Lucas Mendes Gomes

Futebol, análise e previsão de seus resultados

Niterói - RJ, Brasil

Universidade Federal Fluminense

Lucas Mendes Gomes

Futebol, análise e previsão de seus resultados

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador: Prof. Luis Guillermo Coca Velarde

Niterói - RJ, Brasil

Universidade Federal Fluminense

Lucas Mendes Gomes

Futebol, análise e previsão de seus resultados

Monografia de Projeto Final de Graduação sob o título *"Futebol, análise e previsão de seus resultados"*, defendida por Lucas Mendes Gomes e aprovada em , na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Prof. Dr. Luis Guillermo Coca Velarde
Departamento de Estatística – UFF

Prof. Dr. Hugo Henrique Kegler dos Santos
Departamento de Estatística UFF

Prof. Dr. Licinio Esmeraldo da Silva
Departamento de Estatística UFF

Niterói, 12 de dezembro

G633 Gomes, Lucas Mendes
Futebol, análise e previsão de seus resultados / Lucas Mendes Gomes. – Niterói, RJ: [s.n.], 2017.

49f.

Orientador: Prof. Dr. Luis Guilhermo Coca Velarde
TCC (Graduação de Bacharelado em Estatística) – Universidade Federal Fluminense, 2017.

1.Modelos Lineares Bayesianos . 2. Futebol. . I. Título.

CDD 519.2

Resumo

Ao longo dos últimos anos o campo de estudo para a área esportiva vêm crescendo bastante no Brasil. Este trabalho se propõe a estudar um modelo que auxilie na previsão da quantidade de gols marcados por cada time no campeonato brasileiro de futebol. Para isto será utilizado o programa OpenBUGS e serão testados diferentes modelos utilizando modelos lineares generalizados com efeitos aleatórios que representam o ataque e a defesa de cada time, para verificar qual possui a melhor previsão para os dados do campeonato brasileiro de 2016. Foram apresentados 5 propostas de modelo que conseguiram capturar características do desempenho dos times, por exemplo, fator ataque.

Palavras-chaves: Modelos Lineares Bayesianos, Futebol

Agradecimentos

- Primeiramente gostaria de agradecer ao Professor Luis Guillermo Coca Velarde que disponibilizou tempo e conhecimento para a realização deste trabalho
- Agradecer também aos demais professores do colegiado que transmitiram seus conhecimentos ao longos destes anos da graduação auxiliando dando base para este momento
- Aos meus amigos e familiares que me ajudaram ao longo do caminho
- E aos meus pais que sempre me incentivaram

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução p. 10

2 Objetivos p. 12

3 Materiais e métodos p. 13

3.1 Dados . . . . . p. 13

3.2 Metodologia . . . . . p. 15

3.2.1 Modelos lineares generalizados . . . . . p. 16

3.2.2 Proposta de modelo . . . . . p. 17

3.2.3 Estimação de parâmetros . . . . . p. 19

3.2.3.1 Métodos de Monte Carlo em Cadeias de Markov (MCMC) p. 20

3.2.3.2 Escolha de modelo . . . . . p. 21

4 Análise e discussão dos Resultados p. 23

4.1 Modelo 1 . . . . . p. 29

4.2 Modelo 2 . . . . . p. 31

4.3 Modelo 3 . . . . . p. 34

4.4 Modelo 4 . . . . . p. 35

4.5 Modelo 5 . . . . . p. 36

4.6 Algumas observações e discussões sobre os modelos . . . . . p. 41

5 Conclusão p. 45

Referências p. 46

Anexo A – Scripts do OpenBUGS para os modelos utilizados p. 47

Lista de Figuras

1 Gráfico pontuação rodada a rodada . . . . . p. 14

2 Gráfico do total de gols por rodada . . . . . p. 25

3 Gráfico gols feitos acumulados rodada a rodada . . . . . p. 26

4 Gráfico da distribuição de vitórias como mandante e visitante . . . . . p. 27

5 Gráfico da distribuição dos gols no campeonato . . . . . p. 28

6 Gráfico da distribuição dos gols no campeonato separado pelo mando de campo . . . . . p. 28

7 Distribuição a posteriori de $\beta 0$ e $\beta 1$ . . . . . p. 30

8 Cadeias de $\beta 0$ e $\beta 1$ . . . . . p. 30

9 Gráficos do $\beta 2$ . . . . . p. 32

10 Gráficos do $\beta 3$ . . . . . p. 32

11 Cadeias do $\beta 2$ . . . . . p. 33

12 Cadeias do $\beta 3$ . . . . . p. 33

13 Gráficos da previsão para o Palmeiras . . . . . p. 43

14 Gráficos da previsão para o América-MG . . . . . p. 44

Lista de Tabelas

1 Tabela do campeonato . . . . . p. 15

2 Tabela com algumas estatísticas descritivas do campeonato . . . . . p. 23

3 Estatísticas para o Modelo 1 . . . . . p. 30

4 Modelo 2 . . . . . p. 33

5 Modelo 3 . . . . . p. 34

6 Modelo 4 . . . . . p. 35

7 Modelo 5 . . . . . p. 36

8 Resultados do parâmetro $\beta 2$ . . . . . p. 38

9 Resultados do parâmetro $\beta 3$ . . . . . p. 39

10 Resultados do parâmetro $\beta 4$ . . . . . p. 40

11 Função desvio . . . . . p. 43

1 Introdução

As ideias nesta seção foram retiradas de Farias (Farias, 2008) [3], o futebol é considerado, em muitos países, o esporte mais apaixonante e com o maior número de adeptos e espectadores do mundo, historicamente, ao redor do mundo, as equipes ditas "grandes" costumam alternar entre elas o status de campeão nacional e muito raramente estas equipes são rebaixadas para divisões inferiores de seus países. Na maioria dos países existem poucas equipes intituladas "grandes", porém o Brasil foge desse padrão por ter 17 campeões nacionais e diferentes e o título de time "grande" é atribuído a 12 destes times, o que torna o campeonato brasileiro um dos mais disputados e imprevisíveis.

Ao longo dos anos, o futebol se tornou alvo de especialistas que buscam nos números uma maneira de explicar e prever os resultados, aumentando cada vez mais a disponibilidade de dados para análise, porém, dentre os especialistas da área esportiva, o futebol é dito o esporte com menor chance de se prever um resultado pois é onde a chance do mais fraco vencer o mais forte se maximiza em relação a qualquer outro esporte.

Desde o ponto de vista estatístico, alguns trabalhos que utilizam o modelo Linear Dinâmico foram desenvolvidos na área esportiva, como por exemplo, os de Rue e Salvessen (Rue e Salvessen, 2000) [7], Souza e Gamerman (Souza e Gamerman, 2004) [8] e Knorr-Held (Knor-Held, 1997) [6].

Uma temporada de futebol têm cerca de 10 meses de duração e ao longo de período, são jogados diversos campeonatos (entre 2 e 4 campeonatos) e ao longo de uma semana um time pode jogar até 3 vezes, o que aumenta o desgaste do elenco e aumenta a chance de lesões. Esses não são os únicos fatores a influenciarem o desempenho de um time durante uma temporada, outros fatores como troca de técnicos, presença da torcida no estádio, problemas políticos no clube, problemas internos no elenco e atraso de salários,

este último comum no Brasil, são problemas que afetam o dia-a-dia dos clubes e ao final de uma temporada se mostram de grande importância para as pretensões dos clubes. Com os fatores acima citados, é de se supor que ao longo de uma temporada um time oscile entre bons e maus momentos.

Segundo a Confederação Brasileira de Futebol a pontuação do campeonato consiste em 3 pontos em caso de vitória, 1 ponto em caso de empate e 0 pontos em caso de derrota, e caso ocorra empate em número de pontos o desempate ocorrerá seguindo os seguintes critérios:

- Maior número de vitórias
- Maior saldo de gols
- Maior número de gols feitos
- Confronto direto
- Menor número de cartões vermelhos
- Menor número de cartões amarelos
- Sorteio

Como o que define uma partida é o número de gols que os times fazem, é intuitivo pensar que modelando-se o número de gols tenha-se uma melhor previsão dos resultados do campeonato. Para auxiliar a modelar o número de gols feitos pelos times alguns dados poderiam ser levados em conta como o mando de campo, já que times que jogam em casa tendem a ser melhor sucedidos, a posição que os times ocupam na tabela no momento do confronto, já que times melhor posicionados tendem a ganhar de times em pior colocação e informações sobre as rodadas anteriores.

2 Objetivos

- Propor um modelo que relacione o número de gols feitos por uma equipe com diversos fatores, como por exemplo, mando de campo, o ataque e a defesa da equipe, o rival que está sendo enfrentado, entre outros.

3 Materiais e métodos

3.1 Dados

Esse estudo tem como base os dados referentes ao campeonato brasileiro de futebol da série A do ano de 2016. Essa base foi coletada manualmente do site da CBF (CBF, 2016) [2] (www.cbf.com.br/competicoes/brasileiro-serie-a/tabela/2016#.WPQKHekrLIU), rodada após rodada. Neste ano 20 equipes participaram do campeonato que teve como moldes a disputa por pontos corridos, onde todas as equipes se enfrentavam duas vezes em um sistema de turno e returno, sendo assim 38 jogos por equipe e totalizando 380 jogos no campeonato.

O gráfico 1 abaixo mostra como os times oscilam rodada a rodada através da pontuação acumulada.

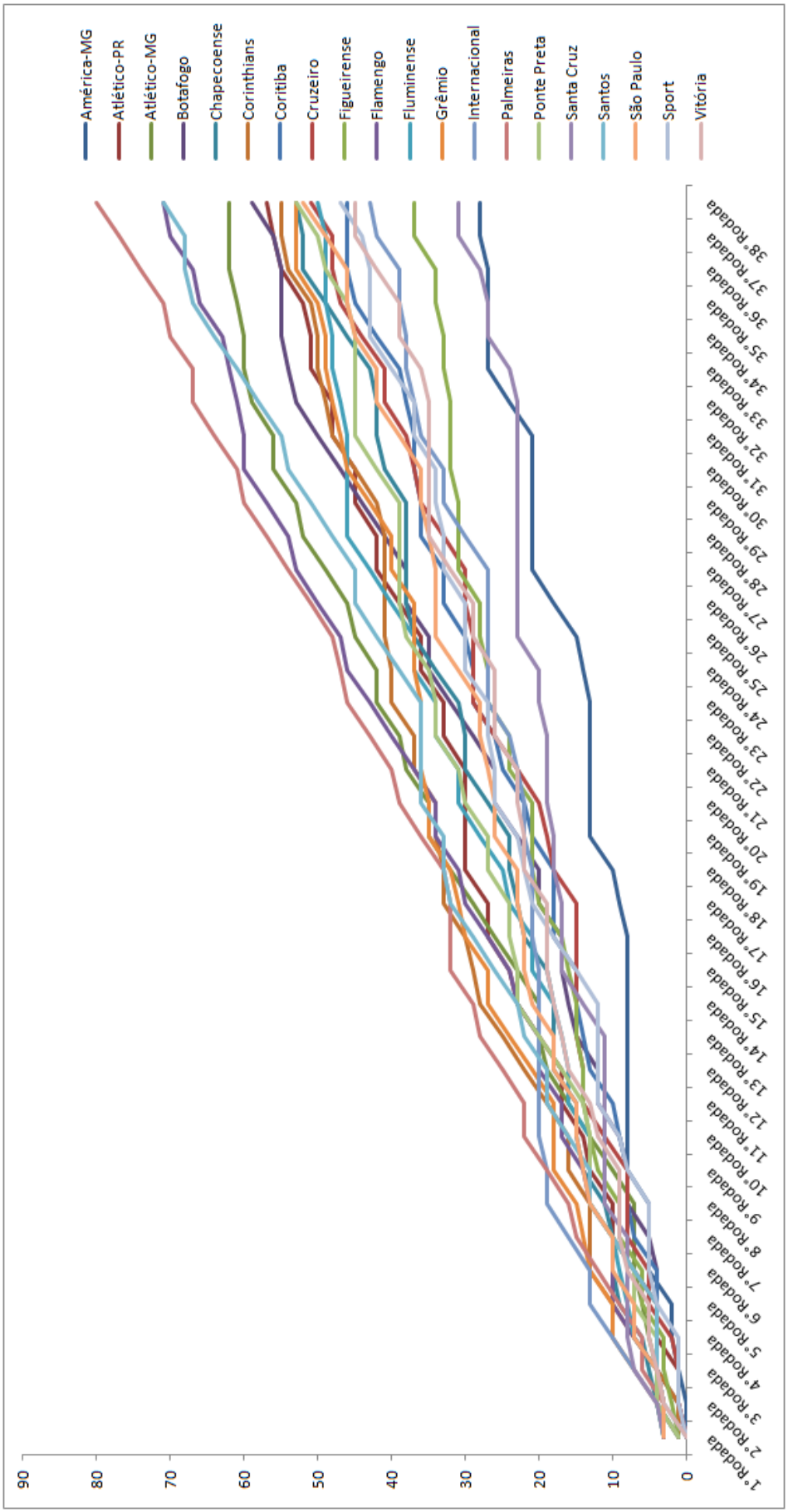

Figura 1: Gráfico pontuação rodada a rodada

O mando de campo e fatores que representam a eficiência de ataque e defesa serão utilizados pois é notável uma diferença entre os primeiros colocados e os últimos no número de gols feitos e no número de gols levados, assim como é de nítida percepção que times que jogam em seus domínios tendem a levar vantagem sobre os times visitantes. Para auxiliar na visualização desses fatos foi montada a tabela 1 com a classificação de cada time ao final do campeonato, os pontos assim como os resultados como mandante e como visitante foram separados para melhor visualização das diferenças de desempenho, esta tabela também possui o número de gols pró, gols contra e o saldo de gols:

Tabela 1: Tabela do campeonato

Posição	Classificação	P	J	V	E	D	PM	VM	EM	DM	PV	VV	EV	DV	GP	GC	SG
1º	PALMEIRAS	80	38	24	8	6	46	14	4	1	34	10	4	5	62	32	30
2º	SANTOS	71	38	22	5	11	47	15	2	2	24	7	3	9	59	35	24
3º	FLAMENGO	71	38	20	11	7	41	12	5	2	30	8	6	5	52	35	17
4º	ATLÉTICO-MG	62	38	17	11	10	42	13	3	3	20	4	8	7	61	53	8
5º	BOTAFOGO	59	38	17	8	13	35	10	5	4	24	7	3	9	43	39	4
6º	ATLÉTICO-PR	57	38	17	6	15	48	15	3	1	9	2	3	14	38	32	6
7º	CORINTHIANS	55	38	15	10	13	37	10	7	2	18	5	3	11	48	42	6
8º	PONTE PRETA	53	38	15	8	15	41	13	2	4	12	2	6	11	48	52	-4
9º	GRÊMIO	53	38	14	11	13	37	11	4	4	16	3	7	9	41	44	-3
10º	SÃO PAULO	52	38	14	10	14	32	9	5	5	20	5	5	9	44	36	8
11º	CHAPECOENSE	52	38	13	13	12	30	8	6	5	22	5	7	7	49	56	-7
12º	CRUZEIRO	51	38	14	9	15	28	7	7	5	23	7	2	10	48	49	-1
13º	FLUMINENSE	50	38	13	11	14	30	8	6	5	20	5	5	9	45	45	0
14º	SPORT	47	38	13	8	17	35	10	5	4	12	3	3	13	49	55	-6
15º	CORITIBA	46	38	11	13	14	34	9	7	3	12	2	6	11	41	42	-1
16º	VITÓRIA	45	38	12	9	17	28	8	4	7	17	4	5	10	51	53	-2
17º	INTERNACIONAL	43	38	11	10	17	32	9	5	5	11	2	5	12	35	41	-6
18º	FIGUEIRENSE	37	38	8	13	17	30	7	9	3	7	1	4	14	30	50	-20
19º	SANTA CRUZ	31	38	8	7	23	23	7	2	10	8	1	5	13	45	69	-24
20º	AMÉRICA-MG	28	38	7	7	24	24	7	3	9	4	0	4	15	23	58	-35

Na tabela 1 acima tem-se: P=Pontos, J=Jogos, V=Vitórias, E=Empates, D=Derrotas, PM=Pontos como mandante, VM=Vitórias como mandante, EM=Empates como mandante, DM=Derrotas como mandante, PV=Pontos como visitante, VV=Vitórias como visitante, EV=Empates como visitante, DV=Derrotas como visitante, GP=Gols pró, GC=Gols contra e SG=Saldo de gols.

3.2 Metodologia

O objetivo do estudo é encontrar uma relação entre o número de gols feitos e diversas variáveis, como por exemplo o mando de campo que possam influenciar a primeira. Para encontrar essa relação inicialmente será utilizado um modelo linear para se encontrar uma equação que relacione as variáveis acima citadas.

A análise de regressão linear consiste na realização de uma análise estatística para verificar se há relação entre uma variável dependente com uma ou mais variáveis independentes,

ou seja, consiste na obtenção de uma equação que procura explicar a variação da variável dependente pela variação da(s) variável(is) independente(s).
O modelo escolhido deve ser coerente com o que acontece na prática. Para isto, deve-se levar em conta as seguintes considerações na hora de se escolher o modelo:

- O modelo selecionado deve ser condizente tanto no grau como no aspecto da curva, para representar o fenômeno em estudo.
- O modelo deve conter apenas as variáveis que são relevantes para explicar o fenômeno. Pelo tipo de problema que este estudo se propõe a resolver, têm-se como proposta de solução a utilização de um modelo linear.

A seguir, será apresentada uma breve descrição das ferramentas utilizadas para a modelagem deste trabalho.

3.2.1 Modelos lineares generalizados

Anteriormente foi mencionada a necessidade de utilizar modelos lineares para relacionar o número de gols com diversas variáveis que expliquem a variação desta resposta.

O modelo inicialmente proposto estabelece que:

$$E[Y_i] = \beta_0 + \beta_1 X_{1,i} + ... + \beta_p X_{p,i}$$

onde $i = 1, ..., n$, Y_i é o número de gols do time i, $X_{1,i}, ..., X_{p,i}$ são p variáveis regressoras para o time i que explicam a variação de Y_i, $\beta_1, ..., \beta_p$ são os efeitos de cada variável regressora e β_0 é chamado de intercepto.

Assumindo que Y_i tem distribuição normal, o modelo pode ser escrito como:

$$\begin{aligned} Y_i &\sim N(\mu_i, \sigma^2) \quad i = 1, ...n \\ \mu_i &= \tau_i \\ \tau_i &= \beta_0 + \beta_1 X_{1,i} + ... + \beta_p X_{p,i} \end{aligned}$$

onde a segunda equação é aparentemente desnecessária, criando um novo parâmetro, τ_i.

A distribuição normal pertence a uma família de distribuições chamada de família

exponencial e, para estas, McCullagh-Nelder (McCullagh-Nelder, 1989) [1] propõem o uso da seguinte generalização:

$$\begin{aligned} Y_i &\sim p(y_i|\theta) \quad i = 1, ...n \\ g(\mu_i) &= \tau_i \\ \tau_i &= \beta_0 + \beta_1 X_{1,i} + ... + \beta_p X_{p,i} \end{aligned}$$

onde $p(y_i|\theta)$ é uma distribuição que pertence à família exponencial, $\mu_i = E[Y_i]$ e g é uma função diferenciável. A primeira equação é a parte probabilística do modelo, a segunda e a terceira definem uma associação entre uma função diferenciável da média de Y_i e a parte sistemática do modelo. Estes modelos são conhecidos pelo nome de Modelos Lineares Generalizados.

No presente trabalho, a variável resposta é o número de gols feitos pelo time i na rodada j, $Y_{i,j}$, que se assume tendo distribuição de Poisson de parâmetro $\lambda_{i,j}$, sendo que esta distribuição pertence à família exponencial, pois ela pode ser escrita como:

$$p(y_i|\lambda_i) = \frac{1}{x_i!} \exp\left(x_i \ln \lambda_i - \lambda_i\right).$$

Assim, o modelo linear generalizado é escrito, de forma geral, como:

$$\begin{aligned} Y_{i,j} &\sim Poisson(\lambda_{i,j}) \quad i = 1, ...n \\ \ln(\lambda_{i,j}) &= \beta_0 + \beta_1 X_{1,i,j} + ... + \beta_p X_{p,i,j} \end{aligned}$$

3.2.2 Proposta de modelo

Este estudo se propõe a fazer uma modelagem para o número de gols que cada time irá marcar em cada rodada. Logo, será explicado o resultado do jogo M x V (Mandante contra visitante).

Para auxiliar na modelagem iremos utilizar fatores aleatórios que podem determinar o comportamento dos times, sendo esses fatores:

- Ataque
- Defesa

- Mando de campo

Utilizando esses fatores no auxílio da modelagem para o jogo M x V assumindo que o número de gols tem distribuição Poisson, tem-se um modelo inicial da seguinte maneira:

- $Y_{i,j} \sim Poisson(\lambda_{i,j})$
- $ln(\lambda_{i,j}) = A_i - D_i + C_i X_{i,j}$

onde

- $Y_{i,j}$ representa o número de gols feitos pelo time i na rodada j
- A_i é um efeito aleatório que representa o fator ataque do time i
- D_i é um efeito aleatório que representa o fator defesa do time i
- C_i representa o efeito do mando de campo do time i
- $X_{i,j}$ é uma variável que indica se o time i tem o mando de campo na rodada j, ou seja,

$$X_{i,j} = \begin{cases} 1 & \text{, se } i \text{ joga em casa na rodada } j \\ 0 & \text{, caso contrário} \end{cases}$$

Este modelo foi definido usando idéias relacionadas com o esporte, por exemplo, quanto melhor o ataque de um time, espera-se um maior número de gols e quanto melhor a defesa do mesmo, menor o número de gols que ele levará, mas também menor a preocupação em fazer gols.

Observar que, como está sendo modelado o número de gols, ao incluir um efeito aleatório associado ao ataque do time, este terá sinal possitivo pois ele *aumenta o número de gols do time*, já o fator defesa atua de forma oposta, diminuindo o número de gols pois, quando o time é forte na defesa, tende a abrir mão de fazer gols, por este motivo, este efeito aparece no modelo com sinal negativo.

Outros modelos foram testados, porém, todos eles tem a mesma estrutura básica que utilizam os modelos lineares generalizados com uma variável regressora (mando de campo)

e efeitos aleatórios. A seguir, são descritos os modelos utilizados.

Os cinco modelos consideram que o número de gols feitos pelo time i na rodada j, $Y_{i,j}$, segue uma distribuição de Poisson com média $\lambda_{i,j}$. A função de ligação usada é $ln(\lambda_{i,j})$. A diferença entre os cinco modelos está no componente sistemático que, além de ter como variável regressora o mando de campo, terá:

Modelo 1: nenhum outro fator além do mando de campo.

Modelo 2: um fator de ataque e outro de defesa do próprio time, que é o modelo anteriormente apresentado.

Modelo 3: um fator de ataque e outro de defesa do próprio time e também um fator de ataque e outro de defesa do rival da rodada.

Modelo 4: um fator de ataque e outro de defesa do próprio time e também um fator de ataque e um fator aleatório relacionado ao rival da rodada.

Modelo 5: um fator de ataque do próprio time e um fator defesa do time rival da rodada.

A expressão formal dos modelos será apresentada no seguinte capítulo, junto com os resultados de cada modelo.

3.2.3 Estimação de parâmetros

A estimação nos modelos foi feita sob uma abordagem bayesiana, o que implica na definição de distribuições a priori para os parâmetros desconhecidos para que, em conjunto com a função de verossimilhança, possam ser determinadas as distribuições a posteriori dos mesmos. Neste trabalho foram usadas distribuições normais com variâncias grandes como forma de definir prioris não informativas para a maioria de parâmetros, porém, para

os fatores ataque e defesa foram utilizadas distribuições normais padrão.

A determinação da distribuição a posteriori é feita utilizando o Teorema de Bayes, através da relação

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

onde $p(\theta|x)$ é a distribuição a posteriori para o parâmetro θ, $p(\theta)$ é a distribuição a priori e $p(x|\theta)$ é a função de verossimilhança. Esta equação é genérica pois, no caso do presente trabalho, θ na verdade é um vetor de parâmetros e x representa o conjunto de dados disponíveis. Um tratado detalhado de inferência Bayesiana, onde se encontra itens como definição de prioris e inferência encontra-se em Migon e Gamerman (Migon e Gamerman, 1999) [4].

As distribuições a posteriori dos modelos considerados neste trabalho são complexas, dificultando a realização de contas analíticas para realizar estimação a partir delas. Nestes casos existe a possibilidade de usar métodos aproximados, especificamente, métodos de Monte Carlo em cadeias de Markov, para obter amostras dos parâmetros provenientes das distribuições a posteriori. Com estas amostras é possível fazer inferência.

3.2.3.1 Métodos de Monte Carlo em Cadeias de Markov (MCMC)

Em inferência bayesiana é comum resumir a informação contida na distribuição a posteriori, usando diversos mecanismos que envolvem, muitas vezes, o cálculo de valores esperados. Quando a distribuição a posteriori é intratável de forma analítica é possível utilizar métodos aproximados que conseguem obter amostras das distribuições de interesse e calcular estimativas amostrais de características da distribuição.

Os métodos de Monte Carlo em cadeias de Markov têm por base realizar um passeio aleatório no espaço paramétrico que converge para a distribuição estacionária, que é a distribuição a posteriori de interesse. Existem dois algoritmos que utilizam esta abordagem: o algoritmo de Metropolis-Hastings e o amostrador de Gibbs.

O algoritmo de Metropolis-Hastings é um método iterativo que, em cada iteração, gera um valor proposto de uma distribuição auxiliar que será aceito, ou não, como vindo

da distribuição de interesse com uma certa probabilidade. Uma distribuição comumente utilizada para gerar o valor proposto é a distribuição a priori.

O amostrador de Gibbs utiliza as distribuições marginais completas para gerar, em cada iteração, um novo vetor de valores da distribuição de interesse. A diferença com o algoritmo de Metropolis-Hastings é que neste, um valor proposto numa certa iteração pode não ser aceito como proveniente da distribuição de interesse, já no amostrador de Gibbs, o valor proposto sempre será aceito.

Ambos métodos requerem de valores inicias para a primeira iteração, assim como de um certo número de iterações a partir do qual se espera que sejam gerados valores provenientes da distribuição de interesse. Este número de iterações é chamado de aquecimento. Neste trabalho foram consideradas 100000 iterações para o aquecimento e depois foram observados os valores gerados nas 10000 iterações seguintes para calcular as estimativas necessárias.

Uma discussão detalhada sobre o tema pode ser encontrada em Gamerman (Gamerman, 1999) [4].

Os algoritmos apresentados aqui encontram-se implementados em diversos pacotes. Neste trabalho foi utilizado o OpenBUGS versão 3.2.3, disponível em `http://www.openbugs.net`. Os scripts utilizados neste trabalho se encontram no Anexo A.

3.2.3.2 Escolha de modelo

Neste trabalho foram propostos 5 modelos. É natural pensar qual deles será o melhor. Gelfand e Ghosh (Gelfand e Ghosh, 1998) [5] definem uma estatística para a escolha de modelos usando uma função de perda que pode ser a função desvio definida em McCullagh-Nelder (McCullagh-Nelder, 1989) [1]. Para o caso deste trabalho em que os dados seguem uma distribuição de Poisson, esta função desvio é definida como

$$Des(\overrightarrow{y}, \overrightarrow{\hat{y}}) = 2 \sum_i \sum_j \left(y_{i,j} \ln \left(\frac{y_{i,j}}{\hat{y}_{i,j}} \right) - (y_{i,j} - \hat{y}_{i,j}) \right)$$

onde $\overrightarrow{y}$ é a matriz que contém os gols marcados por cada time em todas as rodadas, $\overrightarrow{\hat{y}}$ é a correspondente matriz de valores previstos por um modelo. Esta estatística avalia o ajuste do modelo levando em conta a distribuição dos dados de forma que um modelo com bom ajuste apresenta valor pequeno de $Des(\overrightarrow{y}, \overrightarrow{\hat{y}})$.

4 Análise e discussão dos Resultados

Como o objetivo do estudo é modelar o número de gols feitos por cada time para poder prever quem será vitorioso nas partidas, algumas estatísticas descritivas sobre o número de gols rodada a rodada são apresentadas para um melhor entendimento sobre a oscilação que acontece ao longo do campeonato, estas estatísticas descritivas seguem na tabela 2 abaixo:

Tabela 2: Tabela com algumas estatísticas descritivas do campeonato

	Total gols	Média	Mediana	Moda	Variância
Rodada 1	14	0,7	0	0	1,48
Rodada 2	31	1,55	1,5	1	0,58
Rodada 3	29	1,45	1	1	1,21
Rodada 4	19	0,95	1	1	0,47
Rodada 5	25	1,25	1	1	1,67
Rodada 6	22	1,1	1	0	1,67
Rodada 7	28	1,4	1	1	0,88
Rodada 8	28	1,4	1,5	0 e 2	1,31
Rodada 9	32	1,6	2	2	0,99
Rodada 10	21	1,05	1	0	1,63
Rodada 11	28	1,4	1	1	1,62
Rodada 12	37	1,85	2	0,2 e 3	2,13
Rodada 13	23	1,15	1	0 e 1	1,40
Rodada 14	19	0,95	1	0	1,10
Rodada 15	31	1,55	1	1	1,31

Rodada 16	24	1,2	1	1	0,69
Rodada 17	22	1,1	1	0	1,36
Rodada 18	26	1,3	1	0 e 1	1,38
Rodada 19	27	1,35	1	1	0,66
Rodada 20	21	1,05	1	0	1,10
Rodada 21	26	1,3	1	1	0,64
Rodada 22	19	0,95	1	1	0,79
Rodada 23	27	1,35	1	1 e 2	1,19
Rodada 24	32	1,6	1	1	1,83
Rodada 25	19	0,95	1	1	0,68
Rodada 26	20	1	1	0 e 1	0,84
Rodada 27	23	1,15	1	1	1,08
Rodada 28	26	1,3	1	0	1,59
Rodada 29	19	0,95	1	0	1,00
Rodada 30	20	1	0,5	0	1,58
Rodada 31	24	1,2	1	0,1 e 2	1,01
Rodada 32	18	0,9	1	0	0,83
Rodada 33	16	0,8	1	0 e 1	0,59
Rodada 34	26	1,3	1	0	1,91
Rodada 35	20	1	1	1	0,53
Rodada 36	28	1,4	1,5	0	1,62
Rodada 37	21	1,05	1	1	1,42
Rodada 38	21	1,05	1	0	1,79
Campeonato	912	1,2	1	1	1.20

lembrando que a média, a mediana, a moda e a variância são decorrentes dos gols marcados por cada time e não em cada partida.

O gráfico 2 a seguir mostra como o total de gols evoluiu rodada a rodada:

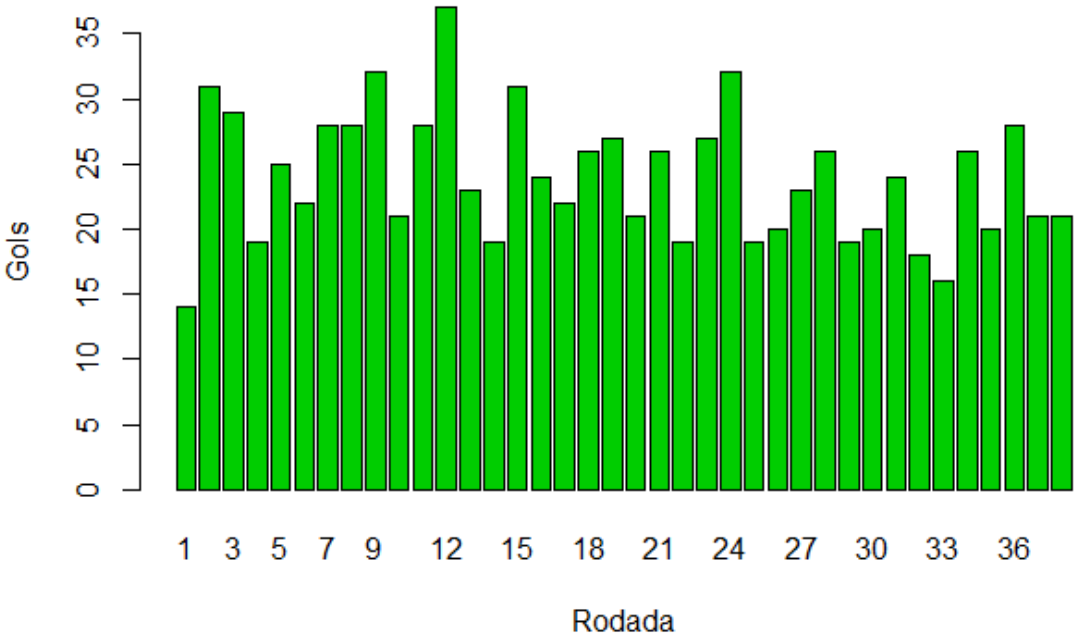

Figura 2: Gráfico do total de gols por rodada

O gráfico 3 abaixo evidencia que ao longo do campeonato é importante o time fazer gols regularmente, visto que os times que fazem mais gols terminam em melhores posições do que os times que fazem poucos gols

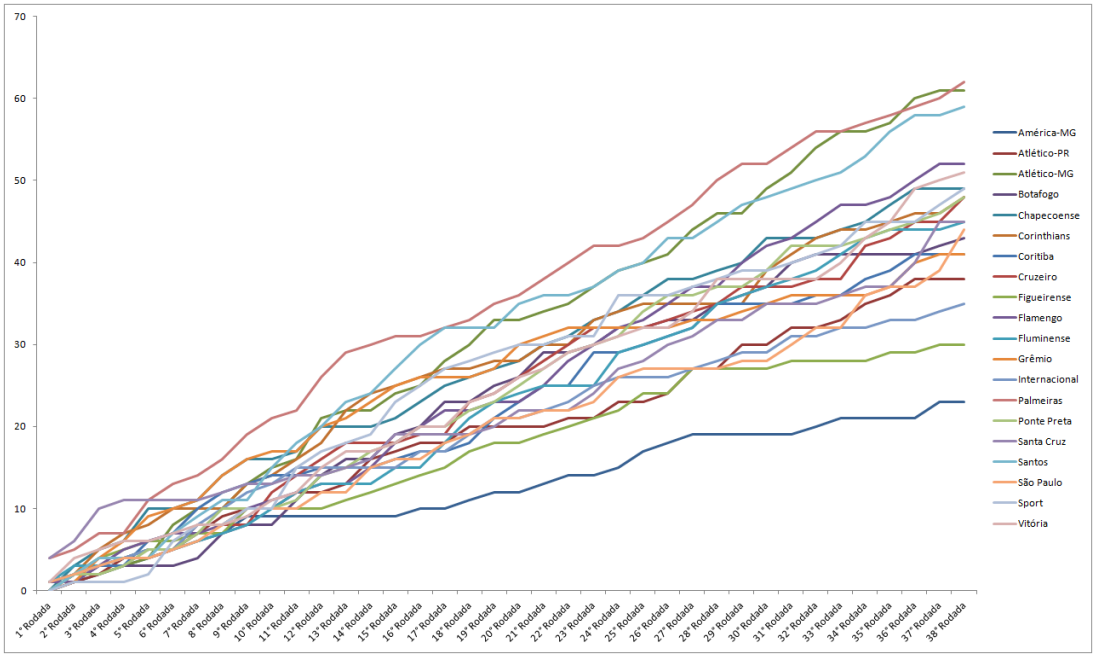

Figura 3: Gráfico gols feitos acumulados rodada a rodada

O gráfico 4 abaixo mostra um pouco melhor a importância em se jogar dentro de seus domínios e com a torcida a seu favor.

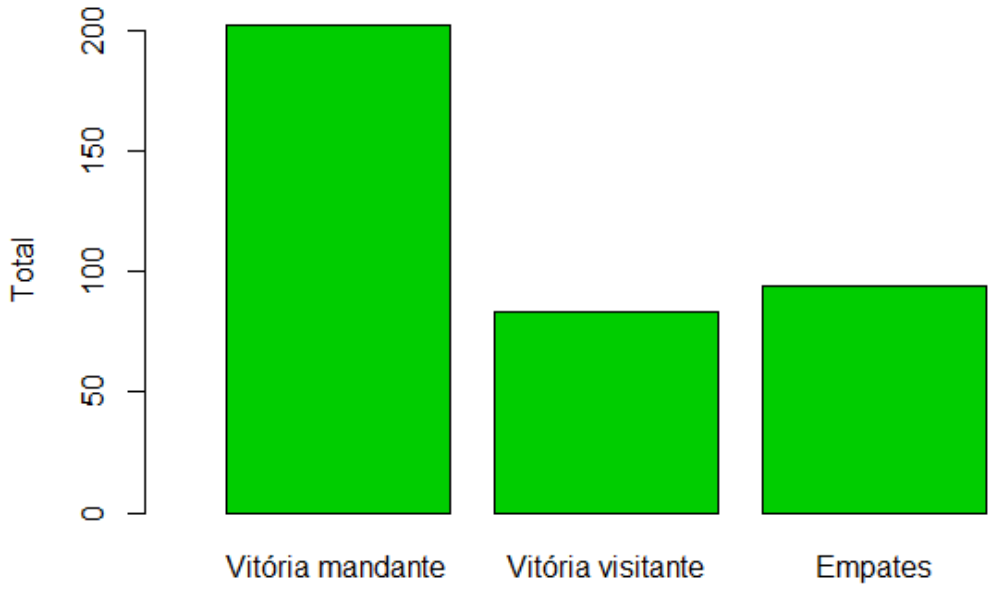

Figura 4: Gráfico da distribuição de vitórias como mandante e visitante

Como o número de gols é o que será modelado, foi feito o gráfico 5 da distribuição dos gols no campeonato para observar a maneira como eles se distribuem:

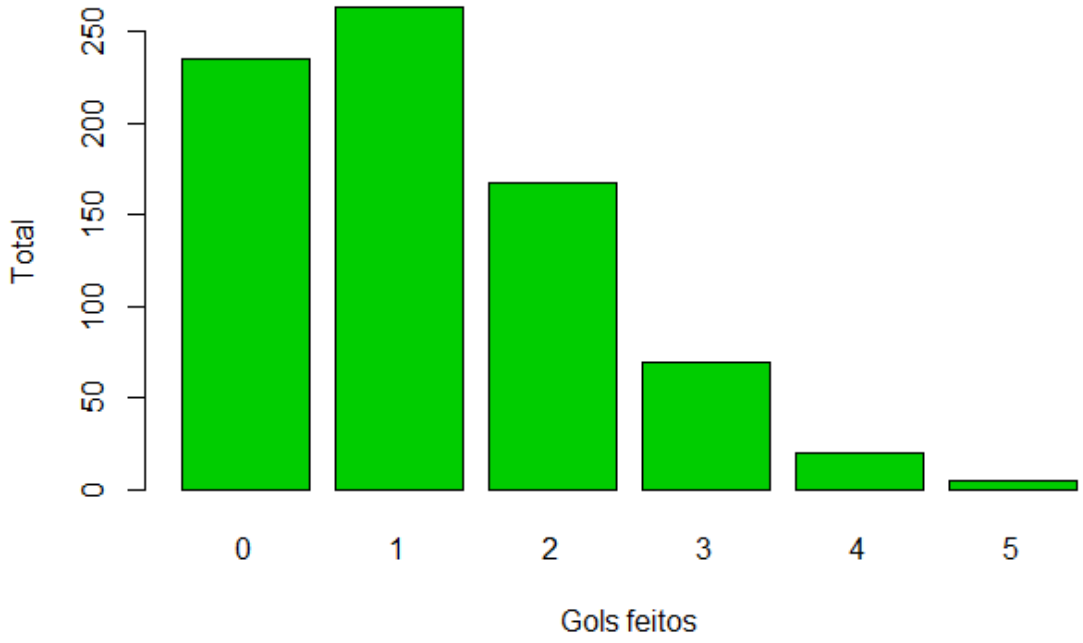

Figura 5: Gráfico da distribuição dos gols no campeonato

O gráfico 5 da distribuição dos gols foi dividido em dois, no gráfico 6 onde mostra a maneira como os gols se distribuem quando o time da casa faz gol e quando o time visitante faz gol:

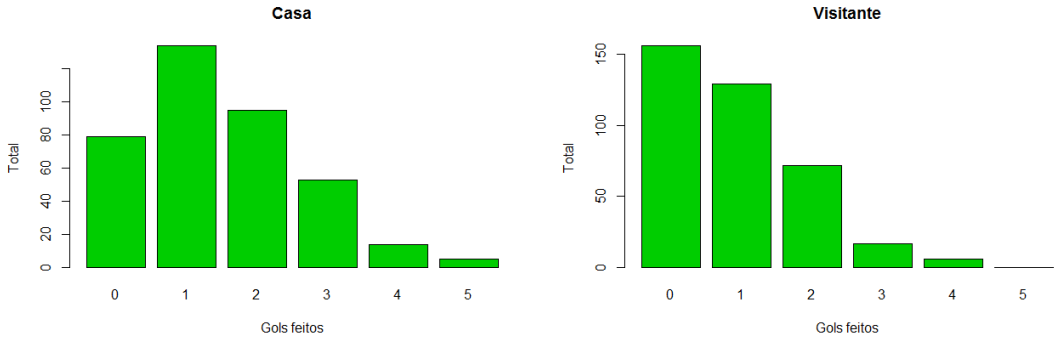

Figura 6: Gráfico da distribuição dos gols no campeonato separado pelo mando de campo

Pode ser observado também que a quantidade de gols feitos pelo time visitante equivale ao gráfico de gols levados pelo time mandante e vice-versa, fazendo assim desnecessário a

utilização dos gráficos de gols levados.

Com os gráficos fica reforçada a necessidade da utilização de um fator que represente o ataque, de outro fator para a defesa e do fator associado ao mando de campo para se modelar a quantidade de gols marcado por cada time a cada rodada.

Os modelos apresentados a seguir foram pensados de forma intuitiva e de acordo com a realidade. Dessa forma serão considerados os fatores ataque e defesa, além do mando de campo como covariável.

Deve-se levar em conta que alguns fatores modelados não são observáveis e por isso foram utilizadas prioris não informativas.

4.1 Modelo 1

Neste modelo apenas é levado em conta o mando de campo, assim, o modelo fica definido da seguinte forma:

- $Y_{i,j} \sim Poisson(\lambda_{i,j})$
- $ln(\lambda_{i,j}) = \beta 0 + \beta 1 * X_{i,j}$
- $\beta 0 \sim N(0; 1000)$
- $\beta 1 \sim N(0; 1000)$
- $X_{i,j}$ é a variável que indica o mando de campo do time i na rodada j

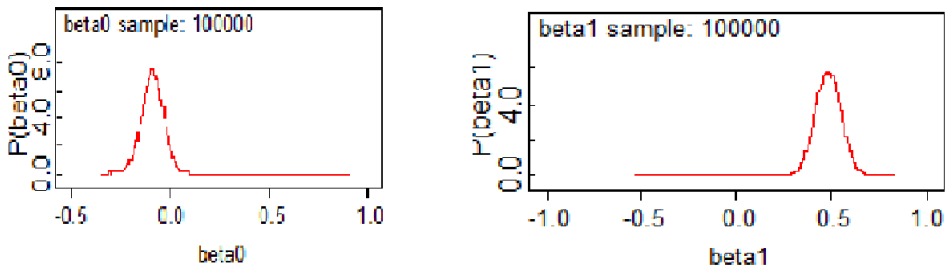

Figura 7: Distribuição a posteriori de $\beta 0$ e $\beta 1$

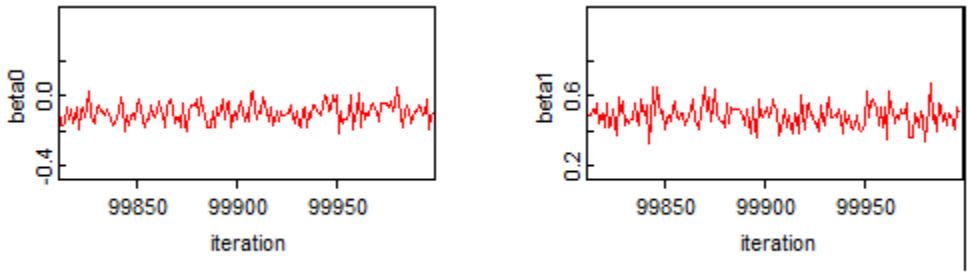

Figura 8: Cadeias de $\beta 0$ e $\beta 1$

Tabela 3: Estatísticas para o Modelo 1

	Média	Desvio Padrão	Intervalo de credibilidade	
$\beta 0$	-0,08889	0,05381	-0,1954	0,01461
$\beta 1$	0,4826	0,06829	0,3492	0,6168

Os valores estimados para os parâmetros deste modelo aparecem na tabela 3. Deve ser observado que o mando de campo tem um efeito positivo e significativo, o que indica que os times com o mando de campo fazem, em média, mais gols do que os times que não tem o mando de campo.

Os histogramas dos valores amostrados, que constam das figuras 7 indicam que as distribuições a posteriori dos parâmetros do modelo seguem distribuições simétricas, centradas nos valores das médias que constam na tabela 3.

Finalmente, os valores gerados nas 10000 iterações consideradas para amostragem formam trajetórias representadas nas figuras 8 que sugerem que o método de MCMC conseguiu convergir para a distribuição a posteriori de interesse.

4.2 Modelo 2

Modelo onde é levado em conta o mando de campo, o fator aleatório do ataque e o fator aleatório da defesa, onde:

- $Y_{i,j} \sim Poisson(\lambda_{i,j})$
- $ln(\lambda_{i,j}) = \beta 0 + \beta 1 * X_{i,j} + \beta 2_i - \beta 3_i$
- $\beta 0 \sim N(0; 1000)$
- $\beta 1 \sim N(0; 1000)$
- $\beta 2_i \sim N(0; 1)$ é o fator aleatório do ataque do time i
- $\beta 3_i \sim N(0; 1)$ é o fator aleatório da defesa do time i
- $X_{i,j}$ é a variável que indica o mando de campo do time i na rodada j

Os resultados para os parâmetros $\beta 2$ e $\beta 3$ se encontram nas tabelas 8 e 9 respectivamente.

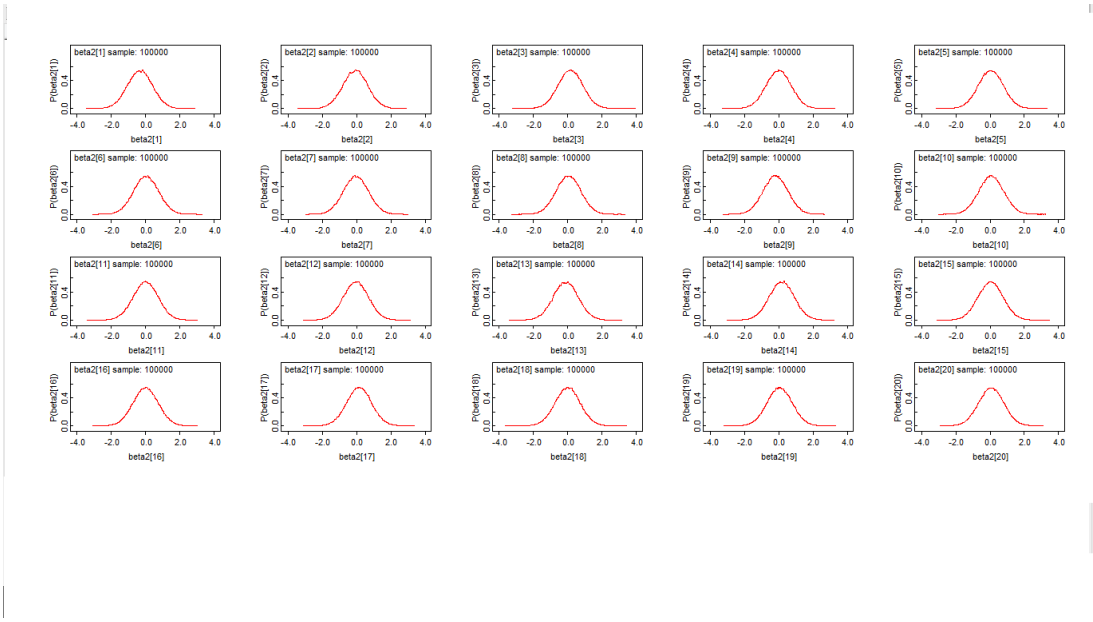

Figura 9: Gráficos do $\beta 2$

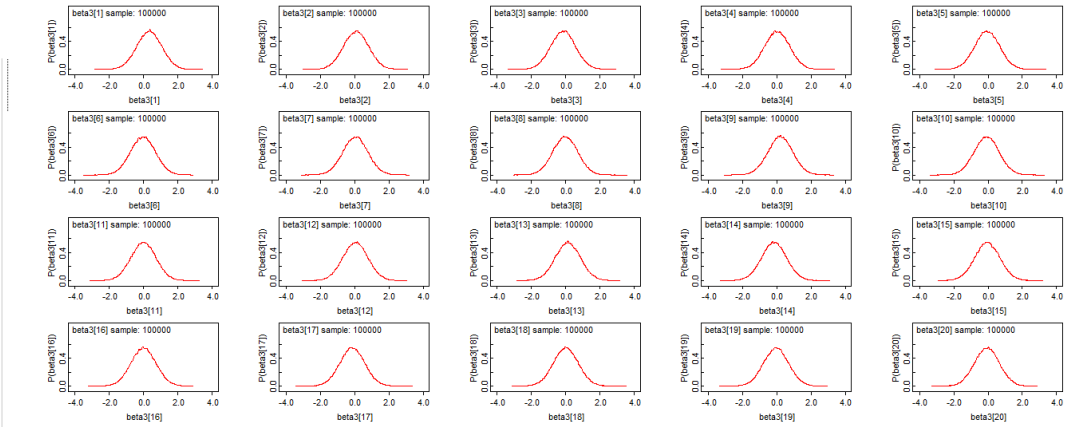

Figura 10: Gráficos do $\beta 3$

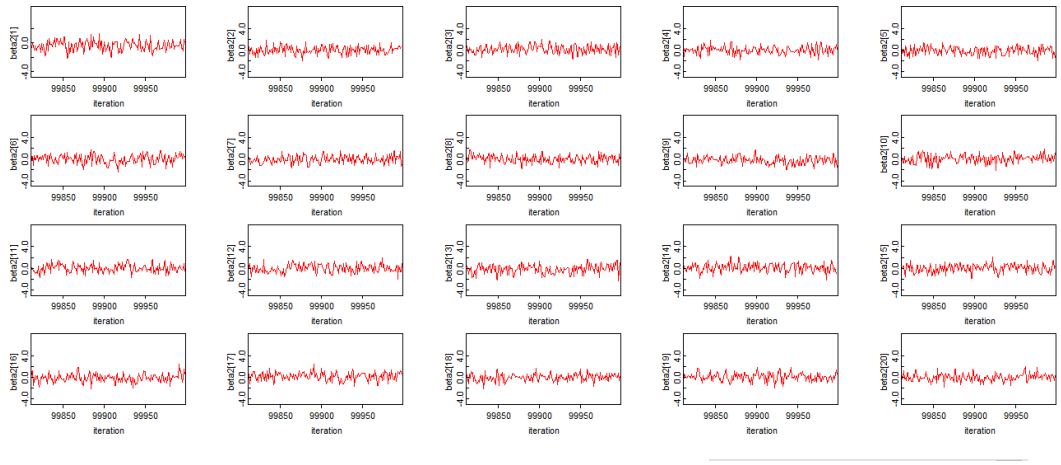

Figura 11: Cadeias do $\beta 2$

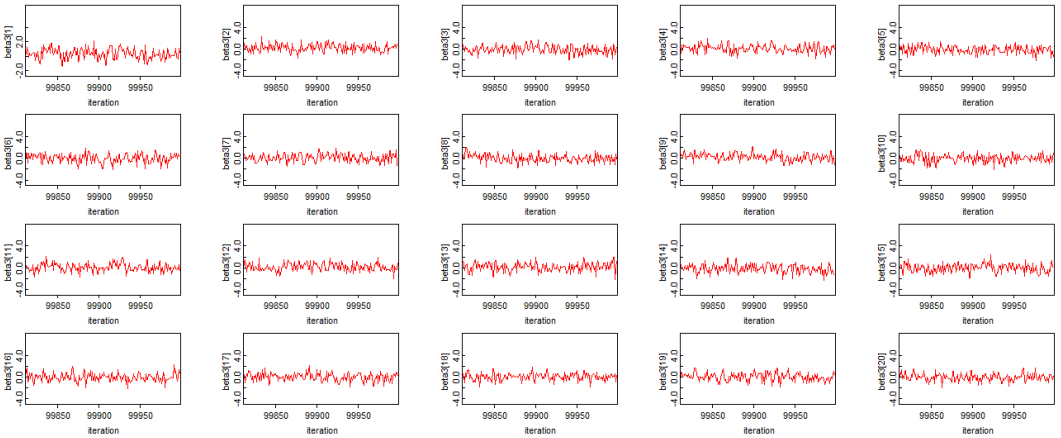

Figura 12: Cadeias do $\beta 3$

Tabela 4: Modelo 2

	Média	Desvio Padrão	Intervalo	
$\beta 0$	-0,114	0,3254	-0,7345	0,5622
$\beta 1$	0,4836	0,06821	0,3501	0,6178

Analogamente como demonstrado para o Modelo 1, como os gráficos são semelhantes tem-se que as distribuições a posteriori dos parâmetros do modelo seguem distribuições simétricas centradas nas médias que constam na tabela 4, os valores de $\beta 0$ e $\beta 1$ convergem e indicam que o mando de campo tem um efeito positivo e significativo.

Tem-se das figuras 9 e 10 que os gráficos de $\beta 2$ e $\beta 3$ mostram uma simetria para todos os betas de todos os times, evidenciando assim que seguem uma distribuição Normal. Já

as figuras 11 e 12 mostram que os valores de $\beta 2$ e $\beta 3$ convergem de acordo com os seus respectivos valores nas tabelas 8 e 9.

4.3 Modelo 3

Modelo onde é levado em conta o mando de campo, o fator aleatório do ataque, o fator aleatório da defesa, o fator aleatório do ataque do time adversário e ao fator aleatório da defesa do time adversário, onde:

- $Y_{i,j} \sim Poisson(\lambda_{i,j})$
- $ln(\lambda_{i,j}) = \beta 0 + \beta 1 * X_{i,j} + \beta 2_i - \beta 3_i - \beta 2_{rival_{i,j}} + \beta 3_{rival_{i,j}}$
- $\beta 0 \sim N(0; 1000)$
- $\beta 1 \sim N(0; 1000)$
- $\beta 2_i \sim N(0; 1)$ é o fator aleatório do ataque do time i
- $\beta 3_i \sim N(0; 1)$ é o fator aleatório da defesa do time i
- $\beta 2_{rival_{i,j}} \sim N(0; 1)$ é o fator aleatório do ataque do rival do time i na rodada j
- $\beta 3_{rival_{i,j}} \sim N(0; 1)$ é o fator aleatório da defesa do rival do time i na rodada j
- $X_{i,j}$ é a variável que indica o mando de campo do time i na rodada j

Tabela 5: Modelo 3

	Média	Desvio Padrão	Intervalo	
$\beta 0$	-0,156	0,3927	-0,939	0,6164
$\beta 1$	0,4834	0,0679	0,3504	0,6169

Assim como no modelo 2, como as figuras são semelhantes, os valores de $\beta 0$, $\beta 1$, $\beta 2$ e $\beta 3$ indicam que as distribuições posteriori dos parâmetros seguem distribuições simétricas centradas em suas respectivas médias.

Os valores gerados nas iterações sugerem que o método de MCMC conseguiu convergir para a distribuição a posteriori de interesse.

Os resultados para os parâmetros $\beta 2$ e $\beta 3$ se encontram nas tabelas 8 e 9 respectivamente.

4.4 Modelo 4

Modelo onde é levado em conta o mando de campo, o fator aleatório do ataque, o fator aleatório da defesa, e o fator aleatório do ataque do time adversário, onde:

- $Y_{i,j} \sim Poisson(\lambda_{i,j})$
- $ln(\lambda_{i,j}) = \beta 0 + \beta 1 * X_{i,j} + \beta 2_i - \beta 3_i + \beta 4_{rival_{i,j}}$
- $\beta 0 \sim N(0; 1000)$
- $\beta 1 \sim N(0; 1000)$
- $\beta 2_i \sim N(0; 1)$ é o fator aleatório do ataque do time i
- $\beta 3_i \sim N(0; 1)$ é o fator aleatório da defesa do time i
- $\beta 4_{rival_{i,j}} \sim N(0; 1)$ é o fator aleatório do ataque do rival do time i na rodada j
- $X_{i,j}$ é a variável que indica o mando de campo do time i na rodada j

Tabela 6: Modelo 4

	Média	Desvio Padrão	Intervalo	
$\beta 0$	-0,1245	0,05422	-0,2318	-0,01911
$\beta 1$	0,4834	0,06822	0,3497	0,6168

Assim como no modelo 3, como as figuras são semelhantes, os valores de $\beta 0$, $\beta 1$, $\beta 2$, $\beta 3$ e $\beta 4$ indicam que as distribuições posteriori dos parâmetros dos modelos seguem distribuições simétricas centradas em suas respectivas médias.

Os valores gerados nas iterações sugerem que o método de MCMC conseguiu convergir para a distribuição a posteriori de interesse.

Os resultados para os parâmetros $\beta 2$, $\beta 3$ e $\beta 4$ se encontram nas tabelas 8, 9 e 10 respectivamente. O parâmetro $\beta 4$ é uma tentativa de medir a eficiência de cada time ao enfrentar os outros.

4.5 Modelo 5

Modelo onde é levado em conta o mando de campo, o fator aleatório do ataque e o fator aleatório da defesa do time adversário, onde:

- $Y_{i,j} \sim Poisson(\lambda_{i,j})$
- $ln(\lambda_{i,j}) = \beta 0 + \beta 1 * X_{i,j} + \beta 2_i - \beta 3_{rival_{i,j}}$
- $\beta 0 \sim N(0; 1000)$
- $\beta 1 \sim N(0; 1000)$
- $\beta 2_i \sim N(0; 1)$ é o fator aleatório do ataque do time i
- $\beta 3_{rival_{i,j}} \sim N(0; 1)$ é o fator aleatório da defesa do rival do time i na rodada j
- $X_{i,j}$ é a matriz com os mandos de campo, onde i=time e j=rodada

Tabela 7: Modelo 5

	Média	Desvio Padrão	Intervalo	
$\beta 0$	-0,1562	0,3294	-0,7861	0,5112
$\beta 1$	0,4837	0,06826	0,3497	0,6178

Assim como no modelo 3, como as figuras são semelhantes, os valores de $\beta 0$, $\beta 1$, $\beta 2$, $\beta 3$ e $\beta 4$ seguem uma distribuição posteriori dos parâmetros seguem distribuições simétricas centradas em suas respectivas médias que contam na tabela 7.

Os valores gerados nas iterações sugerem que o método de MCMC conseguiu convergir para a distribuição a posteriori de interesse.

Os resultados para os parâmetros $\beta 2$ e $\beta 3$ se encontram nas tabelas 8 e 9 respectivamente.

Tabela 8: Resultados do parâmetro $\beta 2$

		Modelo 2			Modelo 3			Modelo 4			Modelo 5		
Times	Gols Pró	$\beta 2$	Intervalo		$\beta 2$	Intervalo		$\beta 2$	Intervalo		$\beta 2$	Intervalo	
Palmeiras	62	0,1588	-1,249	1,575	0,1765	-1,25	1,618	0,1752	-1,232	1,577	0,31	-0,2004	0,8184
Atlético-MG	61	0,1513	-1,275	1,572	0,1683	-1,296	1,623	0,07619	-1,323	1,501	0,3136	-0,196	0,8191
Santos	59	0,1378	-1,29	1,569	0,1506	-1,258	1,604	0,07949	-1,333	1,499	0,2715	-0,2419	0,7801
Flamengo	52	0,07195	-1,362	1,504	0,07298	-1,35	1,511	0,0972	-1,255	1,522	0,1409	-0,3846	0,6588
Vitória	51	0,06625	-1,368	1,49	0,08191	-1,356	1,507	-0,0309	-1,486	1,374	0,1411	-0,3809	0,6568
Chapecoense	49	0,04572	-1,376	1,476	0,004566	-1,392	1,442	0,004089	-1,387	1,393	0,09483	-0,4285	0,6139
São Paulo	49	0,04141	-1,386	1,463	0,02485	-1,42	1,433	0,05791	-1,387	1,515	0,08233	-0,4433	0,6058
Corinthians	48	0,03595	-1,403	1,452	0,04805	-1,385	1,5	0,02834	-1,393	1,43	0,07332	-0,4538	0,5966
Cruzeiro	48	0,03114	-1,399	1,458	0,01998	-1,406	1,442	0,005763	-1,459	1,425	0,07747	-0,4475	0,6023
Ponte Preta	48	0,03232	-1,39	1,455	0,0394	-1,402	1,492	-0,005741	-1,449	1,408	0,0804	-0,4452	0,6026
Fluminense	45	-7,55E-05	-1,438	1,435	0,02251	-1,431	1,434	-0,01814	-1,457	1,412	0,004809	-0,5259	0,5329
Santa Cruz	45	-4,66E-04	-1,429	1,435	0,01592	-1,42	1,453	-0,07321	-1,502	1,385	0,02641	-0,5022	0,5542
Sport	44	-0,0116	-1,448	1,421	-0,007111	-1,431	1,398	-0,0782	-1,509	1,369	-0,004435	-0,5399	0,5222
Botafogo	43	-0,02364	-1,451	1,409	-0,01289	-1,438	1,422	0,008416	-1,462	1,423	-0,04333	-0,5777	0,4861
Coritiba	41	-0,04502	-1,477	1,39	-0,04739	-1,476	1,372	-0,002739	-1,403	1,408	-0,08884	-0,6281	0,4415
Grêmio	41	-0,04637	-1,479	1,389	-0,05207	-1,474	1,381	-0,01545	-1,451	1,465	-0,08598	-0,6263	0,4464
Atlético-PR	38	-0,08412	-1,508	1,341	-0,08774	-1,543	1,324	0,03345	-1,344	1,391	-0,1733	-0,7198	0,3638
Internacional	35	-0,1239	-1,549	1,305	-0,1178	-1,541	1,299	-0,03491	-1,437	1,377	-0,246	-0,7948	0,2985
Figueirense	30	-0,2038	-1,641	1,231	-0,1754	-1,618	1,26	-0,1063	-1,533	1,363	-0,378	-0,9457	0,1809
América-MG	23	-0,333	-1,779	1,11	-0,3226	-1,763	1,128	-0,1464	-1,571	1,269	-0,6346	-1,232	-0,05358

Tabela 9: Resultados do parâmetro $\beta 3$

		Modelo 2			Modelo 3			Modelo 4			Modelo 5		
Times	Gols Contra	$\beta 3$	Intervalo		$\beta 3$	Intervalo		$\beta 3$	Intervalo		$\beta 3$	Intervalo	
Santa Cruz	69	-7,14E-04	-1,436	1,425	-0,01489	-1,434	1,431	0,08825	-1,357	1,528	-0,3303	-0,8328	0,1936
América-MG	58	0,3319	-1,11	1,774	0,3203	-1,11	1,77	0,1876	-1,219	1,599	-0,2215	-0,7324	0,3112
Chapecoense	56	-0,03932	-1,47	1,388	-0,09524	-1,509	1,33	-0,0229	-1,415	1,368	-0,07883	-0,5982	0,4605
São Paulo	55	-0,04339	-1,478	1,389	-0,06221	-1,514	1,364	-0,09484	-1,526	1,376	0,2044	-0,3369	0,7628
Atlético-MG	53	-0,1545	-1,582	1,268	-0,152	-1,597	1,308	-0,07675	-1,475	1,352	-0,09287	-0,6154	0,4459
Vitória	53	-0,05886	-1,48	1,367	-0,06483	-1,51	1,374	-0,02608	-1,474	1,429	-0,1793	-0,6903	0,354
Ponte Preta	52	-0,03286	-1,445	1,386	-0,04595	-1,49	1,406	0,03014	-1,354	1,432	-0,1765	-0,69	0,3554
Figueirense	50	0,1995	-1,228	1,643	0,2042	-1,227	1,641	0,1573	-1,273	1,637	-0,2296	-0,7404	0,2963
Cruzeiro	49	-0,03445	-1,463	1,399	-0,06233	-1,477	1,381	0,01032	-1,417	1,406	-0,1185	-0,6367	0,4176
Fluminense	45	1,11E-05	-1,427	1,435	0,01261	-1,452	1,429	-0,05561	-1,462	1,361	0,03427	-0,4945	0,5856
Grêmio	44	0,0458	-1,385	1,476	0,03201	-1,366	1,462	-0,01087	-1,462	1,452	0,03918	-0,4904	0,5825
Corinthians	42	-0,02979	-1,457	1,398	-0,02945	-1,493	1,421	-0,008975	-1,43	1,403	-0,03627	-0,56	0,5049
Coritiba	42	0,04628	-1,381	1,483	0,03859	-1,376	1,478	-0,01948	-1,433	1,415	0,08522	-0,4469	0,6367
Internacional	41	0,1264	-1,3	1,563	0,1274	-1,284	1,526	0,001194	-1,412	1,425	0,1164	-0,4169	0,6709
Botafogo	39	0,02102	-1,395	1,446	0,02638	-1,373	1,452	-0,03944	-1,506	1,39	0,1077	-0,4243	0,663
Sport	36	0,01119	-1,405	1,446	-0,007165	-1,43	1,403	0,02055	-1,393	1,481	-0,2097	-0,7209	0,3221
Flamengo	35	-0,07332	-1,507	1,361	-0,07366	-1,515	1,377	-0,0867	-1,469	1,324	0,2021	-0,3384	0,765
Santos	35	-0,134	-1,556	1,297	-0,1282	-1,528	1,325	-0,1254	-1,531	1,294	0,06602	-0,4678	0,6163
Atlético-PR	32	0,08411	-1,342	1,51	0,08286	-1,361	1,518	-0,03545	-1,433	1,362	0,3013	-0,2437	0,8679
Palmeiras	32	-0,1633	-1,579	1,266	-0,1397	-1,586	1,31	-0,1714	-1,58	1,223	0,3683	-0,1895	0,9453

Tabela 10: Resultados do parâmetro $\beta 4$

	Média	Intervalo	
Santa Cruz	0,3252	-0,1772	0,8273
Figueirense	0,225	-0,2825	0,7297
América-MG	0,2171	-0,2886	0,7236
Sport	0,2047	-0,303	0,7123
Vitória	0,176	-0,3344	0,689
Ponte Preta	0,1723	-0,3408	0,6866
Cruzeiro	0,1151	-0,3996	0,6327
Atlético-MG	0,08782	-0,4327	0,6051
Chapecoense	0,07421	-0,4452	0,5923
Corinthians	0,03095	-0,4927	0,5494
Fluminense	-0,03943	-0,5634	0,4858
Grêmio	-0,04186	-0,5657	0,4842
Santos	-0,0703	-0,601	0,4555
Coritiba	-0,08934	-0,6179	0,4408
Botafogo	-0,1117	-0,6467	0,42
Internacional	-0,1211	-0,6534	0,4081
Flamengo	-0,2058	-0,7474	0,3311
São Paulo	-0,2092	-0,7484	0,331
Atlético-PR	-0,3067	-0,8538	0,2348
Palmeiras	-0,3723	-0,9306	0,1813

4.6 Algumas observações e discussões sobre os modelos

Para facilitar a observação de se o modelo representa o que aconteceu de verdade no campeonato serão observados 3 pontos: o $\beta 1$ que indica se o fator campo tem relação significativa com os resultados, $\beta 2$ (fator ataque) e $\beta 3$ (fator defesa) do Palmeiras (campeão) e América-MG (último colocado).

Primeiramente observa-se que para todos os modelos $\beta 1$ é positivo e próximo de $0,5$ e $\beta 0$ é sempre próximo de 0, portanto tem-se que é esperado que um time por jogar em casa faça mais gols.

Analisando o fato de alguns times possuírem o mesmo número de gols pró ou o mesmo número de gols contra mas possuírem valores diferentes de fator ataque ou fator defesa, se trata do fato de que alguns times fazem gols em muitas rodadas e outros fazem muitos gols em poucas rodadas, assim, um time possui o ataque mais consistente pois está sempre marcando enquanto o outro time é muito eficaz apenas em algumas partidas, o mesmo vale para a defesa.

Observando a Tabela 8 que diz respeito ao fator ataque dos times, percebe-se que o melhor ataque da competição foi do Palmeiras e analisando os $\beta 2$ do mesmo percebe-se que o Palmeiras possui o melhor fator ataque em todos os modelos. Fazendo a mesma análise para o América-MG que possui o pior ataque do campeonato, nota-se que o mesmo têm o pior fator ataque entre os 20 times em todos os modelos. Assim tem-se indícios de que todos os modelos testados analisam bem o fator ataque.

Analisando agora a Tabela 9 do Modelo 2 ao Modelo 4 o time do América-MG é o time com maior fator defesa apesar de não ser o time mais vazado do campeonato pois levou gol em mais jogos do que a equipe do Santa Cruz (equipe mais vazada) e pela equação dos modelos a equipe com pior defesa possui o maior fator defesa. Por outro lado a melhor defesa possui o menor fator defesa e entre esses modelos o Palmeiras possui a melhor defesa apesar de ter sido vazado tanto quanto o Atlético-PR.

Já no Modelo 5 a pior defesa possui o menor fator defesa e o Santa Cruz possui o

menor fator defesa (o que condiz com o campeonato), enquanto o Palmeiras possui o maior fator defesa, sendo assim, a melhor defesa do campeonato.

Os modelos apresentados neste capítulo apresentam algumas limitações do ponto de vista esportivo como o fato do Modelo 1 mostrar apenas a diferença que o mando de campo faz na quantidade de gols feitos por um time.

O Modelo 2 é um modelo mais completo do que o primeiro pelo fato de apresentar o fator ataque e defesa do time e assim melhorar a previsão da quantidade de gols marcados pelo time a cada rodada, porém o fato de não levar em conta o time adversário não é o melhor modelo para o estudo proposto, afinal o adversário do time faz toda diferença na quantidade de gols marcados e observando apenas os fatores de um dos times seria previsto apenas o saldo do time na partida.

O modelo 3 além dos fatores ataque e defesa do time observado também leva em conta os fatores ataque e defesa do time adversário sendo assim esportivamente é um modelo mais completo que o segundo, porém a utilização do fator ataque e defesa dos dois times intuitivamente retorna a previsão do saldo da partida. Na verdade o número de gols que cada time faz é explicado pela interação dos fatores ataque e defesa do próprio time e do rival da rodada.

O Modelo 4 é um desmembramento do Modelo 3 onde são levados em conta os fatores ataque dos dois times e o fator defesa apenas do time da casa.

O Modelo 5 intuitivamente é o melhor modelo para o que este estudo se propõe a prever já que levando-se em conta se o time é mandante ou visitante, o fator ataque do time e o fator defesa do adversário, imagina-se que dentre os 5 modelos propostos este tenha a melhor previsão para a quantidade de gols marcados por cada time.

Para a escolha do melhor modelo foi calculada a função desvio apresentada no Capítulo anterior e definida em McCullagh-Nelder (McCullagh-Nelder, 1989) [1], obtendo os seguintes valores:

Tabela 11: Função desvio

Modelo 2	Modelo 3	Modelo 4	Modelo 5
-2988,517695	-3011,043244	-3002,434917	-3010,904911

Com isso têm-se que o Modelo 3 foi o que melhor se ajustou aos dados já que gerou o menor valor de desvio e o Modelo 5 ficou bem próximo do Modelo 3 e portanto também seria um bom modelo a ser utilizado.

Com os resultados obtidos, ocorre uma pequena surpresa em relação ao melhor modelo a ser utilizado, pois intuitivamente era esperado que o Modelo 5 se ajustasse melhor aos resultados obtidos na prática, porém o Modelo 3 que possui os fatores ataque e defesa de ambos os times foi o que previu melhor estes resultados.

As figuras 13 e 14 abaixo mostram os valores previstos pelo Modelo 3 para a quantidade de gols rodada a rodada para os times do Palmeiras e América-MG respectivamente:

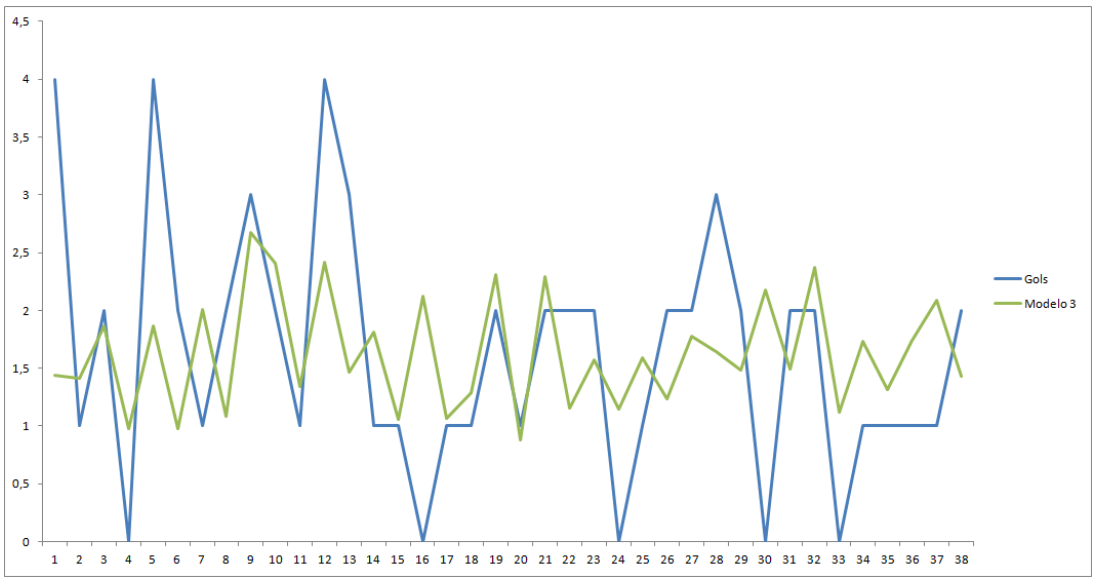

Figura 13: Gráficos da previsão para o Palmeiras

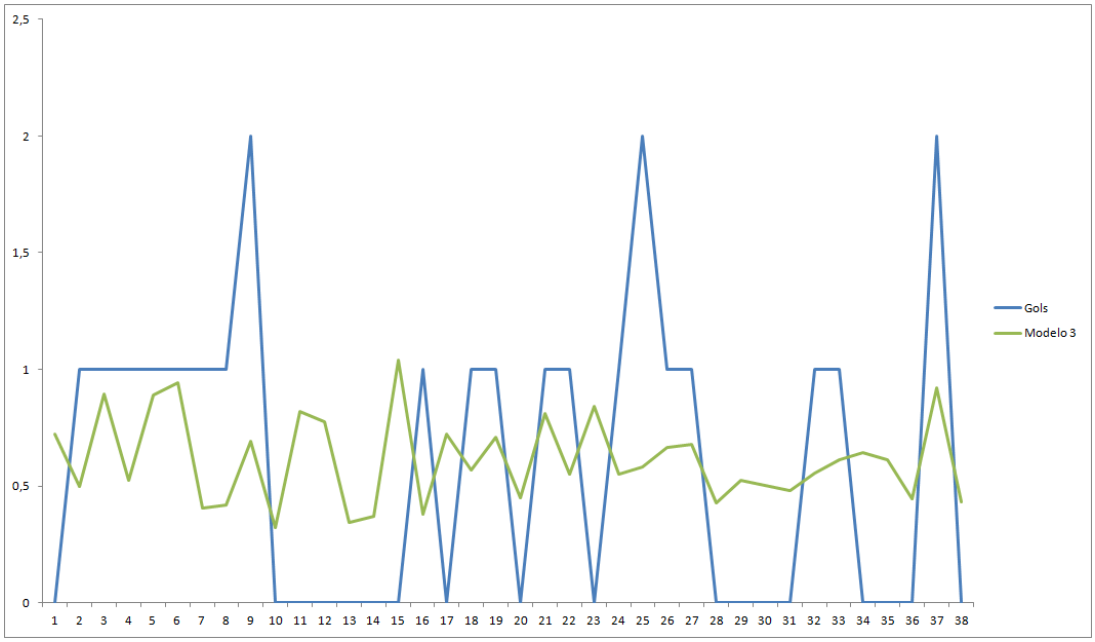

Figura 14: Gráficos da previsão para o América-MG

Pode-se observar que os gráficos não seguem exatamente o que aconteceu na realidade mas que seguem um padrão parecido e mesmo não fazendo boas previsões, o fator ataque consegue reproduzir a tendência dos gols marcados pelos times.

Uma opção para melhorar a qualidade da previsão dos gols poderia ser a utilização de modelos dinâmicos, como proposto por Souza e Gamerman (Souza e Gamerman, 2004) [8], o que constitui uma opção para a continuação deste trabalho.

5 Conclusão

Foram propostos 5 modelos lineares generalizados que incluem os efeitos aleatórios de defesa, ataque além da variável regressora mando de campo. Após testar os 5 modelos, verifica-se que os modelos mesmo não possuindo uma boa previsão eles capturam características dos times como por exemplo o fator ataque. O modelo que melhor se ajustou aos dados foi o Modelo 3 que possui os fatores ataque e defesa tanto do time mandante quanto do time visitante.

Referências

[1] P. McCullagh and. J .A. Nelder. *Generalized. Linear Models.* Chapman and Hall, 1989.

[2] Confederação Brasileira de Futebol. *Tabela de classificação do campeonato brasileiro de 2016.* CBF, 2016.

[3] Fábio Figueiredo Farias. *Análise e Previsão de Resultados de Partidas de Futebol.* UFRJ, 2008.

[4] Dani Gamerman. *Markov Chain Monte Carlo.* Chapman and Hall, 1999.

[5] Ghosh S.K. Gelfand, A.E. *Model choice: a minimun posterior predictive loss approach.* Biometrika, 1998.

[6] Leonhard Knorr-Held. *Dynamic Rating of Sports Teams.* Institute für Statistik, 1997.

[7] O RUE, H e SALVESEN. *Prediction and retrospective analysis of soccer matches in a league.* Norwegian University of Science and Technology, 2000.

[8] D SOUZA JR, O. G. e GAMERMAN. *Previsão de partidas de futebol usando modelos dinâmicos.* Anais do XXXVI SBPO, 2004.

ANEXO A – Scripts do OpenBUGS para os modelos utilizados

Modelo 1:

```
model
    {
       beta0 ~ dnorm(0, 0.001)
       beta1 ~ dnorm(0, 0.001)
       for (i in 1:N) {
for(j in 1:M){
          Y[i, j] ~ dpois(lambda[i,j] )
          log(lambda[i,j]) <- beta0 + beta1* X[i,j]
       }
}

    }
```

Modelo 2:

```
model
    {
       beta0 ~ dnorm(0, 0.001)
       beta1 ~ dnorm(0, 0.001)
       for (i in 1:N) {
for(j in 1:M){
          Y[i, j] ~ dpois(lambda[i,j] )
```

```
          Z[i, j] ~ dpois(lambda[i,j] )
          log(lambda[i,j]) <- beta0 + beta1* X[i,j]+beta2[i]-beta3[i]
       }
beta2[i]~dnorm(0,1)
beta3[i]~dnorm(0,1)
}
    }
```

Modelo 3:

```
model
    {
       beta0 ~ dnorm(0, 0.001)
       beta1 ~ dnorm(0, 0.001)
       for (i in 1:N) {
for(j in 1:M){
          Y[i, j] ~ dpois(lambda[i,j] )
          Z[i, j] ~ dpois(lambda[i,j] )
          log(lambda[i,j]) <- beta0 + beta1* X[i,j]+beta2[i]-beta3[i]-beta2[rival[i
       }
beta2[i]~dnorm(0,1)
beta3[i]~dnorm(0,1)
}
    }
```

Modelo 4:

```
model
    {
       beta0 ~ dnorm(0, 0.001)
       beta1 ~ dnorm(0, 0.001)
       for (i in 1:N) {
for(j in 1:M){
          Y[i, j] ~ dpois(lambda[i,j] )
          Z[i, j] ~ dpois(lambda[i,j] )
```

```
          log(lambda[i,j]) <- beta0 + beta1* X[i,j]+beta2[i]-beta3[i]+beta4[rival[i
       }
beta2[i]~dnorm(0,1)
beta3[i]~dnorm(0,1)
beta4[i]~dnorm(0,1)
}
    }
```

Modelo 5:

```
model
    {
       beta0 ~ dnorm(0, 0.001)
       beta1 ~ dnorm(0, 0.001)
       for (i in 1:N) {
for(j in 1:M){
          Y[i, j] ~ dpois(lambda[i,j] )
          Z[i, j] ~ dpois(lambda[i,j] )
          log(lambda[i,j]) <- beta0 + beta1* X[i,j]+beta2[i]-beta3[rival[i,j]]
       }
beta2[i]~dnorm(0,1)
beta3[i]~dnorm(0,1)
}
    }
```