Swedish football. In Section 4.3 we rank the top players based on their PPS scores, which includes some lesser known players. In Section 4.4, we study the players we found to have significant positive locations differences using the rank-sum tests during our preliminary analysis and compare them to the result of the Bayesian modelling approach to see if we reach similar conclusions.

The aim of this thesis is to analyse player performances in order to find a method and a metric for assessing the ability to perform against difficult opposition. We do this by using the R programming environment for Bayesian inference and by comparing the results of the different models and the test results retrieved during the initial preliminary analysis using non-parametric frequentist methods.

4.1 Models evaluation

To evaluate and compare the models we will compare the Deviance Information Criteria. As mentioned in 3.2.2., Spiegelhalter *et al.* (2002) describes how DIC can be seen as a Bayesian variant of Akaike Information Criterion (AIC). A smaller DIC-value indicates a better fitting model.

The models to evaluate are summarized in Table 4.1, where the complexity of each model can be seen.

Table 4.1 Model summaries. Where "1" represents the player level, "2" represents the team level.

	Model 1	Model 2	Model 3
0	$y_{i,p} \sim N(\alpha_{0p} + \alpha_{1p} \cdot x_i, \sigma_p)$	$y_{i,p} \sim N(\alpha_{0p} + \alpha_{1p} \cdot x_i, \sigma_p)$	$y_{i,p} \sim N(\alpha_{0p} + \alpha_{1p} \cdot x_i, \sigma_p)$
1	$\alpha_{0p} \sim N(\beta_{00} + \beta_{01} \cdot z_p, \tau^2_{\alpha_0})$		
2	$\alpha_{1p} \sim N(\beta_{10} + \beta_{11} \cdot z_p, \tau^2_{\alpha_1})$	$\alpha_{0p} \sim N(\beta_{00}, \tau^2_{\alpha_0})$	

The three models were assessed by using both a binary definition ("big team opponent", "small team opponent") and a non-binary definition (the accumulated number of points by the opposing team at the end of the season) for the X covariate (as according to the "Covariate" column in Table 4.2).

[7] https://www.svenskelitfotboll.se/spelarforeningen-utser-2019-ars-basta-spelare-och-lag/

Both hierarchical models and the non-hierarchical model have better DIC-values when using a binary covariate. Based on this observation, we reject the non-binary covariate definition and focus our study on the binary definition of the "level of opposition" covariate. The non-hierarchical model with the binary covariate appears to have the best fit, but we will have a closer look at the posterior distributions of the models using the binary covariate to decide which model we should choose.

Table 4.2 Bayesian model comparison using the Deviance Information Criteria.

Model	*Covariate*	*DIC*
Model 1.1	Binary	36554
Model 1.2	Non-binary	36566
Model 2.1	Binary	36499
Model 2.2	Non-binary	36656
Model 3.1	Binary	36460
Model 3.2	Non-binary	36631

4.2 Models summary

The player sample we will study in this section consists of the ten top outfield players in the 2019 season of Allsvenskan, selected by the player association[7], but let us have a closer look at their posterior distributions and see if their statuses as "top players" are backed up by the data and our definition of the "big match ability".

Table 4.3 and Table 4.4 show summary statistics for the posterior distributions of the coefficients for the posterior distributions of the top ten players in each model. There is representation from the four top teams in the final league table (Djurgårdens IF, Malmö FF, Hammarby IF and AIK) and players with different positions are included.

Table 4.3 Intercept comparisons between the hierarchical and non-hierarchical regression models. Players are sorted by name.

	Intercept (α_{0p})					
	Model 1.1		Model 2.1		Model 3.1	
PlayerName	*mean*	*sd*	*mean*	*sd*	*mean*	*sd*
Anders Christiansen	14.89	1.86	15.49	2.16	21.94	2.63
Darijan Bojanic	15.72	1.36	15.39	1.63	18.14	1.86
Fouad Bachirou	11.06	1.12	12.07	1.40	12.73	1.69
Marcus Danielson	14.96	1.15	15.28	1.37	17.01	1.50
Mohamed Buya Turay	11.95	1.25	12.35	1.54	13.42	1.97
Muamer Tankovic	14.67	1.42	16.30	1.68	19.52	1.81
Nikola Djurdjic	12.95	1.59	14.37	1.93	18.39	2.53
Per Karlsson	10.16	0.56	10.12	0.68	10.09	0.71
Rasmus Bengtsson	12.75	1.14	12.14	1.45	12.92	1.78
Sebastian Larsson	13.80	1.23	12.69	1.45	13.69	1.74

According to Table 4.3 we can see that the player with the highest intercept mean are different in the models. Darijan Bojanic in model 1.1, Muamer Tankovic in model 2.1 and Anders Christiansen in model 3.1. These players along with Markus Danielsson seem to have the most consistent high intercept means across the models, indicating that they are the players among the top ten who performed the best against “small teams”. The player in the other end of the spectrum with the lowest intercept mean, Per Karlsson, also has the lowest variance, indicating that he is the most consistent player against “small teams”.
The means and standard deviations of the intercept are generally lower for these players in model 1.1 compared to model 2.1, which in turn are lower compared to model 3.1. This is likely due to shrinkage towards the population mean in model 2.1 and the population- and team level group means in model 1.1.

Table 4.4 Regression coefficient comparisons between the hierarchical and non-hierarchical regression models. Players are sorted by name.

	”Level of opposition” regression coefficient (α_{1p})					
	Model 1.1		Model 2.1		Model 3.1	
PlayerName	*mean*	*sd*	*mean*	*sd*	*mean*	*sd*
Anders Christiansen	-1.36	0.73	0.93	4.38	-5.43	4.28
Darijan Bojanic	-1.22	0.67	1.08	2.89	-1.67	2.96
Fouad Bachirou	-1.58	0.69	-4.07	2.24	-4.74	2.45
Marcus Danielson	-1.38	0.69	-1.12	2.41	-2.83	2.46
Mohamed Buya Turay	-1.47	0.7	-1.93	2.79	-2.98	3.05
Muamer Tankovic	-1.54	0.71	-5.82	3.08	-9.02	3.04
Nikola Djurdjic	-1.54	0.73	-4.87	3.95	-8.9	4.15
Per Karlsson	-1.39	0.56	-1.25	1.06	-1.22	1.09
Rasmus Bengtsson	-1.21	0.65	0.8	2.42	0.05	2.65
Sebastian Larsson	-1.07	0.72	3.47	2.6	2.46	2.78

According to Table 4.4, the player with the highest “big match” regression coefficient (α_{1p}) mean in both models is Sebastian Larsson. The high “big match effect” indicates that he is the player in the sample that improves his performances the most against “big teams” compared to his performances against “small teams”. This is also indicated by his PPS score in Table 4.5, which shows that the probability of him playing better against “big teams” is 8% based on model 1.1, 91% based on model 2.1 and 82% based on model 3.1.
Muamer Tankovic, who has the highest intercept mean, also has the lowest regression coefficient mean and PPS score. This indicates that he is very unlikely to play better against “big teams” compared to “small teams. Given his high consistent intercept mean, this could be an indication of overperformance against “small” teams rather than underperformance against “big” teams. The estimations are considerably lower, but also more precise in model 1.1 compared to the other models. This can be explained by the

nature of the hierarchy of model 1.1, where we have used informative priors for the “level of opposition” regression coefficient based on the “team level” group means.

The positive probability space indicating the probability of a player performing better against a “big team” (the PPS score) is also affected by the shrinkage of the intercept. All the players in the top ten have higher PPS scores in model 3.1 compared to model 2.1, and all the players except for Per Karlsson have higher PPS scores in model 2.1 compared to model 3.1.

Table 4.5 PPS scores. Players are sorted by name.

	PPS scores		
	Model 1.1	Model 2.1	Model 3.1
PlayerName	*PPS*	*PPS*	*PPS*
Anders Christiansen	0.0400	0.5779	0.0973
Darijan Bojanic	0.0478	0.6420	0.2827
Fouad Bachirou	0.0091	0.0369	0.0279
Marcus Danielson	0.0307	0.3131	0.1205
Mohamed Buya Turay	0.0243	0.2400	0.1580
Muamer Tankovic	0.0124	0.0341	0.0023
Nikola Djurdjic	0.0199	0.1092	0.0182
Per Karlsson	0.0115	0.1154	0.1266
Rasmus Bengtsson	0.0447	0.6303	0.5105
Sebastian Larsson	0.0796	0.9106	0.8183

In Figure 4.**1**.1 we can see the posterior distributions visualized and better understand the PPS score and the differences between players and models. A divider at value zero has been added to the posterior plots in order to simplify the interpretation of the PPS. The PPS-scores (the probability space on the positive side of the posterior plots) in model 1 are lower compared to the others, but the “big match effect” estimates are also much more precise.

It is worth noting that even though the PPS scores differ significantly between the models, Figure 4.1 shows that the midfielders retain the same rank relative to each other in all three models. In Table 4.5 we can see that among the attackers in the sample, Buya Turay have the highest PPS scores (0.0243; 0.2400; 0.1580), followed by Djurdjic (0.0199; 0.1092; 0.0182) and Tankovic with the smallest PPS scores (0.0124; 0.0341; 0.0023). The same tendency can be observed comparing the defenders in the sample (Danielsson, Karlsson, Bengtsson).

The 90% credible interval in the posterior plots in Figure 4.1 visualizes the difference in consistency of the “big match effect” and how it affects the PPS. Christiansen has a lower “big match effect” than Bachirou, yet his PPS is greater due to his larger credible interval range.

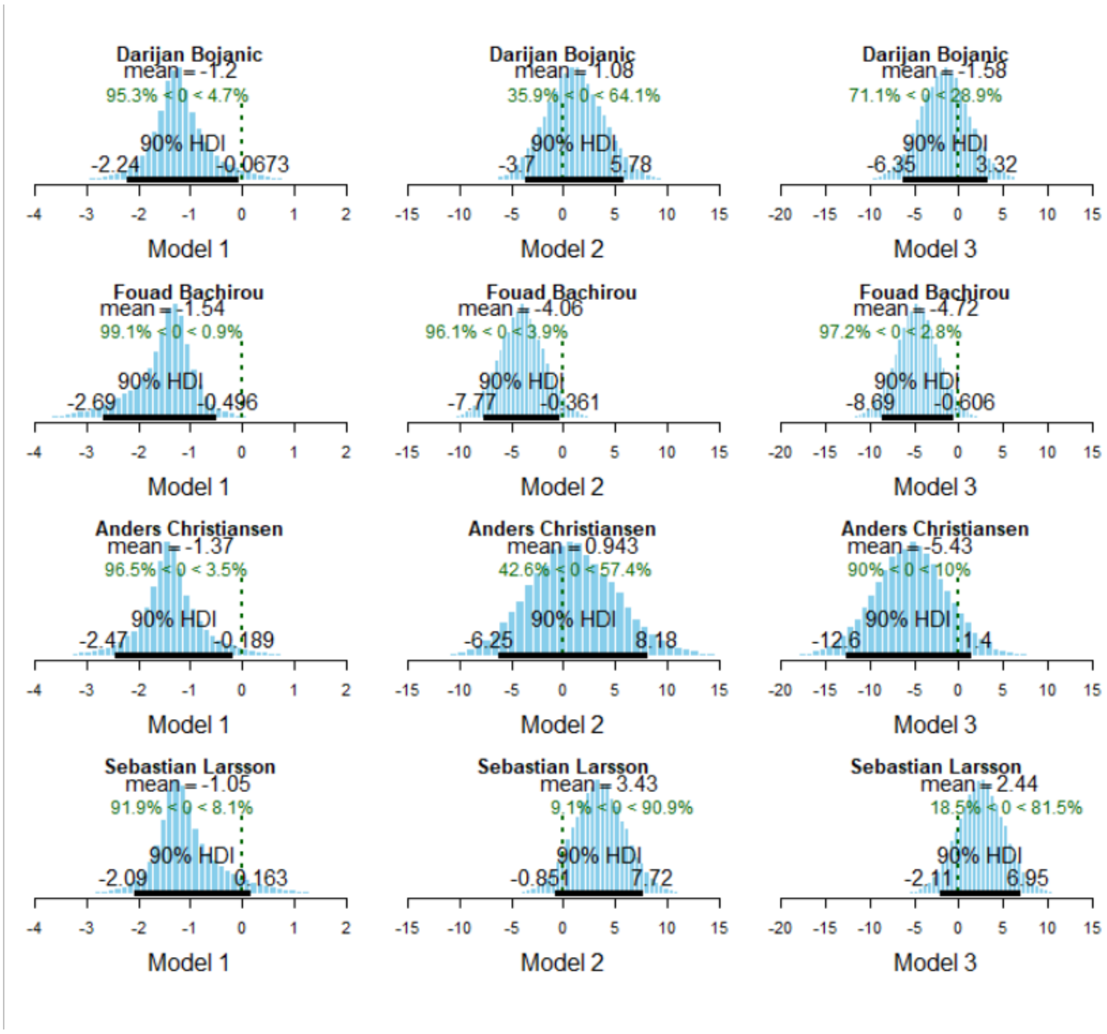

Figure 4.1 Posterior distributions of the "big match effect" for the top four midfielders using model 1, model 2 and model 3 with the binary definition of the “level of opposition” covariate.

4.3 Top players based on PPS scores

In this section we will rank the players from the entire league based on their PPS scores using the three different models and display the top 10.

Table 4.6 Top players based on the probability of performing better against "big teams" compared to "small teams".

Top PPS scores using the different models								
Model 1.1			Model 2.1			Model 3.1		
#	*Name*	*PPS*	*#*	*Name*	*PPS*	*#*	*Name*	*PPS*
1	Tobias Karlsson	0.16	1	Erik Berg	0.99	1	José León	0.99
2	Tom Pettersson	0.14	2	José León	0.98	2	Erik Berg	0.98
3	José León	0.14	3	Omar Eddahri	0.95	3	Jordan Attah Kadiri	0.97
4	Behrang Safari	0.13	4	Tobias Karlsson	0.94	4	Tobias Karlsson	0.96
5	Yaser Kasim	0.11	5	Tom Pettersson	0.93	5	Omar Eddahri	0.93
6	Kasper Larsen	0.10	6	Behrang Safari	0.92	6	Behrang Safari	0.93
7	Christoffer Carlsson	0.10	7	Kristijan Miljevic	0.92	7	Christoffer Carlsson	0.92
8	Filip Dagerstål	0.10	8	Sebastian Larsson	0.91	8	Yaser Kasim	0.92
9	Karol Mets	0.10	9	Yaser Kasim	0.89	9	Tom Pettersson	0.92
10	Erik Berg	0.10	10	Alex Dyer	0.88	10	Johan Blomberg	0.90

From Table 4.6 we can see that while the actual PPS scores are different, some of the players consistently place in the top ten of all the models. For example, José León is top 3 for all the models, while Omar Eddahri is top 5 in model 2.1 and model 3.1, but not top 10 for model 1.1.

Model 1.1 differs from the other two models in that the maximum PPS-values are much lower than for the other two models. In model 2.1 and 3.1 the top players have a PPS of 0.99, which would suggest that they almost always will perform better against a “big team”.

4.4 Non-parametric comparison

In the Mann-Whitney rank-sum test, we used a significance level of 0.9. That resulted in 9 players which demonstrated a significant difference of performance against “big teams”. The R-code is presented in Appendix C.2. From the non-hierarchical regression model, the posteriors for the same players are obtained and presented in Table 4.7.

Table 4.7 Comparison of players with a significant result in the rank-sum test for positive location difference with the result from regression model. Players are sorted by their p-value of their rank-sum tests.

		Rank-sum		Model 1.1			Model 2.1			Model 3.1		
	M	*PM*	*P-val*	*mean*	*PPS*	*%*	*mean*	*PPS*	*%*	*mean*	*PPS*	*%*
José León	12	4.01	0.01	-0.85	0.14	99	3.31	0.98	99	3.67	0.99	100
John Björkengren	28	2.73	0.02	-1.11	0.08	95	1.78	0.81	96	2.06	0.83	94
Erik Berg	10	7.62	0.03	-1.08	0.10	97	7.37	0.99	100	7.37	0.98	99
Kirill Pogrebnyak	17	2.41	0.05	-1.31	0.03	73	0.04	0.51	75	2.56	0.75	90
Tobias Karlsson	26	2.66	0.05	-0.77	0.16	100	1.93	0.94	99	2.16	0.96	99
Jordan Attah Kadiri	10	17.79	0.08	-1.33	0.04	86	5.19	0.86	96	11.00	0.97	99
Tom Pettersson	22	2.04	0.08	-0.85	0.14	99	2.21	0.93	99	2.16	0.92	98
Behrang Safari	16	2.31	0.09	-0.90	0.13	99	2.20	0.92	98	2.39	0.93	98
Yaser Kasim	11	2.01	0.09	-0.95	0.11	99	1.77	0.89	98	2.16	0.92	98

In Table 4.7, “PM” indicates the pseudo median difference between the observations of the two samples using the rank-sum test described in 2.2. Comparing the pseudo medians of the players to the means of the “big match” effect of each model, we can see that they seem to correlate the most with model 3.1. The column “PPS” represents the positive probability space, as defined in section 3.3. “M” denotes the number of matches played during the 2019 season and “%” indicates the league percentiles in which the players fall ranked by their PPS-scores in each separate model.

The PPS percentiles of the players are very similar across the models, with all players placing among the highest in the league. The exceptions are Kirill Pogrebnyak, who still places no lower than in the 51th percentile (in model 2.1). Through studying the data, we found that Pogrebnyak along with Jordan Attah Kadiri (who also scored low in the hierarchical models) were the players with the by far largest variance in performances. They both had a couple of extremely high scoring performances, but usually scored rather poorly compared to league average in the rest of their matches.
This observation points to the advantage of the hierarchical models where, thanks to the priors, some of that variance is taken into account.

5 Discussion

We began our study by exploring the problem and the data at hand through simple rank-sum tests during our preliminary analysis phase. This gave us an indication that the “big match ability” could be measured. We were able to identify players with significant positive locations differences, which we later compared using the Bayesian models. The non-parametric tests are easy to perform, but the result is not as easy to interpret as the results of the Bayesian inference. Furthermore, it does not satisfy our objectives of constructing a method and a metric to quantify the ability and to compare players based on the ability.

We ruled out the use of the non-binary definition of the “big match” covariate partly due to the lesser fit indicated by the DIC scores, but also because the binary covariate is easier to understand, explain and validate.

The DIC values suggests that we should use the non-hierarchical model (model 3.1) with the binary covariate, but we also need to take the interpretability of the model results into account.

The non-informative priors in model 3.1 makes it the most relevant model to compare to the rank-sum test method since they both “let the data speak for itself”. It is also between these two we see the most correlation between p-values of the rank-sum test and PPS percentile (see Table 4.7). This indicates that if we want to assess players solely based on the observations with no prior information, the rank-sum approach could be considered a viable frequentist heuristic method to find players that overperformed against “big teams” (given that the sample size is large enough and that sensitivity to variance is not a concern). The PPS scores obtained from the non-hierarchical Bayesian model is however superior as a metric, thanks to its interpretability and the fact that we can use it to rank and compare all the players in the population.

The non-hierarchical model is not dependent on any population- or group priors, which makes it simple to apply and its result reproducible using any number of subjects from the population. But the lack of informative priors is also its weakness; with no prior information, the posteriors are only based on the observational data. For some of the players in our dataset, the number of observations might not be enough to consider the non-hierarchical model fitting.
We therefore suggest that the non-hierarchical model should be primarily used when assessing and comparing players that played most matches during the season of interest. It may also be highly applicable when analysing data with larger amounts of individual observations, such as the raw event data or data describing movement patterns.

The hierarchical models (model 1 and model 2) incorporate prior knowledge about the distributional aspects of the response variable. Instead of assuming that the performance score and the “big match” effect can take any value at the same probability, we make some realistic predictions to get us closer to the truth. Model 2 assumes that the distribution of performance scores of each individual player is restricted by the performance distribution in the population. Model 1 adds an additional level of constriction, by which we assume that the distributional parameters of performance scores and the “big match” effect of players depend on the players’ team level. These hierarchies arguably get us closer to predicting the “true” parameters and the latent ability, but they naturally make the posteriors less representative of the raw observations by skewing them towards the group- and population priors. For example, we ended up with no PPS scores above 0.16 in model 1.1, even though 123 players received a PPS score over 0.50 in model 3.1.

While we believe our models can be used to estimate and compare the “big match ability” of different players, we suggest that the hierarchical aspects of the data (and football players in general) should be further explored in order to find an appropriate model with appropriate shrinkage that can make more accurate assessments and predictions.

The total score per minute played as a measure of performance can be questioned. For instance, it does not take the team performance into account: a score of 500 can be considered high in a match where the team total score is 2000, but low in a match where the team total is 10000.
To deal with this issue, we could normalize the performance scores based on the total performance score of all teammates in the match. This should make the response variable more representative of the relative level of performance and how much a player contributed to the match result and might improve the predictive qualities of the model.
It also disregards the effects of playtime: a substitute that only plays the last 20 minutes of the match could potentially play more intensively and gain a higher average performance score compared to a player that is on the field for the entire 90 minutes.

The PPS score can also be considered problematic as a metric since it only measures the “big match effect” relative to the players’ own baseline performance level. A bad player that plays better (but still bad) against big teams would still be considered a ”big match player” due to his/her high PPS score. This underlines the importance of using the PPS score in combination with other metrics, such as predicted performance scores, in order to assess how well a player will actually play against tough opposition.

6 Conclusions

In this thesis, we explored the objective of trying to find a method and a metric for assessing the ability to perform against difficult opposition by analysing different covariates and linear models. We introduced a novel metric we call PPS and showed how it can be applied as a measure of the "big match" potential defined by the probability that a player will perform better against a "big team" rather than against to "small team".

The strength of the PPS score as a metric lies in its simplicity. If we believe in the model from which it is obtained, it can be put to great use when comparing players. Say for example that a top club that is a title contender is looking for a new player acquisition. They need a new player and have narrowed it down to two individuals that have equal physical and technical abilities. Being able to assess their ability to perform when it matters the most for this club (in close games and against difficult opposition) would be a great help to make the best decision on who to recruit.

The delimitations mentioned in 1.5 suggests there are many topics related to our study that could be explored.
We purposefully limited our study to the match context defined by the level of opposition and chose a simple approach to assign this variable to the different matches. A more sophisticated method could be applied where the level of opposition is defined by a range of different relevant variables that describes the match context. As shown by the model comparisons in Table 4.2, using the final league points of the opposing team does not capture all aspects of what makes a match "big". Different opposing teams are not automatically equally difficult for all teams. Rivalry could be one such factor, but so could tactics of the opposing team.
Recent form of both teams could be included in the model to account for temporary trends in performance and missing key players (due to injuries or suspensions) may be incorporated into the categorization.

But in order to capture the true latent "big match player" ability, we should add other types of match contexts to our player assessments. One such context that we could model in a similar way to what we have done with "level of opposition" in this thesis would be "match pressure". A pre-game pressure covariate similar to the "match importance" index presented by Bedford and Schembri (2006) could be modelled based on the implicit effects of the potential match outcomes. Much like Bransen *et al.* (2019) suggests, a and an in-game pressure covariate could be modelled based on the probabilities of match outcomes given the time left and relative goal difference.

Once an adequate model of the ability has been established, further topics related to its implications could be studied, such as transfer sums and the long-term effects on club success based on match outcomes and overall seasonal performance.

In the meantime though, we believe the PPS score metric is relevant enough to be used for making predictions. And if we have learned anything about this first foray into Bayesian statistics, it is that if we start making some predictions we might come a bit closer to the truth.

References

Aggarwal, C. C. & Reddy, C. K. eds. (2014). *Data clustering: algorithms and applications*. Boca Raton: CRC Press.

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

Beetz, M., Hoyningen-Huene, N. von, Kirchlechner, B., Gedikli, S., Siles, F., Durus, M. & Lames, M. (2009). ASPOGAMO: Automated Sports Games Analysis Models. *Int. J. Comp. Sci. Sport*, 8(1).

Bransen, L., Robberechts, P., Van Haaren, J. & Davies, J. (2019). Choke or Shine? Quantifying Soccer Players' Abilities to Perform Under Mental Pressure. In: ESPN, *MIT Sloan Sport Analytics Conference*. Boston, USA 6-7 March 2020. Enschede: SciSports.

Bedford, A. & Schembri, A. J. (2006). A Probability Based Approach for the Allocation of Player Draft Selections in Australian Rules Football. *Journal of Sports Science & Medicine*, 5(4), pp. 509–516.

Egidi, L. & Gabry, J. (2018). Bayesian hierarchical models for predicting individual performance in soccer. *Journal of Quantitative Analysis in Sports*, 14(3), pp. 143–157, doi:10.1515/jqas-2017-0066.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition.* Hoboken: CRC Press.

Gelman, A. & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4), pp. 457–472.

Hodges, J. L. & Lehmann, E. L. (1963). Estimates of Location Based on Rank Tests. *Annals of Mathematical Statistics*, 34(2), pp. 598–611, doi:10.1214/aoms/1177704172.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan*. Edition 2. Boston: Academic Press.

Link, D. & de Lorenzo, M. F. (2016). Seasonal Pacing - Match Importance Affects Activity in Professional Soccer. *PLoS ONE*, 11(6), pp. 1–10, doi:10.1371/journal.pone.0157127.

Murphy, K. P. (2007). *Conjugate Bayesian analysis of the Gaussian distribution*. Technical report, University of British Columbia.

Peralta Alguacil, F. J. (2019). Modelling the Collective Movement of Football Players. Master Thesis, Uppsala University.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Presented at the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Vienna, Austria.

Raftery, A. E. & Lewis, S. M. (1992). [Practical Markov Chain Monte Carlo]: Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo. *Statistical Science*, 7(4), pp. 493–497.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Linde, A. van der (2002). Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(4), p. 583.

Appendix A: The Twelve player performance assessment model.

Twelve primarily make use of logistic regression and other supervised machine learning methods to calculate the probability that different actions lead to a goal. Here we outline the method we use for passes.

All the matches used to train the model are broken down into sequences of possession, i.e., fragments of the game during which one of the teams holds the possession of the ball without losing the ball and without any stops in the play (due to fouls, throw-ins, offsides, etc.). A chain was considered broken and another chain begun whenever the opposition team made two consecutive touches of the ball.

Once all actions are allocated to a possession chain then two logistic regressions are fitted in order to assign a value to each pass. The first regression is obtained by assigning each possession chain a value between 0 (if the play ends without a shot) and 1 (if the sequence finishes with a goal).

$$P(shot|x_1, x_2, y_1, y_2)$$

This gives the probability of a pass (defined by its starting and ending co-ordinates on the pitch) leading to a shot. A second regression is then used to compute the probability of a shot leading to a goal (i.e. to obtain the expected goal value of the shot).

$$P(goal|shot, x_1, x_2, y_1, y_2)$$

Multiplying these two probabilities for every shot gives the probability that a pass of with certain starting and ending co-ordinates and qualifiers is likely to result in a goal. It is this value which we call the pass impact.

Similar methods are used for the value of dribbles. A reverse process is then used for tackles, headers and so based on how the action reduces the oppositions probability of scoring.

Appendix B: Tables

Table B.1 The final score of Allsvenskan 2019, GF = Goals For, GA = Goals Against, GD = Goal Difference.

Position	Team	Played	Wins	Draws	Loss	GF	GA	GD	Points	BigTeam
1	Djurgårdens IF	30	20	6	4	53	19	34	66	1
2	Malmö FF	30	19	8	3	56	16	40	65	1
3	Hammarby IF	30	20	5	5	75	38	37	65	1
4	AIK	30	19	5	6	47	24	23	62	1
5	IFK Norrköping	30	16	9	5	54	26	28	57	1
6	BK Häcken	30	14	7	9	44	29	15	49	1
7	IFK Göteborg	30	13	9	8	46	31	15	48	1
8	IF Elfsborg	30	11	10	9	44	45	-1	43	0
9	Örebro SK	30	9	6	15	40	56	-16	33	0
10	Helsingborgs IF	30	8	6	16	29	49	-20	30	0
11	IK Sirius	30	8	5	17	34	51	-17	29	0
12	Östersunds FK	30	5	10	15	27	52	-25	25	0
13	Falkenbergs FF	30	6	7	17	25	62	-37	25	0
14	Kalmar FF	30	4	11	15	22	47	-25	23	0
15	GIF Sundsvall	30	4	8	18	31	50	-19	20	0
16	AFC Eskilstuna	30	4	8	18	23	55	-32	20	0

Appendix C: Code

C.1 JAGS model definitions

```
# Model 1: Full hierarchical model:
mod_string_mod1 =
  "model {
    # Likelihood:
    for ( i in 1:N ) {
      y[i] ~ dnorm(b0[player[i]] + b1[player[i]] * x[i], tau.p[player[i]])
    }
    # Player priors:
    for ( p in 1:Nplayers ) {
      b0[p] ~ dnorm(b0.tl[teamLevel[p]], tau.b0.tl[teamLevel[p]])
      b1[p] ~ dnorm(b1.tl[teamLevel[p]], tau.b1.tl[teamLevel[p]])
      tau.p[p] ~ dgamma(0.001, 0.001)
    }
    # Team level priors
    for ( tl in 1:2) {
      b0.tl[tl] ~ dnorm(b0.l, tau.b0.l)
      b1.tl[tl] ~ dnorm(0, 0.001)
      tau.b0.tl[tl] ~ dgamma(0.001, 0.001)
      tau.b1.tl[tl] ~ dgamma(0.001, 0.001)
    }
    # League priors:
    tau.b0.l ~ dgamma(0.001, 0.001)
    b0.l ~ dnorm(0, 0.001)
    }
"
# Model 2: Reduced hierarchical model:
mod_string_mod2 =
"model {
    # Likelihood:
    for ( i in 1:N ) {
      y[i] ~ dnorm(b0[player[i]] + b1[player[i]] * x[i], tau.p[player[i]])
    }
    # Player priors:
    for ( p in 1:Nplayers ) {
      b0[p] ~ dnorm(b0.l, tau.b0.l)
      b1[p] ~ dnorm(0, 0.001)
      tau.p[p] ~ dgamma(0.001, 0.001)
    }
    # League priors:
    tau.b0.l ~ dgamma(0.001, 0.001)
    b0.l ~ dnorm(0, 0.001)
    }
"
# Model 3: Normal regression model:
mod_string_mod3 =
"model {
    # Likelihood:
    for ( i in 1:N ) {
      y[i] ~ dnorm(b0[player[i]] + b1[player[i]] * x[i], tau.p[player[i]])
    }
    # Player priors:
    for ( p in 1:Nplayers ) {
      b0[p] ~ dnorm(0, 0.001)
      b1[p] ~ dnorm(0, 0.001)
      tau.p[p] ~ dgamma(0.001, 0.001)
    }
  }
"
```

Figure C.1 JAGS model definitions.

C.2 Running the model

```
#####################################################################
## Load libraries
#####################################################################
if (!require(runjags)) install.packages('runjags')
if (!require(MCMCvis)) install.packages('MCMCvis')
library(runjags)
library(MCMCvis)
library("DBDA2E-utilities")
```

Figure C.2 Loading libraries and data.

```
#####################################################################
## Create parameter posterior summary tables for the intercept and slope:
#####################################################################
posteriorSummary <- function(codaSamplesObject, modelName, sourceDF, playerIds) {
  summary_b0<-MCMCsummary(codaSamplesObject, probs=c(0.025, 0.05, 0.5, 0.95, 0.975),
    params=c("b0"), round = 4, Rhat=TRUE, n.eff=TRUE)
  summary_b0$playerId<-playerIds
  addToMCMCSummary(summary_b0, sourceDF, "b0") -> summary_b0
  summary_b1<-MCMCsummary(codaSamplesObject, probs=c(0.025, 0.05, 0.5, 0.95, 0.975),
    params=c("b1"), round = 4, Rhat=TRUE, n.eff=TRUE)
  summary_b1$playerId<-playerIds
  addToMCMCSummary(summary_b1, sourceDF, "b1") -> summary_b1
  addB1PPS(summary_b1, codaSamplesObject) -> summary_b1
  assign(paste(modelName, "summary_b0", sep="_"), summary_b0, envir = .GlobalEnv)
  assign(paste(modelName, "summary_b1", sep="_"), summary_b1, envir = .GlobalEnv)
}
```

Figure C.3 Function for creating posterior summary tables.

```
#####################################################################
## Add additional data to the summary tables:
#####################################################################
addToMCMCSummary <- function(summaryDF, SourceDF, parameter) {
  summaryDF$playerId<-playerId
  for (i in 1:length(summaryDF$playerId)) {
    summaryDF$Name[i]<-as.character(SourceDF$PlayerName[match(summaryDF$playerId[i],
      SourceDF$PlayerId)])
    summaryDF$Team[i]<-as.character(SourceDF$Team[match(summaryDF$playerId[i],
      SourceDF$PlayerId)])
    summaryDF$matchesPlayed[i]<-matchesPlayed$Freq[match(summaryDF$playerId[i],
      matchesPlayed$Var1)]
  }
  summaryDF$`90%CredibilityIntervalRange`<-summaryDF$`95%`-summaryDF$`5%`
  summaryDF$`95%CredibilityIntervalRange`<-summaryDF$`97.5%`-summaryDF$`2.5%`
  return(summaryDF)
}
```

Figure C.4 Function for adding additional information to posterior summary tables.

```
######################################################################
## Add PPS to the b1 summary table:
######################################################################
addB1PPS <- function(summaryDF, codaSamples) {
  for (i in 1:length(summaryDF$playerId)) {
    b1<-paste("b1[", i, "]", sep="")
    paramSampleVec<-as.matrix(codaSamples[, b1])
    summaryDF$positiveProbabilitySpace[i]<-sum( paramSampleVec > 0 ) / length(
      paramSampleVec )
  }
  rm(paramSampleVec, b1)
  return(summaryDF)
}
```

Figure C.5 Function for adding the PPS scores to a summary table.

```
######################################################################
## Create posterior plots for a combination of models and players:
######################################################################
B1ModelComparison <- function(codaSampleObjectNames=c("model_1", "model_2", "model_3"),
players=c(157816, 86249, 76466, 19057), sourceDF) {
  par(mfrow=c(length(players),length(codaSampleObjectNames)))
  par(mar=c(5,1,1,1), oma=c(2,0,2,0))
  for (i in players) {
    prow <- which(playerId == i, arr.ind=TRUE)
    b1<-paste("b1[", prow, "]", sep="")
    pname<-sourceDF$PlayerName[match(i, sourceDF$PlayerId)]
    for (j in 1:length(codaSampleObjectNames)) {
      codaSamples<-get(codaSamplesObjectNames[j])
      plotPost( codaSamples[, b1], main=pname, cenTend="mean", compVal=0,
        credMass= 0.9, xlab=bquote(names(codaSampleObjectNames)[j]))
    }
  }
  par(mfrow=c(1,1))
  par(mar=c(5.1, 4.1, 4.1, 2.1), oma=c(0,0,0,0))
}
```

Figure C.6 Function for creating posterior plots.

```
######################################################################
## Create trace plots for a combination of models and players:
######################################################################
TracePlots <- function(codaSampleObjectNames=c("codaSamples_mod1", "codaSamples_mod2",
"codaSamples_mod3"), players=c(), sourceDF) {
  par(mfrow=c(length(players),length(codaSampleObjectNames)))
  par(mar=c(5,1,1,1), oma=c(2,0,2,0))
  for (j in 1:length(codaSampleObjectNames)) {
    codaSamples<-get(codaSampleObjectNames[j])
    for (i in players) {
      prow <- which(playerId == i, arr.ind=TRUE)
      b0List[i]<-paste("b0\\[", prow, "\\]", sep="")
      b1List[i]<-paste("b1\\[", prow, "\\]", sep="")
    }
    MCMCtrace(codaSamples, params = b0List, type="trace", ind=FALSE,
      ISB = FALSE, pdf = FALSE, Rhat=FALSE)
    MCMCtrace(codaSamples, params = b1List, type="trace", ind=FALSE,
      ISB = FALSE, pdf = FALSE, Rhat=FALSE)
  }
  par(mfrow=c(1,1))
  par(mar=c(5.1, 4.1, 4.1, 2.1), oma=c(0,0,0,0))
}
```

Figure C.7 Function for plotting trace plots.

```
#####################################################################
## Function for running the model:
#####################################################################
runModel = function( modelNr, playerMatchData, binaryCovariate=TRUE, chains=3,
thinningInterval=2, maxRunTimeMinutes=30, output=TRUE) {
  if (modelNr %in% c(1,2,3)) {
    y<-playerMatchData$TotalScorePM # regressand
    N<-length(y) # number of observations
    xb<-playerMatchData$OpponentBigTeam # regressor: level of opposition (binary)
    xnb<-playerMatchData$OpponentLeaguePoints # regressor: level of opposition (non-binary)
    player<-as.factor(playerMatchData$PlayerId) # factor categorized by player ID
    playerId<-as.numeric(levels(player)) # all unique player IDs
    teamLevel<-as.factor(playerMatchData$TeamBigTeam)
    Nplayers<-nlevels(player) # Number factors in the level
    matchesPlayed<-as.data.frame(table(playerMatchData$PlayerId))
    if(binaryCovariate) { x<-xb } else { x<-xnb  }

    if (modelNr==1) {
      data_jags <- list(y=y, x=x, player=player, teamLevel=teamLevel,
        Nplayers=Nplayers, N=N)
      mod_string <- mod_string_mod1
      params <- c("b0", "b1", "tau.p", "b0.tl", "tau.b0.tl",  "b1.tl",
        "tau.b1.tl", "b0.l", "tau.b0.l")
    }
    else if (modelNr==2) {
      data_jags <- list(y=y, x=x, player=player, Nplayers=Nplayers, N=N)
      mod_string <- mod_string_mod2
      params <- c("b0", "b1", "tau.p", "b0.tl", "b0.l", "tau.b0.l")
    }
    else if (modelNr==3) {
      data_jags <- list(y=y, x=x, player=player, Nplayers=Nplayers, N=N)
      mod_string <- mod_string_mod3
      params <- c("b0", "b1", "tau.p")
    }

    modelName<-paste("model", modelNr, ifelse(binaryCovariate==TRUE, "BC", "NBC"), sep="_")
    maxRuntime<-paste(maxRunTimeMinutes,"m",sep="")

    results <- autorun.jags(model=mod_string, data=data_jags,
      max.time=maxRuntime, monitor=params,
                      n.chains=chains, thin=thinningInterval, method="rjags")
    DIC<-extract(results, "dic")
    assign(paste(modelName, "DIC", sep="_"), DIC, envir = .GlobalEnv)
    codaSamples <- as.mcmc.list(results)
    assign(paste(modelName, "codaSamples", sep="_"), codaSamples, envir = .GlobalEnv)
    write.jagsfile(results, paste("jagsFile_", modelName, ".txt", sep=""), remove.tags =
      TRUE, write.data = TRUE, write.inits = TRUE)
    posteriorSummary(paste(modelName, "codaSamples", sep="_"), modelName)
  }
  else {
    print("Warning! Unknown model number, no model was run!")
  }
}
```

Figure C.8 Main function for running the model

Appendix D: Trace plots

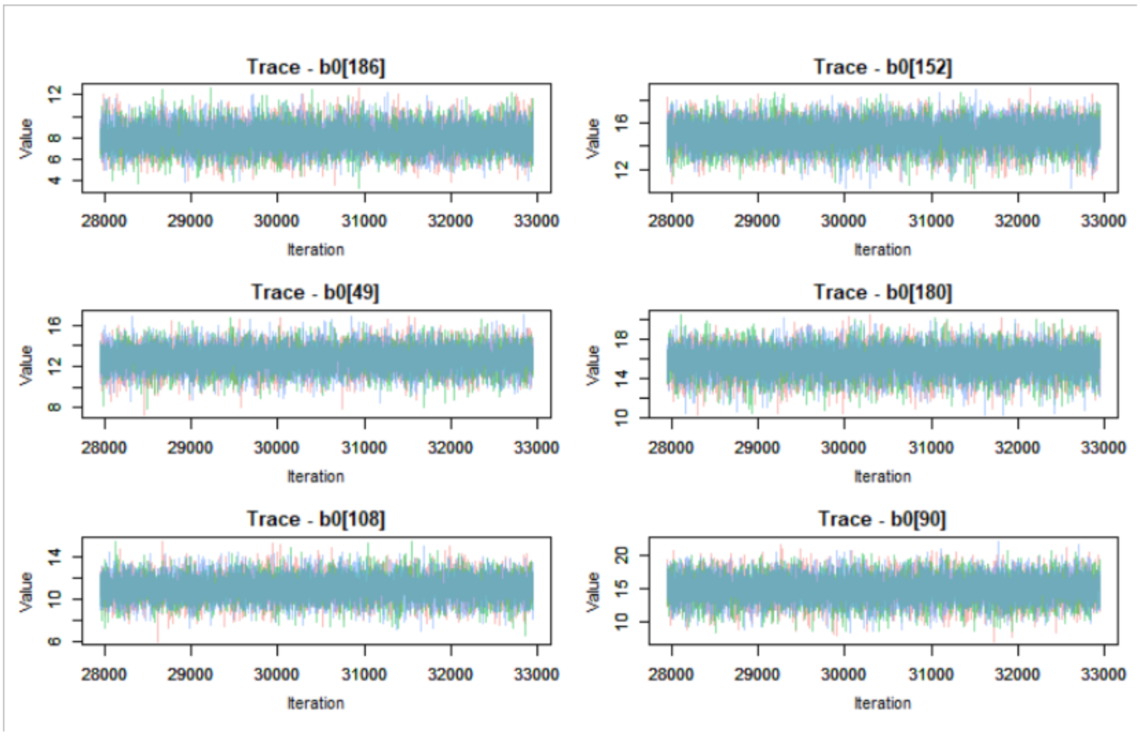

Figure D.1 Trace plots for the intercept in model 1.1.

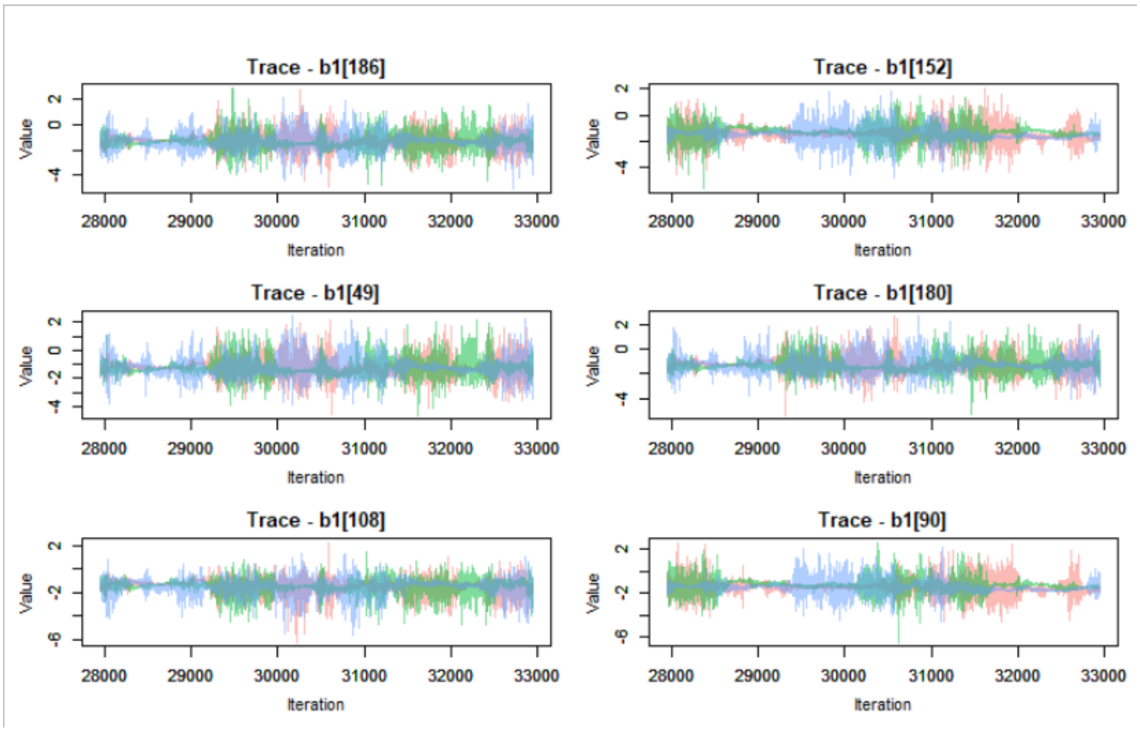

Figure D.2 Trace plots for the "level of opposition" regression coefficient in model 1.1.

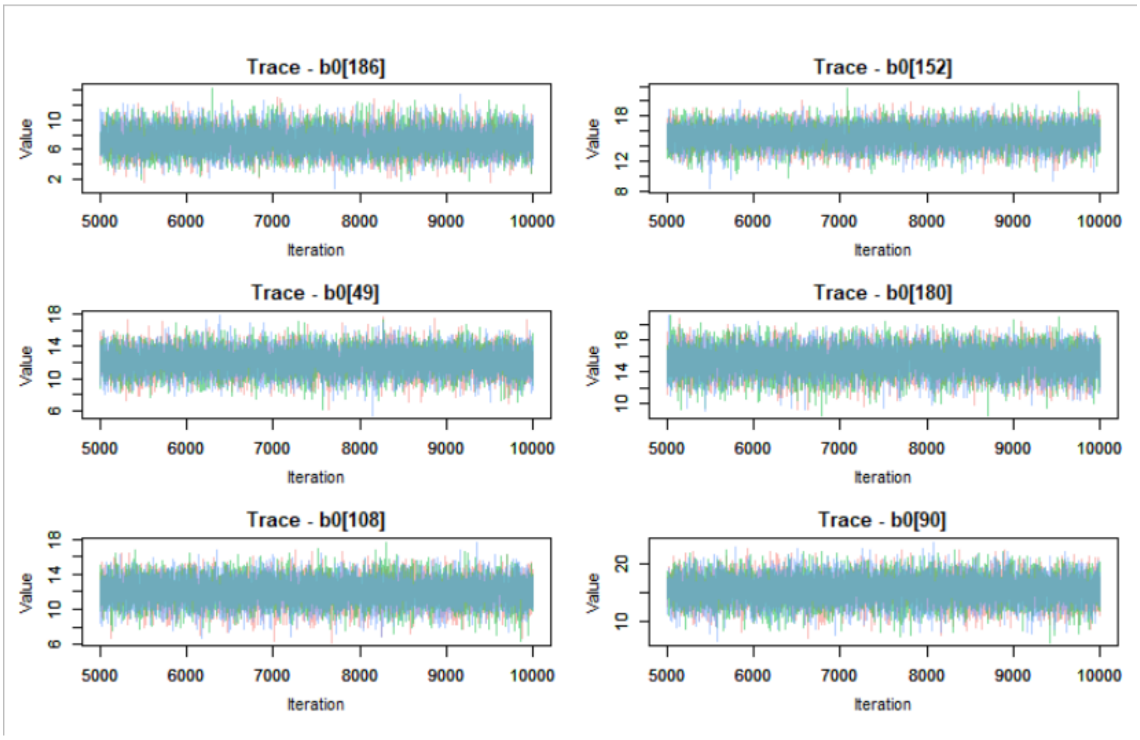

Figure D.3 Trace plots for the intercept in model 2.1.

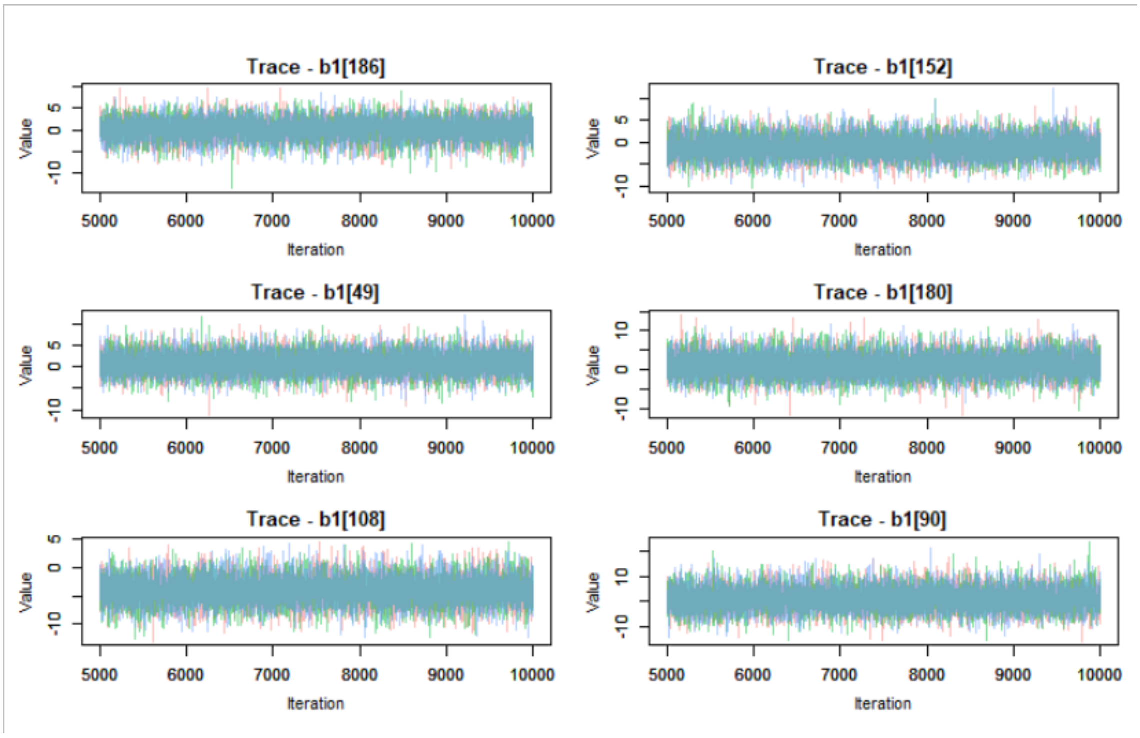

Figure D.4 Trace plots for the "level of opposition" regression coefficient in model 2.1.

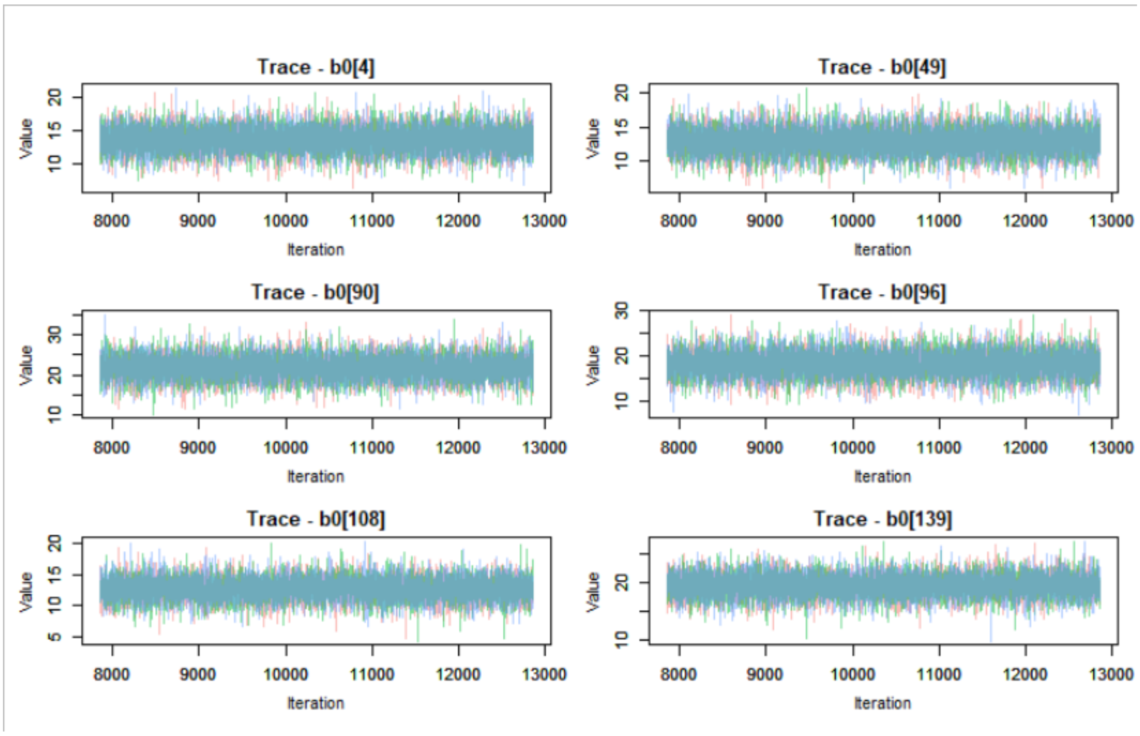

Figure D.5 Trace plots for the intercept in model 3.1.

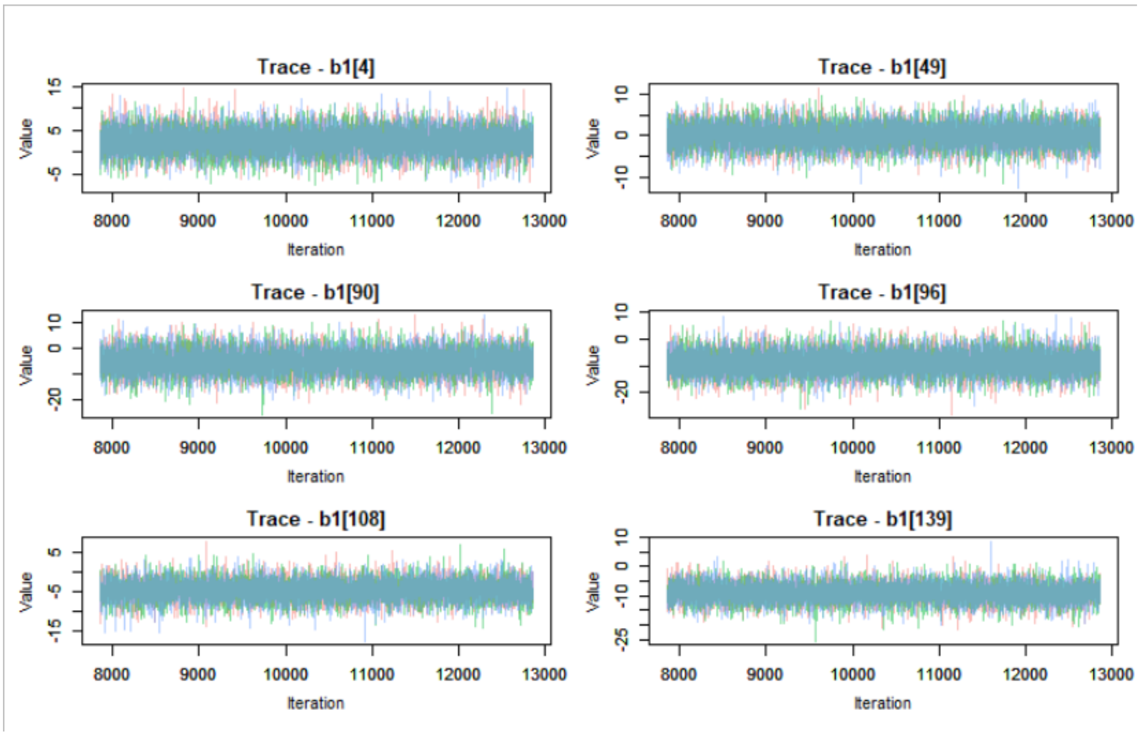

Figure D.6 Trace plots for the "level of opposition" regression coefficient in model 3.1.