axioms

MDPI

Article

Is Football/Soccer Purely Stochastic, Made Out of Luck, or Maybe Predictable? How Does Bayesian Reasoning Assess Sports?

Leonardo Barrios Blanco [1], Paulo Henrique Ferreira [2], Francisco Louzada [3] and Diego Carvalho do Nascimento [1,*]

[1] Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Copiapó 1530000, Chile; leonardo.barrios.2020@alumnos.uda.cl

[2] Department of Statistics, Federal University of Bahia, Salvador 40170110, Brazil; paulohenri@ufba.br

[3] Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos 13566590, Brazil; louzada@icmc.usp.br

* Correspondence: diego.nascimento@uda.cl

Abstract: Predicting the game score is a well-explored duty, using mathematical/statistical models. Nonetheless, by adopting a Bayesian methodology, this study aimed to estimate probabilistically the Chilean Premier League teams' position, considering them a hierarchical structure. This approach enabled the evaluation of the main Chilean championship that provides the major soccer players for the national team. Thus, a countable (Poisson) regression structure was considered to explain each match as a combination of home advantage, added to the power of attack and defense of each team and considering their performance in the championship as an independent game. We were able to quantify the relationship across the defense and attack of each team and, in addition, were able to group/verify the performance of the entirety of the 2020 Chilean Premier League. For the model validation, we saved the last five games for the model prediction and we found that, in this league, the teams presented a statistical significance in the attack factors, which influences the scores (goals); however, all the teams showed low defense power and we have also found that playing at home or away did not present a game advantage. Our model was able to predict the Chilean league position table, with precision on the top five positions, and from the 6–11 positions there was a small shift (close performance in the championship) caused by the similarity of the expected number of goals, which implied the same position on the rank. This type of model has been shown to be very competitive for the soccer championship prediction.

Keywords: Hierarchical Bayesian model; Poisson regression; soccer game prediction; Chilean Premier League

check for updates

Citation: Barrios Blanco, L.; Ferreira, P.H.; Louzada, F.; Nascimento, D.C. Is Football/Soccer Purely Stochastic, Made Out of Luck, or Maybe Predictable? How Does Bayesian Reasoning Assess Sports? *Axioms* **2021**, *10*, 276. https://doi.org/10.3390/axioms10040276

Academic Editors: Hari Mohan Srivastava and Kevin H. Knuth

Received: 1 September 2021
Accepted: 21 October 2021
Published: 26 October 2021

1. Introduction

The mathematical systematization of sports performance and talent search as a crucial mechanism for obtaining good players for large professional teams is known in both sports science and in the literature on sports management [1]. In European and American contexts, the sports industry has a longer tradition in collecting statistical data, such as for baseball, soccer and basketball [2,3]. In recent years, digital economy has brought new technologies (player analysis) that have offered additional advantages, such as predicting player pairings or an athlete's performance under specific conditions, all these based on classical statistical tools.

However, several sports organizations traditionally rely almost exclusively on the human experience for the talent process, even though in the past decades statistical computations were more heavily used [4]. In soccer, it is still believed that experts in the field, such as coaches, managers and talent scouts, can effectively convert collected data into usable knowledge. For example, Louzada et al. [5] proposed the creation of performance indicators based on multivariate statistical analysis. From this tool, one can quantify the performance of young players and take into account expert opinion for talent search or personalized training.

In this direction, Bayesian methods make it possible to combine the information contained in the observed data and the subjectivity of different experts in sports [6]. The knowledge acquired from the experts will be added as a prior probabilistic distribution as, for example, these elements may consider aspects such as the expected recruitment of a player (to a certain team) or his/her physical performance.

The soccer national leagues, seen from a global perspective, generates results on a daily basis, which makes the amount of data that can be studied increase [7]. That passion, which has various economic and social implications, is of great interest in order to see the linking of probabilistic models for the quantification of their results (and performance measurement). In this sense, statistical modeling applied to sports is a popular tool that has promoted various investigations [8–10]. Two of the main aspects to investigate in the world of soccer are who will be the winner of the match? Or what will the number of goals scored by each team be? There have been various methodologies and probabilistic distributions used to carry this goal out, for example, in different studies the Binomial, Negative Binomial and Poisson distributions have been used, in addition to using both the classical and Bayesian inference perspectives.

Moreover, the Poisson distribution has been widely accepted as a suitable model for this kind of prediction; in particular, this model is often used because independence is assumed between the goals scored by the home team and the visitor (given its simplicity, which demands the estimation of a single parameter). For example, a soccer match can be seen as two models (Y_1, Y_2), in which Y_1 is the number of goals scored by the home team and Y_2 is the number of goals scored by the visitor, though adopting the independence of the bivariate Poisson in which the relevant parameters are constructed as the combination of the attack and the defense of the teams.

For instance, Santana et al. [11] adopted the Poisson distribution to estimate the probability of the vector (victory, tie, defeat) of each team on a given match. They demonstrated the effectiveness of this proposed model whereas a modified model (Poisson autoregressive model with exogenous covariates) was used to estimate the result of a given match, compared to previously adopted multinomial logistic regression and support vector machines models, which do not inform the score of the match (and the discrete probability model does).

In statistical methodology, Bayesian methods allow the inclusion of information from an investigator's preliminary tests, allowing combination with the observed data. For instance, linear models can be understood as a combination of observed variables and latent effects, which can show hidden patterns, and can be adopted as a hierarchical Bayesian model to enable the estimation of individual effects in groups.

The present study proposed to develop a hierarchical Bayesian model to estimate the positions of the teams in the general table of the Chilean League 2020, saving the last five rounds as a prediction. The adopted model was based on the number of goals of each team to measure the attack power and defense both at home and away, and also the effect of home advantage. Hierarchical models are widely used as they are a natural way of taking into account the relationships between variables, assuming a common distribution for a set of relevant parameters that are believed to be related to one another.

The structure of this article is as follows: Section 2 provides the basis for the model proposal, Section 3 describes the proposed model and how the data were manipulated, Section 4 describes the results in terms of parameter estimation and prediction of a new outcome, and finally Section 5 highlights the respective conclusions and possible improvements to be made to following investigations.

2. Theoretical Framework

The sports world has always been interested in having a certainty of winning, since private teams invest large amounts of money to achieve good results. There are many aspects that are covered in the context of sports, such as the physical condition of the athletes, the maintenance of the place where the sport is developed, as well as having the best technical staff. This investment should ensure that it returns monetary gains.

A particular case is soccer tournaments, in which it seems difficult to predict the outcome of a game or who will be the winner in a certain championship, because there are many factors involved, such as those previously exposed (athletes, fields and technical personnel) as well as the presence of the public on the fields, injuries of the players, among others. Various studies have been able to predict winners of tournaments and matches, with different statistical models [10,12,13]. An example of this is the large betting houses that, despite their work being based on "chance", use models to benefit themselves and not the user.

Since the beginning of the 20th century, much data have been collected and, currently, the amount of data that can be collected is much higher, due to current technology. There have been several investigations into whether the result of soccer is luck or can be predicted [7].

In this sense, seeking to model a soccer game, a game can be represented with the number of goals of the gth team as a combination of the home versus away teams as Y_{g1} and Y_{g2}, respectively, in which the observed elements (the amount of goals) are bivariate variables $Y = (Y_{g1}, Y_{g2})$ that can be modeled as independent events. Moreover, in a championship, the league is structured in such a way that every team will confront each other twice, once as the home team and another as the away team. In this case, the random variable that counts the number of goals is discrete, therefore the Poisson distribution would naturally be adopted as:

$$Y_{gj} \mid \theta_{gj} \sim \text{Poisson}(\theta_{gj}), \tag{1}$$

conditioned to the parameters $\boldsymbol{\theta} = (\theta_{gi}, \theta_{gj})$, which represent the play intensity of the home team (i) and visitor team (j), respectively.

For instance, let us consider a univariate discrete random variable Y that follows a Poisson distribution [14] with parameter θ, represented by $Y \sim \text{Poisson}(\theta)$, with probability mass function given by:

$$P(Y = y \mid \theta) = \frac{e^{-\theta}\theta^{y}}{y!}, \quad \text{for } y = 0, 1, 2, \ldots, \tag{2}$$

which has the property that the expectation and the variance are equal, that is, $\mathbb{E}[Y] = \mathbb{V}\text{ar}[Y] = \theta$.

Extending this univariate parameterization as a regression structure that associates explanatory factors to the composition of the parameter θ, taking $\boldsymbol{X} \in \mathbb{R}^n$ a vector of independent variables, we have that a regression model could be written as $\log(\mathbb{E}[Y|\boldsymbol{X}]) = \boldsymbol{\beta X}$, in which $\boldsymbol{\beta} \in \mathbb{R}^k$ denotes a vector of regression coefficients. Thus, the extension of this statistical model, incorporating an explanatory vector structure ($\boldsymbol{X}$) to estimate the parameter θ, is:

$$\theta := \mathbb{E}[Y \mid \boldsymbol{X}] = e^{\boldsymbol{\beta X}}. \tag{3}$$

Then,

$$P(Y = y \mid \boldsymbol{X}, \boldsymbol{\beta}) = \frac{e^{y\boldsymbol{\beta X}} e^{-e^{\boldsymbol{\beta X}}}}{y!}. \tag{4}$$

If a bivariate discrete variable is considered, (Y_1, Y_2), with independence, then each of the variables following a Poisson distribution are seen as $P(Y_1, Y_2) = P(Y_1)P(Y_2)$. Additionally, if a Bayesian approach is adopted, the parameters are random variables, and one can estimate this regression model parameters, $(\theta_{gi}, \theta_{gj})$, as random variables effect (of each gth team) and assuming a log-linear as the (canonical) link function as:

MODEL 1 (NON-HIERARCHICAL):

$$\log(\theta_{g1}) = \beta_{\text{home}} + \beta_{1(g)}\text{att}_{\text{home}(g)} + \beta_{2(g)}\text{def}_{\text{away}(g)}, \tag{5}$$

$$\log(\theta_{g2}) = \beta_{3(g)}\text{att}_{\text{away}(g)} + \beta_{4(g)}\text{def}_{\text{home}(g)}. \tag{6}$$

The parameter β_{home} represents the home advantage for the game taking place at home, which is assumed as constant for each team and match. The parameters β_1 and β_3 represent the ability to attack, and β_2 and β_4 the defense of each team. The elements considered $\{(\text{home}(g), \text{away}(g)), \forall g = 1, \ldots, n\}$ indicate the n teams which are playing in the league.

Thus, for each of the parameters $(\beta_{1(g)}, \beta_{2(g)}, \beta_{3(g)}, \beta_{4(g)})$ indicated in the first modeling approach, which will enable us to estimate a common attack and defense structure among the teams in a championship, varying the scale parameters, however, also represent the league's characteristics [15]. Nevertheless, a second modeling approach is to consider the common shared structure added by a hierarchical level, which can design the attack strength of each team (regardless of playing at home or away), as well as the defense strength. Moreover, in the second model, the parameters $(\beta_{1(g)} = \beta_{3(g)}$ and $\beta_{2(g)} = \beta_{4(g)})$ are centered though carrying hierarchical levels per team. For instance, the attack of each gth team, $\text{att}_{\text{home}(g)}$ and $\text{att}_{\text{away}(g)}$, can be seen as a random variable distributed with mean μ_{att} and precision τ_{att}.

MODEL 2 (HIERARCHICAL-Two levels):

$$\log(\theta_{g1}) = \beta_{\text{home}} + \beta_{\text{att}(g1)}\text{att}_{(g1)} + \beta_{\text{def}(g2)}\text{def}_{(g2)}, \tag{7}$$

$$\log(\theta_{g2}) = \beta_{\text{att}(g2)}\text{att}_{(g2)} + \beta_{\text{def}(g1)}\text{def}_{(g1)}. \tag{8}$$

The Bayesian modeling technique naturally adopts a hierarchical structure, enabling us to add levels to the structure that estimate each group's influence (marginal) and their combination (global), that is, it also enables us to explain the common structure among those groups. These hierarchical models make it possible, in a natural way, to treat the levels' data, avoid overfitting, and easily combine with custom decision analysis tools [16]. Figure 1 summarizes the relationships between these parameters, and the statistical structure adopted to estimate the expected number of goals for each match.

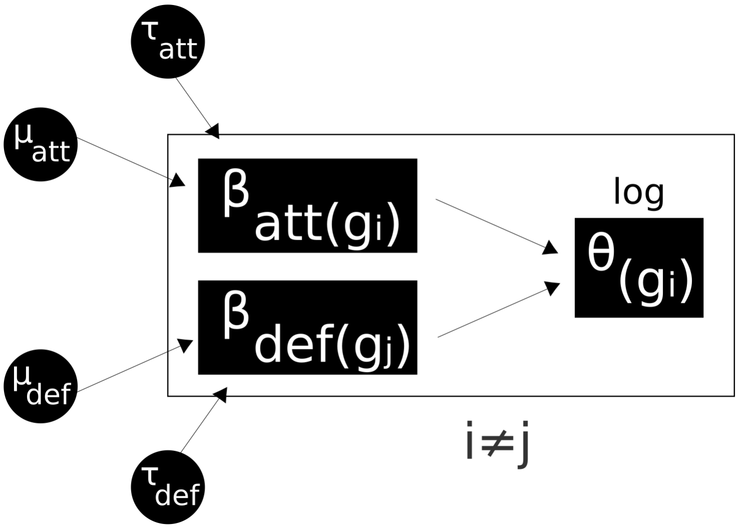

Figure 1. Visualization of the hierarchical model adopted to explain a league performance.

The term hierarchical model is addressed by models with two or more levels in their random variable (as a single contribution). Additionally, it also assumes that all abilities (attacks and defenses) are derived from a common structure (league), or are mostly very similar. Through the Bayesian approach, the experts' understandings are used as a priori information (although not a trivial task).

Hamiltonian Monte Carlo (Hmc) with Rstan

Under the Bayesian paradigm, statistical models can be calculated analytically from their posterior distributions, with reference to their parameters. However, deterministic integration methods are usually used to approximate these posterior distributions, which present a high complexity, especially under the presence of high dimensions.

Any unknown parameter (θ) is associated with a probability function ($\pi(.)$), which can be determined by Bayes' theorem:

$$\pi(\theta \mid \text{Data}) = \frac{\pi(\theta)\pi(\text{Data} \mid \theta)}{\pi(\text{Data})} = \frac{\pi(\theta)\pi(\text{Data} \mid \theta)}{\int \pi(\theta)\pi(\text{Data} \mid \theta)d\theta}. \tag{9}$$

The first study that proposed a numerical approach for the posterior distribution $\pi(\theta|\text{Data})$ relating this integration problem and using the simulation based on states of molecule systems as a solution, was proposed by Metropolis [17] and its updates are made via random walk. This technique is known as Markov chain Monte Carlo (MCMC). Consider a large independent and identically distributed sample ($T_1, ..., T_n$), in which $T_i \sim \pi(t|\text{Data})$, the posterior distribution will be the result of any function g, via the law of large numbers:

$$\mathbb{E}[g(T_i)] = \int g(t)\pi(t \mid \text{Data})dt = \int g(\theta)\pi(\theta \mid \text{Data})d\theta = \mathbb{E}[g(\theta) \mid \text{Data}]. \tag{10}$$

Later, combining the idea of numerical approximation of the posterior distribution via resampling, Alder and Wainwright [18] presented an alternative to the random walk following Newton's laws of motion as Hamiltonian dynamics, which improved the estimation of the states, originating the Monte Carlo hybrid method or the Hamiltonian Monte Carlo (HMC). In order to search for the position of the variables and obtain inferences towards θ, the method adds an auxiliary variable (φ) called "momentum". Therefore, the parameter space is explored through partial derivatives of the Hamilton's equation converging much faster. For further investigation, please consult Brooks et al. [19].

For Bayesian analysis, we used the integration of the programs R and Stan using the `Shinystan` package in R, which offers a variety of graphs and metrics tools, related to the convergence of the Monte Carlo chains, such as: η_{eff}/N, *mcse/sd*, *HMC/NUTS* and autocorrelation of each of the parameters, detailed by chain.

The metric η_{eff}/N is composed by the total sample N and the effective sample size η_{eff}, in which the latter is calculated as:

$$\eta_{eff} = \frac{n}{1 + \sum_{k=1}^{\infty} \rho_k(\theta)}, \tag{11}$$

with n being the total sample size and $\rho_k(\theta)$ is the lag k autocorrelation of parameters (θ's) (for further information, see [19]). The η_{eff}/N indicates the proportion of the parameters whose samples have effective size less than a certain percentage with respect to the total sample (regardless some significance level, α). It is also analyzed as each percentage of effective size impacting in each parameter.

We can also find the metric *mcse/sd*, in which *mcse* is the Monte Carlo standard error, which is the application of the delta method to the Monte Carlo estimates of the posterior variance [19], and *sd* is the standard deviation of the parameter's posterior. Thus, *mcse/sd* indicates the proportion of parameters whose standard errors are greater than a certain percentage with respect to the Monte Carlo estimates of the posterior variance [20].

Another common metric is the *HMC/NUTS*, in which *HMC* represents the Hamiltonian Monte Carlo, which is an MCMC method that uses the derivatives of the density function of the sample to generate efficient procedures for the parameter's posterior [20]; and

NUTS (No-U-Turn Sampler) automatically selects an $n\_leapfrog$ (number of jump steps performed during Hamiltonian simulation) at each iteration to run posteriorly without doing unnecessary work. The idea is to avoid the random walk behavior that arises in the Metropolis or Gibbs samplers when there is correlation in the posterior distribution [20].

From these metrics, the `Shinystan` package [21] generates a table considering each chain separately and together, in which, for the *HMC/NUTS*, it shows the mean, standard deviation, maximum and minimum of the following properties of the chains:

- *accept_stat*: For the HMC without NUTS, it is the Metropolis' standard acceptance probability. A value closer to one is better (robust);
- *stepsize*: The integrator used in the Hamiltonian simulation. If the value is large, it will be imprecise and reject too many proposals; and if it is small, it will take too many small steps, which will cause long simulation times per interval;
- *treedepth*: A $treedepth = 0$ means that the first jump step is immediately rejected and returns to the initial state;
- *n_leapfrog*: The number of jump steps performed during the Hamiltonian simulation. If its value is small, the sample will become a random walk; and if it is large, the algorithm will work more in each iteration;
- *divergent*: The number of jump transitions with divergent error. This number is the average of divergence at each iteration;
- *energy*: The Hamiltonian value in each sample. The energy diagnostic for HMC quantifies whether the tails of the posterior distribution are heavy or not.

3. Methodology

3.1. Data

This study adopted a hierarchical Bayesian model (presented in the previous theoretical section) for data from the 2020 Chilean soccer league, First Division, which had a record of 314 games. The main objective was to verify the admissibility of this model for predicting the statistical performance of the 2020 Chilean soccer championship.

The Chilean professional soccer league is organized by the National Professional Soccer Association from Chile. The league is divided into three categories, which for the 2020 season consisted of: 18 teams in the first division (or Premier League), 15 teams in the second category called "First B", and 12 teams in the third category called "Second Professional Division". In total, the professional soccer league in Chile is made up of 45 teams.

In the First Division or Premier League, the teams play all against each other, that is, they confront twice in a way that each team will play once at home and once as a visitor. For each game, those teams obtain points to position themselves in a table, with the first place occupied by the one with the most points. The score is as follows: three points for the winner of the match, or one point for each team in the case of a tie.

The 2020 edition of the Premier league began on 24 January 2020 and ended on 17 February 2021. It should be noted that the 2020 season was affected by the COVID-19 (Coronavirus Disease 2019) pandemic, which caused the season to end on January 2021, though it would generally end in December of the same year as it started. This pandemic affected both the organization of the event (its calendar), as well as the possibility for fans to watch their teams (up close) and to be able to support them within each team's stadium.

The weights of being a visitor or at home in a game are considered to be very important in the literature [22], thus the home advantage will be expected given the support of the team's fans rather than when it is a visitor team. Hence, the validation of statistical models previously adopted for prediction should be tested in the Chilean Premier League.

Thus, most of the data collected in soccer matches were from before the pandemic, when the fans were able to enter the soccer field and support their team. Therefore, this study intended to make the prediction of the 2020 season and check up the effectiveness of the public support.

3.2. Description of the Sports (Hierarchical) Model

The hierarchy adopted in this study was used to estimate the potential of each team from the 2020 Chilean Premier League, by centering the championship's power of attack and defense as a composition (in common) of each team. This approach considered the performance, taking into account only the number of goals scored in each match as a performance of a home team versus the visitor team. This stochastic event (number of goals) was modeled using the countable regression from the generalized linear model (GLM), which adopts the log link function for a Poisson variable distribution, presented in Section 2. The adopted priors, in model 1, were non-informative with distributions:

$$\beta_{\text{home}} \sim \text{Normal}(0, 0.0001), \tag{12}$$
$$\beta_{1(g)} \sim \text{Normal}(0, 0.0001), \tag{13}$$
$$\beta_{2(g)} \sim \text{Normal}(0, 0.0001), \tag{14}$$
$$\beta_{3(g)} \sim \text{Normal}(0, 0.0001), \tag{15}$$
$$\beta_{4(g)} \sim \text{Normal}(0, 0.0001). \tag{16}$$

In our hierarchical model, it is shown as a regression structure in which the parameters $(\theta_{g1}, \theta_{g2})$ are associated with the attack capacity (att) and defense (def) of each team. Moreover, all teams share a common performance that composes the 2020 Chilean championship. In addition, the league has an attack (μ_{att}) and defense (μ_{def}) strength. That is, each gth team is described as:

$$\beta_{\text{home}} \sim \text{Normal}(0, 0.0001), \tag{17}$$
$$\beta_{\text{att}(g)} \sim \text{Normal}(\mu_{\text{att}}, \tau_{\text{att}}), \tag{18}$$
$$\beta_{\text{def}(g)} \sim \text{Normal}(\mu_{\text{def}}, \tau_{\text{def}}). \tag{19}$$

Therefore, each team will compose the estimation of the four parameters, which correspond to: two parameters for the attack (att_g), observed when the team is visiting or at home, in the same way as the two parameters associated with the defense (def_g).

The total number of attack parameters is 18, related to the total number of teams in the league. These parameters are distributed around the mean μ_{att} with the dispersion of τ_{att}. The structure is obtained given the hierarchical model, since it considers the performance of each individual team, and their behavior composes the Chilean Premier League additionally as a higher level. Then, the same reasoning was applied for the defense ability of each team. All the priors adopted were non-informative with distributions:

$$\mu_{\text{att}} \sim \text{Normal}(0, 0.0001), \tag{20}$$
$$\tau_{\text{att}} \sim \text{Gamma}(0.1, 0.1), \tag{21}$$
$$\mu_{\text{def}} \sim \text{Normal}(0, 0.0001), \tag{22}$$
$$\tau_{\text{def}} \sim \text{Gamma}(0.1, 0.1). \tag{23}$$

In the following section, we present the structures of the adopted championship (2020 Chilean First Division league), and the results obtained from the proposed Bayesian hierarchical modeling.

4. Results

The 2020 Chilean First Division league is made up of 18 teams, which play against each other twice in a season (one at home and one away). This study took out the 34 rounds that occurred from 24 January 2020 to 17 February 2021, and the last five games were saved for validation. In the last rounds, 314 games were played, and the summarization of the scored goals through this championship is described, per team, in Figure 2.

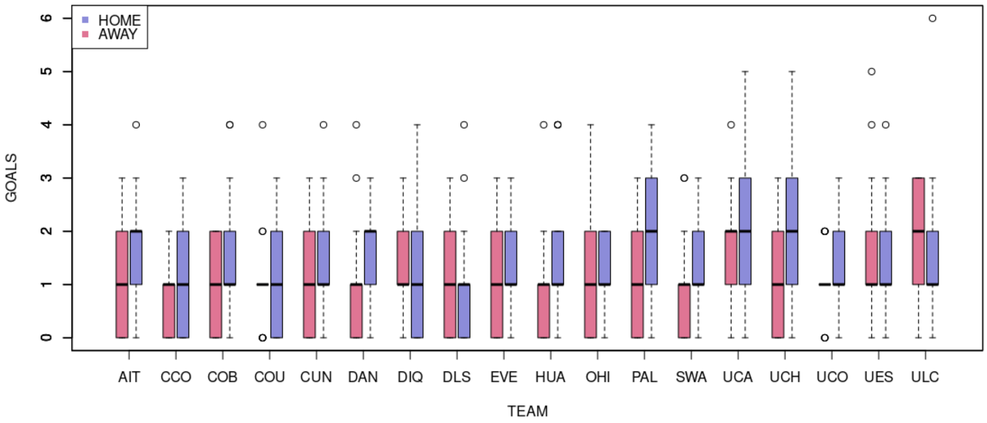

Figure 2. Teams' performance through the 2020 Chilean Premier Soccer League. Despite the group being a home player or visitor, most of the teams' performance was on a single goal (on median, or 50% of the time). At home, the teams with better performance were PAL, UCA, UCH, DAN, and AIT. As visitors, better performances were ULC and UCA teams.

Considering the theoretical model of Poisson regression, we take as the response variable the number of goals scored by each team, explained by a home game factor added to their attack power and defense. The `Shinystan` package containing graphical and numerical summaries of parameters was adopted for the model and convergence diagnostics, and implemented in the R software for the analysis of the parameters with the HMC algorithm of strings with the size being 2000 iterations, with warmup = 1000 and thin = 1, with four independent chains. All statistical analyses of this study consider a credibility level of 50% (as statistical significance).

Both theoretical models, presented in Section 2, were implemented and adopted the leave-one-out (LOO) cross-validation with moment matching aiming the model selection. The *loo* package was adopted, based on the generated quantities from the STAN codes [23]. Table 1 shows that the *loo* results across the models are similar, and using the loo model comparison (function *loo_compare*), the expected log pointwise predictive density (*elpd*) difference between model 1 versus model 2 was −0.1 with a standard error (SE) of 0.6 (not differing statistically). Therefore, we chose model 2 because it contains fewer parameters.

Table 1. Model selection using leave-one-out cross-validation with moment matching.

	Statistic	Estimate	SE
Model 1 (Non-hierarchical)	elpd_loo	−15.4	2.3
	p_loo	0.5	0.3
	looic	30.7	4.7
Model 2 (Hierarchical)	elpd_loo	−15.4	2.1
	p_loo	0.2	0.1
	looic	30.8	4.1

In Table 2, we see, in detail, the descriptive statistics of the posterior distributions of the parameters of model 2 (hierarchical), which describe the expected scoring of the teams resulting in the final positions for the championship of 2020 (in which the last rounds will serve as a comparison and were not considered in the estimation process). It is possible to observe that the expected results of the three teams with the greatest attack power, foreseen by the adopted model, are in the first four positions of the final result table of the championship. Due to the model adopted, they are in the first four positions of the final result table of the championship. It should be noted that it was expected by the model that the other team in the top four is the Universidad de Chile (UCH), which we can place as having the fourth best attacking position in the graph of Figure 3.

Table 2. Betas Maximum a Posteriori (MAPs) summary estimation for theoretical model 2.

	Posterior Mean	**Posterior Quantile 25%**	**Posterior Quantile 75%**
HOME	1.37×10^8	-5.96×10^9	6.45×10^9
UCA_att	**0.4999**	**0.4071**	**0.5947**
ULC_att	**0.4185**	**0.3278**	**0.5124**
UES_att	**0.3616**	**0.2715**	**0.4532**
UCH_att	**0.2636**	**0.1724**	**0.3557**
PAL_att	**0.2535**	**0.1608**	**0.3487**
AIT_att	**0.2208**	**0.1292**	**0.3168**
COB_att	**0.1965**	**0.1087**	**0.2902**
HUA_att	**0.183**	**0.0868**	**0.2799**
DAN_att	**0.1583**	**0.0623**	**0.2573**
SWA_att	**0.142**	**0.0482**	**0.2429**
OHI_att	**0.1107**	**0.0162**	**0.2067**
CUN_att	**0.1061**	**0.0184**	**0.1972**
DIQ_att	0.0711	−0.0231	0.1681
UCO_att	0.0687	−0.024	0.1657
EVE_att	0.0519	−0.0445	0.1513
DLS_att	0.0082	−0.0872	0.1085
CCO_att	−0.0311	−0.1284	0.071
COU_att	−0.059	−0.1584	0.0434
UCO_def	−0.0105	−0.0136	0.0055
ULC_def	−0.0042	−0.0091	0.0066
COU_def	−0.0097	−0.0134	0.0051
COB_def	−0.0013	−0.0086	0.0085
UCA_def	0.0084	−0.0051	0.0134
UCH_def	0.0104	−0.0042	0.0137
DIQ_def	−0.0114	−0.0143	0.0044
DLS_def	−0.0002	−0.0085	0.0085
PAL_def	−0.008	−0.0121	0.0059
UES_def	−0.0254	−0.0249	0.0022
SWA_def	−0.0249	−0.0242	0.0023
HUA_def	−0.0077	−0.0117	0.0057
CUN_def	−0.0235	−0.0213	0.0025
CCO_def	−0.0051	−0.0104	0.0069
OHI_def	0.0014	−0.007	0.0086
EVE_def	−0.0058	−0.0096	0.0065
DAN_def	−0.0044	−0.0095	0.0068
AIT_def	−0.0219	−0.0202	0.0024
μ_{att}	$1.12E \times 10^8$	-6.56×10^8	6.71×10^09
τ_{att}	0.2729	0.2288	0.311
μ_{def}	1.08×10^8	-6.68×10^9	6.84×10^9
τ_{def}	0.0343	0.0047	0.0452

In Figure 3, we are able to see the prediction of attack and defense of the 18 teams in the league, in which we can observe that the level of the defense for each team individually was statistically equal to zero, which means that there is no significant result in the defense among the teams. However, if it was obtained that the attack of the teams, such as the Universidad Católica (UCA), Unión La Calera (ULC) and Unión Española (UES), have greater attack power according to the model. In addition, the prediction error of the model is displayed, with regards to the attack prediction of the teams, that is, due to the size of the attack it is possible to present the championship positions, according to the adjusted Poisson regression model.

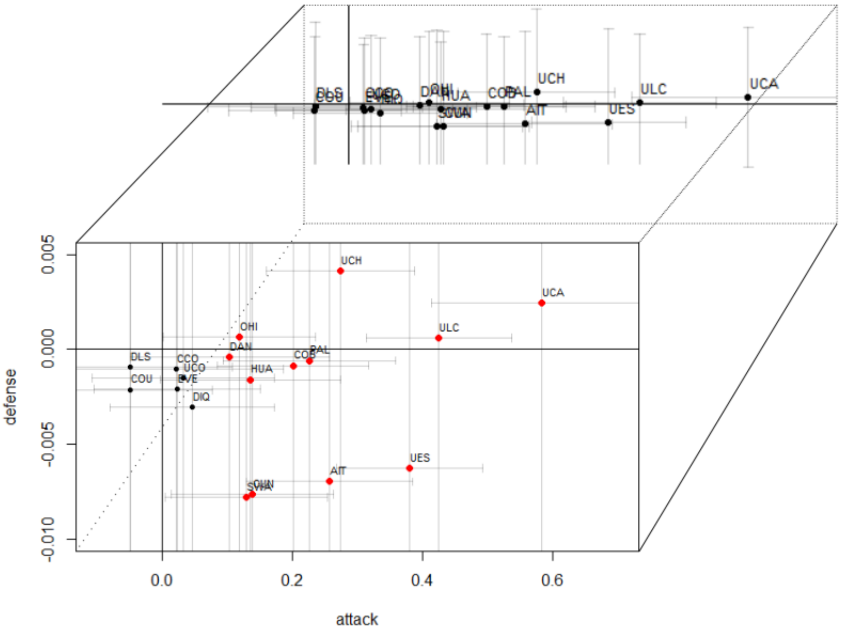

Figure 3. Prediction of the Chile 2020 championship, according to the adjusted model, saving the last 5 rounds of 2020 games. The points in red are the teams that had statistical significance, being only their attacking power significant.

Additionally, from observation of Figure 4, it is possible to notice the attack average of each team, and we can see that the first five teams with the highest attacking average are the same first five teams who have been better ranked at the end of the season, and positions 5 to 11 are also the same teams although not in the same position.

In Table 2, in addition to the 11 teams that have a significant statistic, there is also the Audax Italiano, which has a significant value and is in position 6 of the prediction table, whereas it is in position 14 of the final table of the 2020 championship. However, its defense is one of the lowest, thus we can interpret this as an outlier.

Although most of the teams are between positions 11 and 18 of the final table of the championship, their attack and defense prediction errors are high compared to the other 11, which may lead to thinking that this is a factor that alters the program.

On the other hand, we can see in Figure 5 how the *Home* parameter is distributed, which represents the attack and defense of each team at home, and looks symmetrically distributed. In this way, we can say that the *Home* parameter did not intervene in the definition of a match in the 2020 Chilean Premier League.

Figure 6 shows four HMC chains of the *Home* parameter, in which in order to determine convergence it is convenient to observe that the four chains are consistent and to check that the chains are not stuck in unusual regions. Therefore, we can say that the chains converge.

We can also observe the deviation represented by τ_{def} and τ_{att}. For the case of defense, we have an asymmetry that is close to zero. On the other hand, for attack, we have an average of the dispersion between 0.2 and 0.3.

The use of the `Shinystan` package facilitates the analysis of the results due to the intuitive and easy interaction of this package, and by adopting the *launch_shinystan* function, a parsing interface related to the strings from HMC, you will be able to view various graphs and parameter estimates in more detail.

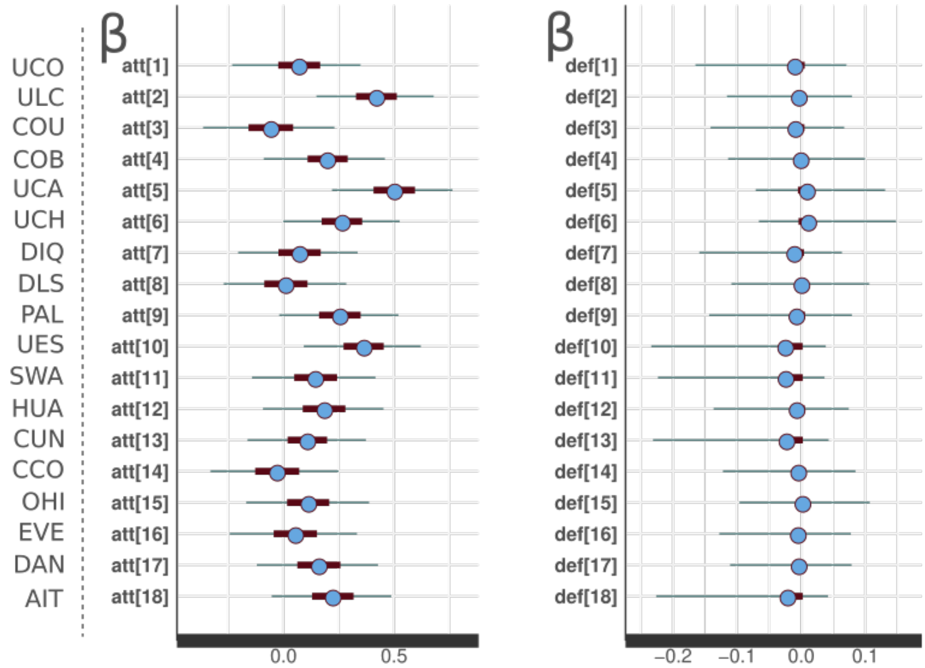

Figure 4. Betas Maximum a Posteriori (MAPs) probability estimates and 50% credibility intervals (50% CIs) for the attack parameters on the left, and the defense parameters on the right for each team.

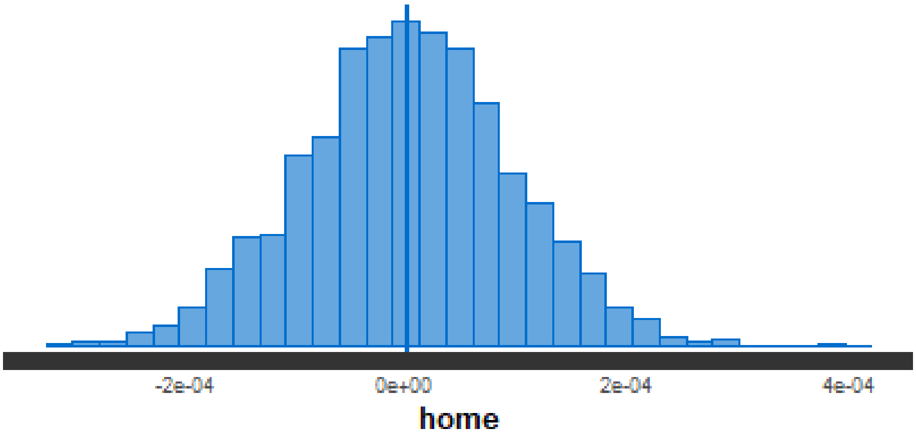

Figure 5. A posteriori distribution of the *Home* parameter for the hierarchical model.

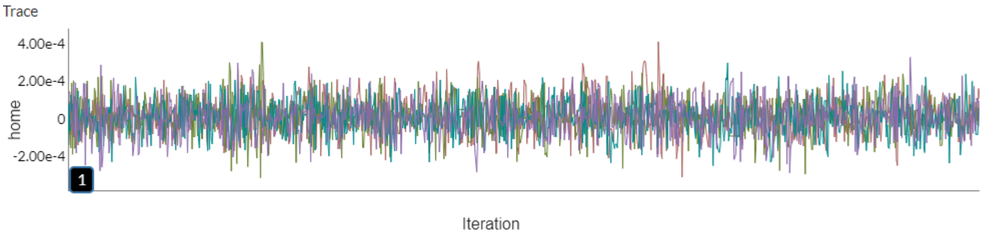

Figure 6. HMC chains of the *Home* parameter.

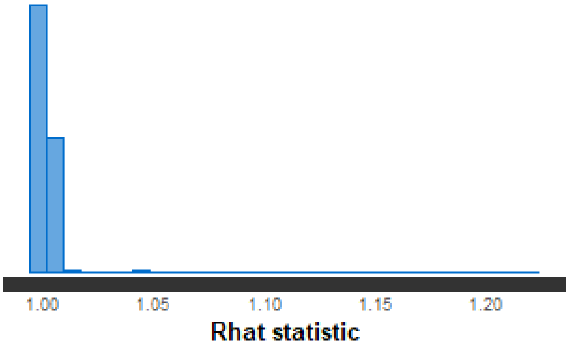

Figure 7. R statistics of the HMC chains of the *Home* parameter.

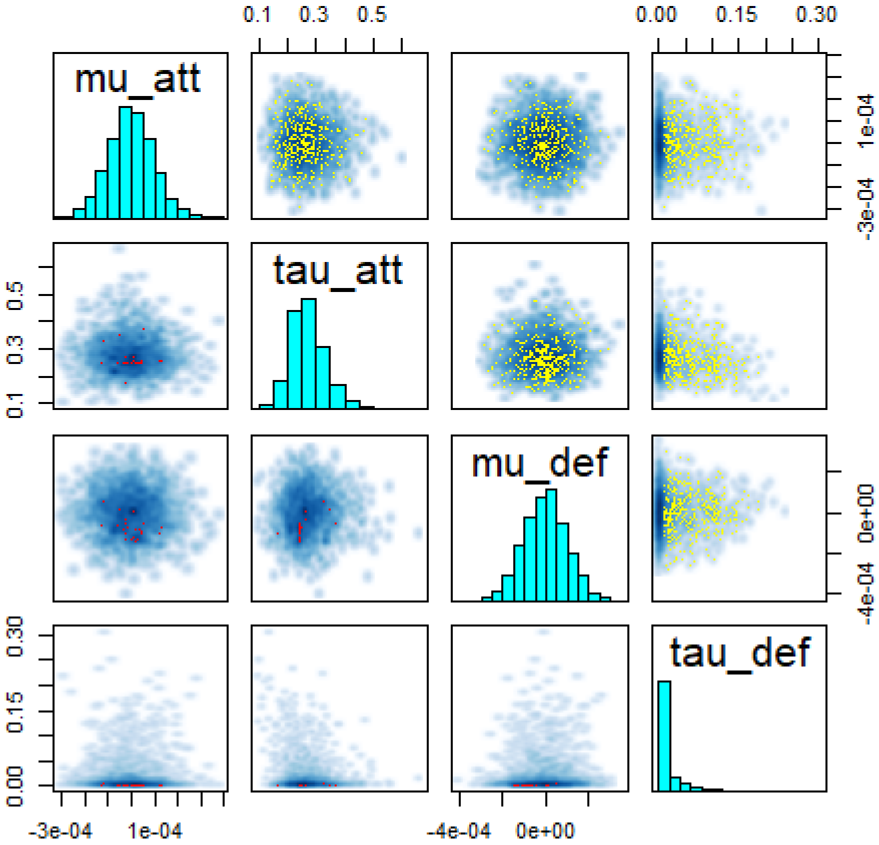

Figure 8. Distribution of defense and attack parameters in general in the Chilean league (μ and τ). The main diagonal shows the histograms of the posterior distributions of the offense and defense of the 2020 Chilean First Division league. The upper triangular part shows evidence of independence, from the HMC, among the parameter estimates. The lower triangular part shows results associated with the bivariate distributions of the predicted posteriors.

One of the first visualizations would be the HMC chains adopted from the model, also allowing a separate or simultaneous display on the same graph. The HMC strings, approximated by the NUTS algorithm, are computationally more expensive than Metropolis or Gibbs. However, they are more efficient, and therefore the samples can be small [24].

The convergence of the chains is intuitively represented when targeting the same region, that is, there is a behavior similar to a convergence point (value), and the variance between the chains is approximately equal to the mean variance within the chains (homoscedastic), also quantified by the estimated *R_hat* ($\hat{R}$) statistic, which is close to 1 [24]. As an example of convergence of the *Home* parameter, Figure 6 demonstrates the obtained strings in which each color refers to different strings, and the beginning of each chain was randomly selected from the Normal (0, 0.01) distribution. In addition, the chains look visually stationary achieving the convergence measure, showing just a random noise around the mode of the posterior for the *Home* parameter, confirmed by the statistics of the $\hat{R}$ value concentrated at 1 (less than the threshold 1.01), as shown by Figure 7.

`Shinystan` also makes it possible to view all the posteriors of the parameters adopted by the statistical model, enabling a better understanding of what is being analyzed and, in the case of seeing any anomaly, to be able to correct it or to do some other test. For each parameter that intervenes in the model, a graph of the posterior distribution is generated, and, as shown in Figure 8, the posterior distributions for the μ_{def}, τ_{def}, μ_{att} and τ_{att} parameters obtained by the HMC strings. From these graphs, we could see that the *Home* parameter is distributed in a symmetric way (in Figure 5), and that the μ_{def} and μ_{att} parameters have slight skewness (in Figure 8). The precision parameters τ_{def} and τ_{att} are involved with attack parameter dispersion and suggest a significance, compared to defense, since the distribution of τ_{def} is around zero.

In view of their empirical distributions, Figure 8 shows that, in the upper triangle, there is no trend in the posterior parameters, which makes them look unrelated (μ_{att}, τ_{att}, μ_{def} and τ_{def}). It is also important to highlight that the parameter (μ_{att}) can approximate a discussion related to the attacking average of the Chilean league (μ_{def}) and the defense average in a way that represents the championship as the main one in Chile. Observing the figure, it is in accordance with the previous one since the defense average is zero, that is, it does not contribute to the prediction. However, the attack average is non-zero (specifically for some teams being statistically significant).

In the lower triangular part of Figure 8, we see the predictive posterior of the parameters involved, from which we see that it adjusts to the expected values.

It can also be obtained individually, that is, per team, associated with the posterior distributions of defense and attack. This is an important aspect if you want to observe the performance of only one computer (obtained by adopting the hierarchical model).

Some of the metrics that were extracted from the `Shinystan` interface are presented in Table 3, in which the *accept_stat* metric validates the convergence of each of the strings in a general way, since there are values close to one.

Table 3. Metrics provided by the `Shinystan` parameter on chain convergence for the hierarchical model.

Chain	*Accept_stat*	*Stepsize*	*Treedepth*	*N_deapfrog*	*Divergent*	*Energy*
All chains	0.7922	0.0265	7.0215	158.1773	0.0158	494.9190
Chain 1	0.7909	0.0224	7.4130	218.1270	0.0060	498.7200
Chain 2	0.7583	0.0276	6.9330	125.8300	0.0050	494.4215
Chain 3	0.7272	0.0260	6.6970	145.0860	0.0510	463.3970
Chain 4	0.8924	0.0297	7.0430	143.6660	0.0010	523.1376

The *stepsize* values are small. Therefore, the program performed long simulation times for the interval. In the case of *treedepth*, the values are different from zero and it does not hit the maximum value, which is shown to be an efficient No-U-Turn-Sampler.

The metrics of n_{eff}/N and $mcse/sd$ were also extracted, indicating that the following parameters: def_1, def_6, def_7, def_{10}, def_{11}, def_{13}, def_{18}, τ_{def}, log-posterior, have an effective sample size less than 10% of the total sample size; and that the following parameters: def_{10}, def_{11}, def_{13}, def_{18}, τ_{def}, have a Monte Carlo standard error greater than 10% of the posterior standard deviation. Finally, for the development of this research, the autocorrelation metric was necessary, which shows the relationship of the parameters in each of the strings, and after their analysis, all chains presented signals of convergence.

Our models predicted correctly most of the Chilean final ranking positions, using the last five matches for prediction. Table 4 shows the median prediction, generated from the adjusted Poisson derived from the four independent chains which generated quantities blocks, based on both adopted models (hierarchical and non-hierarchical) which presented quite similar results.

Table 4. Median predicted goals, from the 4 chains, adopting the last 5 games of the 2020 Chilean championship.

MATCH	HOME			AWAY		
HOME × AWAY	GOAL	MODEL 1	MODEL 2	GOAL	MODEL 1	MODEL 2
DLS × AIT	0	1	1	2	1	1
UCO × UCA	1	1	1	2	2	1
PAL × COU	2	1	1	2	1	1
EVE × HUA	1	1	1	0	1	1
OHI × CCO	1	1	1	1	1	1

For example, the first five positions predicted by the model are: Universidad Católica (UCA), Unión La Calera (ULC), Unión Española (UES), Universidad de Chile (UCH) and Palestino (PAL), in an order based on a calculated attack, whereas the general table remained: Universidad Católica (UCA), Unión La Calera (ULC), Universidad de Chile (UCH), Unión Española (UES) and Palestino (PAL), which, in the final table, is affected by the league rules that indicate that, at the time of equalizing the points, the position is defined by the goal difference, and since our model is based on the goals scored, it will show the one that scores more goals at the top of the table. The highest one in the table is the one that scores more goals. Therefore, by organising the attack value of the teams in decreasing order, and comparing with the final table of the league (Table 5), our model manages to predict most of the positions exactly when it comes to the highlighted teams.

On the other hand, positions 7 to 12 of Table 2, according to the estimates of the Poisson model, correspond to positions 6 to 11 of the final table of positions for the Chilean league (Table 5). It happens that, since the difference between one position and the other is one point, then the positions predicted by model 2 were not equivalent to the final result.

Table 5. Final standings of the 2020 Chilean Premier League. The positions highlighted, in bold style, were predicted correctly by the adjusted Bayesian hierarchical model.

Position	Team Name	Points
1	**Universidad Católica (UCA)**	65
2	**Unión La Calera (ULC)**	57
3	**Universidad de Chile (UCH)**	52
4	**Unión Española (UES)**	52
5	**Palestino (PAL)**	51
6	**Deportes Antofagasta (DAN)**	48
7	**Cobresal (COB)**	47
8	**Huachipato (HUA)**	46
9	**Curicó Unido (CUN)**	46
10	**O´Higgins (OHI)**	45
11	**Santiago Wanderers (SWA)**	44
12	Everton (EVE)	43
13	Universidad de Concepción (UCO)	41
14	**Audax Italiano (AIT)**	41
15	Deportes La Serena (DLS)	39
16	Colo Colo (CCO)	39
17	Deportes Iquique (DIQ)	38
18	Coquimbo Unido (COU)	35

5. Conclusions

The Bayesian hierarchical structure modeling adopted in this study was shown to be adequate for describing the Chilean Premier League, with the attack and defense power of each team, but enabled the analyses of the entire Chilean Premier League (attack and defense power). In the 2020 Chilean Premier League, the definition of the championship, based on the model 2, depends on the power of the team to attack (β_{att}) rather than its defense (β_{def}), statistically. This attribute has been estimated considering both individual and group characteristics, allowed by the hierarchical modeling. Despite the home advantage debated in the literature [22], this study did not find any statistical significance of this parameter.

In this sense, our adopted model attended the goals of each team for determining its defense and attack potential, and enabled the determination of 11 of the 18 positions of the final Chilean premier championship ranking. Moreover, this study took into account a simple regression structure model. Thus, many other aspects were not considered (and could also be tested), such as the geographical region where the game is played, the performance of players, and their nationalities. Further works may consider three levels of hierarchy by adding the players' layer. This type of model has proved to be very competitive for the prediction of results in championships, in which all teams play against each other in a tournament, with no need to predict the uncertainty on the elimination due to each phase. Another possibility is to adopt the championship as a dynamic process and, through time-varying parameters, accommodate the league's evolution round-by-round [25].

For instance, such modeling can be used to assess the decision-making for the team's strategic office and coach, unraveling their weaknesses and strengths, thus supporting, through quantified inferences, the targeting of improved technical skills. This modeling can also be designed for betting on the winning team (expected score).

Author Contributions: Conceptualization, project administration, methodology and funding acquisition, D.C.d.N.; software, data curation and formal analysis, L.B.B.; writing—original draft preparation, D.C.d.N. and L.B.B.; writing—review and editing, P.H.F. and F.L. All authors have read and agreed to the published version of the manuscript.

Funding: Francisco Louzada acknowledges support from the Brazilian institutions FAPESP and CNPq.

Data Availability Statement: All the programming scripts and data set are fully available at https://github.com/ProfNascimento/GLM_Chilean_Soccer_League.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

HMC	Hamiltonian Monte Carlo
MCMC	Markov chain Monte Carlo
NUTS	No-U-Turn Sampler
MAP	Maximum a Posteriori
CIs	Credibility Intervals
GLM	Generalized Linear Model
LOO	leave-one-out
elpd_loo	LOO expected log pointwise predictive density
p_loo	LOO effective number of parameters
looic	LOO Information Criterion
SE	Standard Error
att	attack capacity
def	defense capacity
θ	Poisson parameter related to the expected number of events of each team
β_{home}	Home advantage parameter for the team playing at home
β_1 and β_3	MODEL 1—Attack ability parameters from home and away
β_2 and β_4	MODEL 1—Defense ability parameters from home and away
β_{att}	MODEL 2—Attack ability parameter (for each team)

β_{def}	MODEL 2—Defense ability parameter (for each team)
μ_{att}	MODEL 2—2020 Chilean Premier attack (mean) location parameter
μ_{def}	MODEL 2—2020 Chilean Premier defense (mean) location parameter
τ_{att}	MODEL 2—2020 Chilean Premier attack variability parameter
τ_{def}	MODEL 2—2020 Chilean Premier defense variability parameter
UCA	Universidad Católica team
ULC	Unión La Calera team
UCH	Universidad de Chile team
UES	Unión Española team
PAL	Palestino team
DAN	Deportes Antofagasta team
COB	Cobresal team
HUA	Huachipato team
CUN	Curicó Unido team
OHI	O´Higgins team
SWA	Santiago Wanderers team
EVE	Everton team
UCO	Universidad de Concepción team
AIT	Audax Italiano team
DLS	Deportes La Serena team
CCO	Colo Colo team
DIQ	Deportes Iquique team
COU	Coquimbo Unido team

References

1. Radicchi, E.; Mozzachiodi, M. Social talent scouting: A new opportunity for the identification of football players? *Phys. Cult. Sport* **2016**, *70*, 28. [CrossRef]
2. Schumaker, R.P.; Solieman, O.K.; Chen, H. *Sports Data Mining*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2010; Volume 26.
3. Morgulev, E.; Azar, O.H.; Lidor, R. Sports analytics and the big-data era. *Int. J. Data Sci. Anal.* **2018**, *5*, 213–222. [CrossRef]
4. Beal, R.; Norman, T.J.; Ramchurn, S.D. Artificial intelligence for team sports: A survey. *Knowl. Eng. Rev.* **2019**, *34*, e28. [CrossRef]
5. Louzada, F.; Maiorano, A.C.; Ara, A. iSports: A web-oriented expert system for talent identification in soccer. *Expert Syst. Appl.* **2016**, *44*, 400–412. [CrossRef]
6. Santos-Fernandez, E.; Mengersen, K.L.; Wu, P. Bayesian methods in sport statistics. In *Wiley StatsRef: Statistics Reference Online*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2019; pp. 1–8.
7. Anderson, C.; Sally, D. *The Numbers Game: Why Everything You Know about Soccer Is Wrong*; Penguin: London, UK, 2013.
8. Baio, G.; Blangiardo, M. Bayesian hierarchical model for the prediction of football results. *J. Appl. Stat.* **2010**, *37*, 253–264. [CrossRef]
9. Lee, A.J. Modeling scores in the Premier League: Is Manchester United really the best? *Chance* **1997**, *10*, 15–19. [CrossRef]
10. Suzuki, A.K.; Salasar, L.E.B.; Leite, J.; Louzada-Neto, F. A Bayesian approach for predicting match outcomes: The 2006 (Association) Football World Cup. *J. Oper. Res. Soc.* **2010**, *61*, 1530–1539. [CrossRef]
11. Santana, H.; Ferreira, P.H.; Ara, A.; Louzada, F.; Suzuki, A.K. Modelagem Estatística e de Aprendizado de Máquina: Previsão do Campeonato Brasileiro Série A 2017. *MatemáTica EstatíStica Foco* **2019**, *7*, 42-a.
12. Constantinou, A.C.; Fenton, N.E.; Neil, M. pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowl.-Based Syst.* **2012**, *36*, 322–339. [CrossRef]
13. Hervert-Escobar, L.; Hernandez-Gress, N.; Matis, T.I. Bayesian based approach learning for outcome prediction of soccer matches. In *International Conference on Computational Science*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 269–279.
14. Poisson, S.D. *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile*; Bachelier: Cambridge, MA, USA, 1837.
15. Gelade, G.A. The influence of team composition on attacking and defending in football. *J. Sport. Econ.* **2018**, *19*, 1174–1190. [CrossRef]
16. Moreno, E.; Martínez, C. Bayesian and frequentist evidence in one-sided hypothesis testing. In *TEST*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 1–20.
17. Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092. [CrossRef]
18. Alder, B.J.; Wainwright, T.E. Studies in molecular dynamics. I. General method. *J. Chem. Phys.* **1959**, *31*, 459–466. [CrossRef]
19. Brooks, S.; Gelman, A.; Jones, G.; Meng, X.L. *Handbook of Markov Chain Monte Carlo*; CRC Press: Boca Raton, FL, USA, 2011.
20. Betancourt, M. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv* **2017**, arXiv:1701.02434.

21. Gabry, J.; Simpson, D.; Vehtari, A.; Betancourt, M.; Gelman, A. Visualization in Bayesian workflow. *J. R. Stat. Soc. Ser. Stat. Soc.* **2019**, *182*, 389–402. [CrossRef]
22. CJ, D.; Chakravarty, A. Team Contingent or Sport Native? A Bayesian Analysis of Home Field Advantage in Professional Soccer. *J. Bus. Anal.* **2021**, *4*, 67–75.
23. Vehtari, A.; Gelman, A.; Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **2017**, *27*, 1413–1432. [CrossRef]
24. Muth, C.; Oravecz, Z.; Gabry, J. User-friendly Bayesian regression modeling: A tutorial with rstanarm and shinystan. *Quant. Methods Psychol.* **2018**, *14*, 99–119. [CrossRef]
25. Crowder, M.; Dixon, M.; Ledford, A.; Robinson, M. Dynamic modelling and prediction of English Football League matches for betting. *J. R. Stat. Soc. Ser. Stat.* **2002**, *51*, 157–168. [CrossRef]