Trabalho Final

Raquel Yuri da Silveira Aoki

14 de Dezembro 2015

Disciplina: Modelos Gráficos Probabilísticos

Professor: Renato Martins Assunção

1 Introdução

O trabalho final é baseado no artigo [1], em que Baio e Blangiardo desenvolveram um modelo hierárquico Bayesiano no futebol com dois objetivos: modelar o número de gols marcados em cada jogo de um campeonato.

O modelo do artigo foi ajustado aos resultados dos jogos do campeonato *Italian Serie A* na temporada 1991-1992. No trabalho final da disciplina de Modelos Gráficos Probabilísticos será feita a tentativa de ajustar o modelo nos dados da Séria A do Campeonato Brasileiro de 2014.

Nesse relatório final será apresentada uma descrição da base de dados na Seção 2, a modelagem utilizada na Seção 2 e as considerações finais podem ser encontradas na Seção 4.

2 Base de Dados

A base de dados com os 380 jogos da Série A do Campeonato Brasileiro de 2015 foi coletada no site *www.tabeladobrasileirao.net/2014/* e suas 5 primeiras linhas são mostradas na Tabela 1. A coluna 'jogo' é o número do jogo, seguido pelos nomes dos times mandante e visitante e seus respectivos placares.

Nos jogos da Série A do Campeonato Brasileiro de 2014, foram marcados em média 2.26 gols por jogo, sendo que o time mandante marcou em média 1,42 gols e o visitante 0,84 gols. Nas Tabelas 5 e 6 mostradas no Anexo é possível ver a frequência de gols que cada time fez como visitante ou mandante.

Tabela 1: Parte dos dados que serão utilizados

Jogo	**Mandante**	**Visitante**	**Gols**	
			Man.	**Vis.**
1	Internacional	Vitória	1	0
2	Fluminense	Figueirense	3	0
3	Chapecoense	Coritiba	0	0
4	Atlético-MG	Corinthians	0	0
5	São Paulo	Botafogo	3	0

A Tabela 2 mostra a média de gols de cada time como mandante e visitante. Observa-se que somente o time Bahia tem média de gols como visitante maior que a média de gols como mandante. Foi feito um teste t pareado para verificar se a diferença entre a média de gols como mandante e visitante dos times é igual a 0, e ao nível de 5% de singnificância o teste indica que deve-se rejeitar essa hipótese nula (valor-p = 0,00).

Tabela 2: Média de gols de cada time como mandante e visitante.

Equipe	**Média de Gols**	
	Mandante	**Visitante**
Atlético-MG	1.47	1.21
Atlético-PR	1.26	1.00
Bahia	0.74	0.89
Botafogo	1.26	0.37
Chapecoense	1.26	0.79
Corinthians	1.68	0.89
Coritiba	1.26	0.95
Criciúma	1.00	0.47
Cruzeiro	2.26	1.26
Figueirense	1.26	0.68
Flamengo	1.53	0.89
Fluminense	2.11	1.11
Goiás	1.58	0.42
Grêmio	1.26	0.63
Internacional	1.95	0.84
Palmeiras	1.11	0.68
Santos	1.32	0.89
São Paulo	1.68	1.42
Sport	1.21	0.68
Vitória	1.21	0.74

3 Modelo

No artigo original, são ajustados dois modelos de regressão de Poisson com um *framework* Bayesiano. Em um dos modelos, é feito o ajuste no número de gols feitos pelos times como mandantes, e no outro, como visitantes. Um dos motivos para fazer dois modelos, é que como visto na Seção 2, o número de gols marcados pelos times na condição de mandante e visitante é significativamente diferente.

Considere que g representa o g-ésimo jogo(g=1,...,G=380) e que no g-ésimo jogo o número de gols marcados pelo dos time mandante e visitante é representado por y_{g1} e y_{g2} respectivamente. Assim:

$$y_{gj|\theta_{gj}} \sim Poisson(\theta_{gj}) \tag{1}$$

em que os parâmetros $\Theta = (\theta_{g1}, \theta_{g2})$ representam o escore da intensidade do g-ésimo jogo para o time mandante (j=1) e o visitante (j=2) respectivamente.

A formulação desses parâmetros foi feita de acordo com o usado na literatura [2] e com o artigo de Baio e Blangiardo, sendo assumido um modelo de efeito aleatorio log-linear:

$$\begin{aligned} log\theta_{g1} &= home + att_{h(g)} + def_{a(g)} \\ log\theta_{g2} &= att_{a(g)} + def_{h(g)} \end{aligned} \tag{2}$$

em que att e def representam o ataque e a defesa do g-ésimo jogo. No caso do time mandante, o efeito depende do ataque do time mandante ($h()$), da defesa do time visitante($a()$) acrescido de uma vantagem por jogar em casa ($home$); já o efeito do time visitante depende do ataque do time visitante e da defesa do time mandante. A Figura mostra a representação em DAG do modelo hierárquico.

Resumidamente, para cada um dos G jogos, será calculado o valor de θ_{g1} e θ_{g2}. Feito isso, são ajustados dois modelos bayesianos de Poisson, o primeiro usa como variável resposta o número de gols dos times mandantes e o efeito θ_{g1}; o segundo usa o número de gols dos times visitantes e o efeito θ_{g2}.

3.1 Distribuições *a priori*

Em [1], os autores ajustam o modelo usando duas configurações de distribuições *a priori* sobre os efeitos aleatorios. Entretanto, essas configurações não eram muito informativas e não geraram resultados satisfatórios. Por isso, a informação dada *a priori* sobre

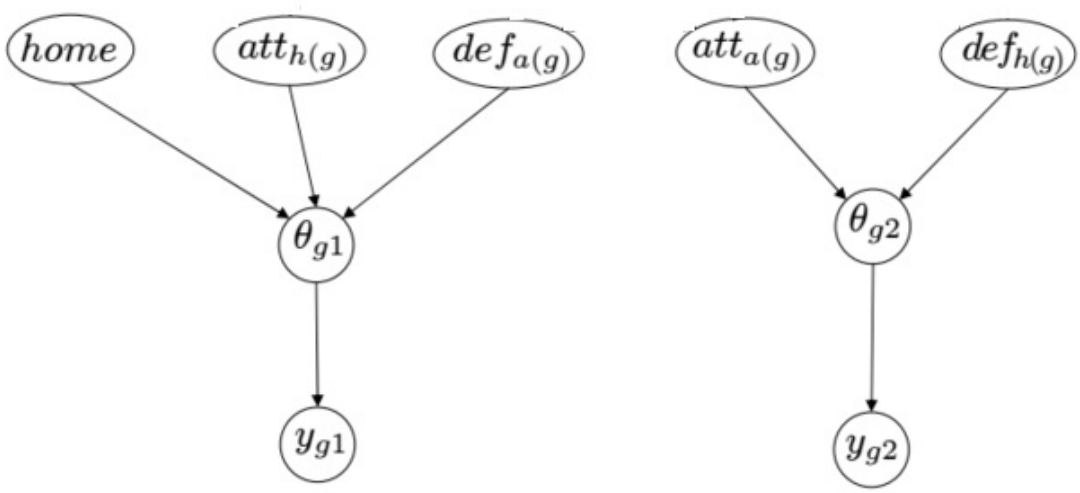

Figura 1: A representação em DAG do modelo hierárquico

o ataque e a defesa dos times mandante e visitante foi trocada para uma mais informativa.

Nessa nova configuração *a priori*, os termos de ataque e a defesa utilizandos na formula para calcular θ para cada time mandante e visitante são gerados aleatoriamente de uma distribuição Normal com desvio padrão de 0.01. A média dessa distribuição varia de acordo com a situação do time no jogo (mandante ou visitante) e seu calculo é mostrado na Tabela 3:

Tabela 3: Formula para o calculo da média de cada time no ataque e na defesa, como mandantes e visitantes.

Termo	Situação	Calculo da média
Ataque	Mandante	$\frac{gols_{mar.mand_i} - mean(gols_{mar.mand})}{sd(gols_{mar.mand})}$
	Visitante	$\frac{gols_{mar.vis_i} - mean(gols_{mar.vis})}{sd(gols_{mar.vis})}$
Defesa	Mandante	$\frac{gols_{sof.mand_i} - mean(gols_{sof.mand})}{sd(gols_{sof.mand})}$
	Visitante	$\frac{gols_{sof.vis_i} - mean(gols_{sof})}{sd(gols_{sof.vis})}$

Na Tabela 3, se um time i está no ataque e é visitante, a média da sua normal é calculada como o total de gols feitos pelo time i no campeonato como visitante, menos a média do total de gols marcados por todos os times como visitantes dividido pelo desvio padrão do total de gols marcados por todos os times como visitantes.

Note que essa distribuição *a priori* pode utilizar os gols marcados em Campeonatos Brasileiros de anos anteriores, como 2013, 2012, ou mesmo outros campeonatos, como

Copa do Brasil. Entretanto, é exigido um maior tratamento dos dados, pois a configuração dos times de um ano para outro dos campeonatos (ou entre os campeonatos) pode ser diferente. Para evitar esse problema, nessa análise foi usado os gols marcados no Campeonato Brasileiro de 2014, mesmo sabendo que essa pode não ser a melhor solução.

3.2 Ajuste do Modelo

O *software* utilizado para fazer as análises foi o R, com o auxilio do pacote $hSDM$, que é especialmente desenvolvido para modelos de distribuições hierárquicos Bayesiano. A função usada foi $hSDM.poisson$, com um *burnin* de 1000 e mcmc igual a 5000.

Após fazer os ajustes dos dois modelos com os dados provenientes do Campeonato Brasileiro 2014 Série A, foram tomados os valores preditos para cada observação, isto é, para cada um dos 380 jogos do campeonato e foi montada uma nova base de dados com os resultados dos jogos preditos pelo modelo. Com essa nova tabela de resultados preditos dos jogos, foi calculado como seria o resultado do campeonato e comparado com o observado, para uma validação da predição *Posteriori*.

A Figura 2 mostra um gráfico de dispersão entre a posição de cada time observada e predita e a linha vermelha representa o *Lowess*. Nota-se que a maioria dos times ficou a uma distância média de 2.5 colocações do observado, sendo que 3 times ficaram exatamente na colocação observada. A correlação entre a posição observada e predita foi de 0.84 e a posiçao de cada time real e predita pode ser vista na Tabela 4.

Para cada time, foi feito um gráfico com os pontos acumulados ao longo das rodadas do campeonato observado e predito. Os resultados podem ser vistos nas Figuras 3 e 4. Algumas equipes tiveram seu desempenho bem ajustado pelo modelo, como o Atlético-MG, Atlético-PR, Fluminense, Internacional e Santos; algumas outras equipes, o modelo não conseguiu fazer previsões de modo que acompanhassem bem os valores observados, como pode ser visto no Corinthians e Grêmio.

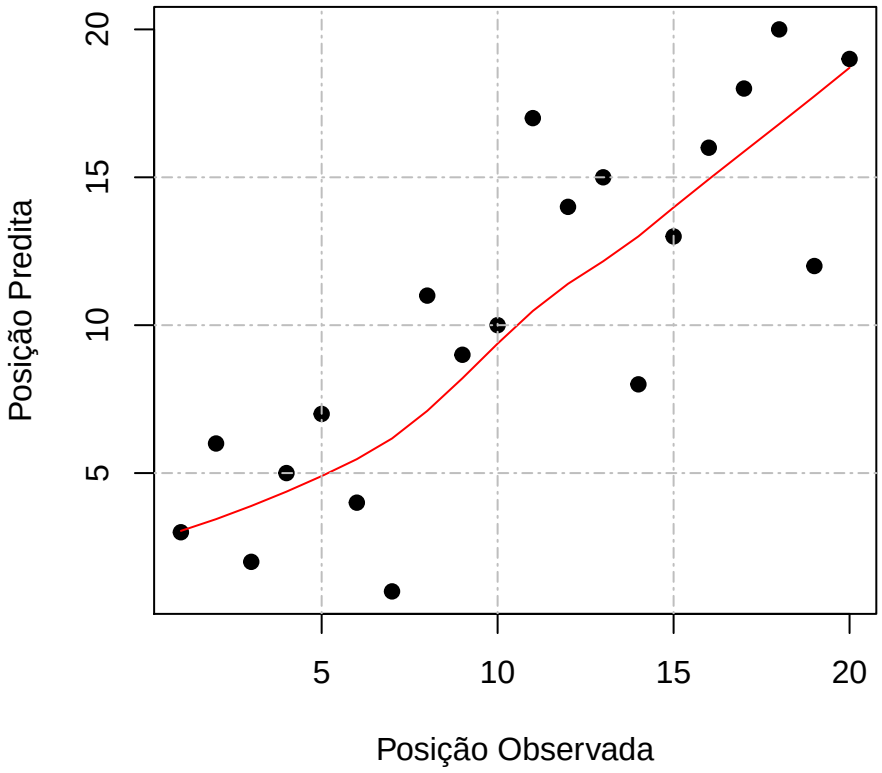

Figura 2: Correlação entre a colocação observada dos times e a predita pelo modelo para o Campeonato Brasileiro de 2014.

Tabela 4: Comparação das posições preditas e observadas do Campeonato Brasileiro 2014

Time	**Colocação**	
	Observada	**Predita**
Cruzeiro	1	3
São Paulo	2	6
Corinthians	3	2
Internacional	4	5
Atlético-MG	5	7
Fluminense	6	4
Grêmio	7	1
Atlético-PR	8	11
Santos	9	9
Flamengo	10	10
Sport	11	17
Coritiba	12	14
Figueirense	13	15
Goiás	14	8
Chapecoense	15	13
Palmeiras	16	16
Vitória	17	18
Bahia	18	20
Botafogo	19	12
Criciúma	20	19

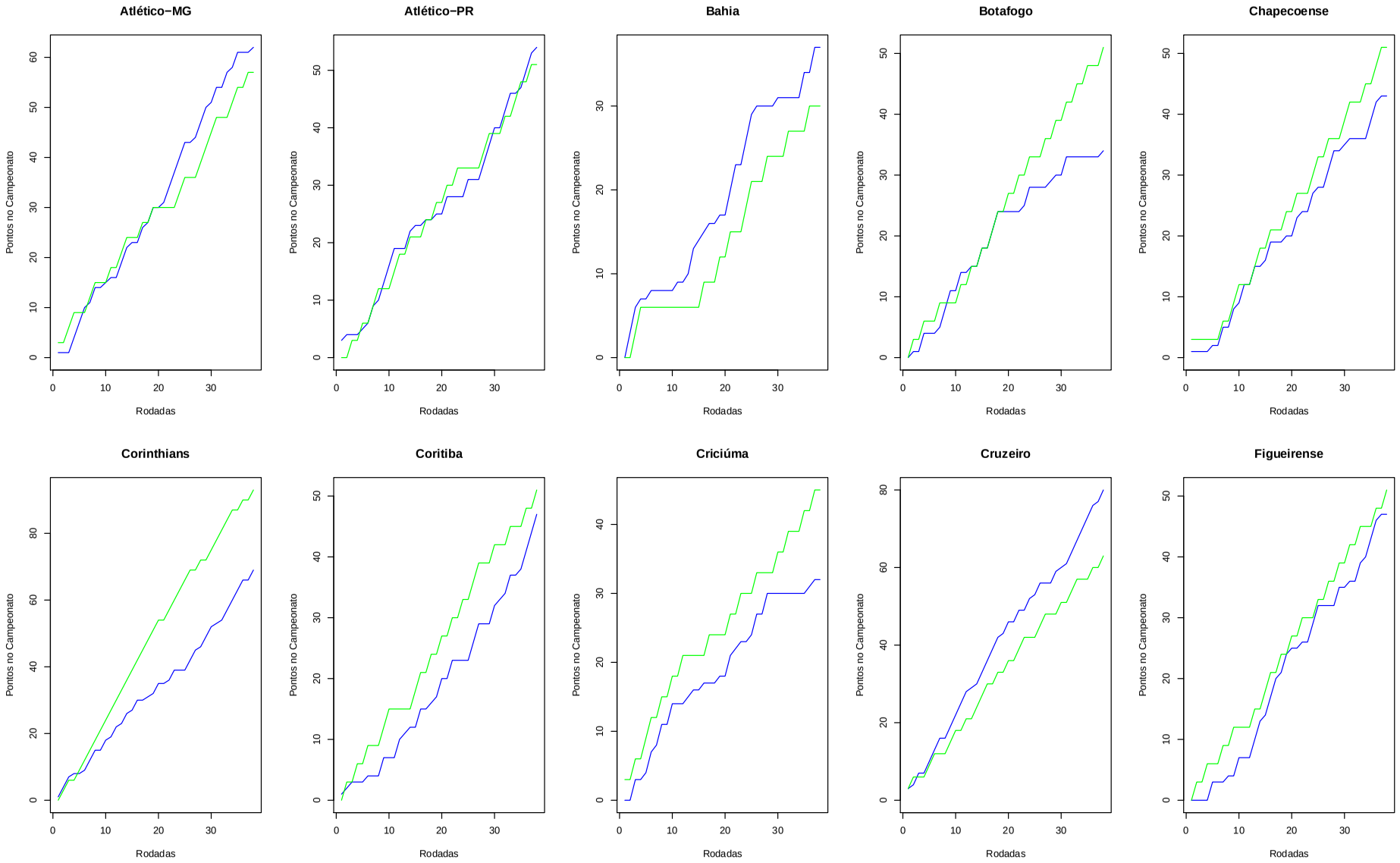

Figura 3: Validação da predição da posteriori: a linha azul representa os pontos acumulados observados no Campeonato Brasileiro de 2014 e a linha verde o predito pelo modelo - PARTE 1.

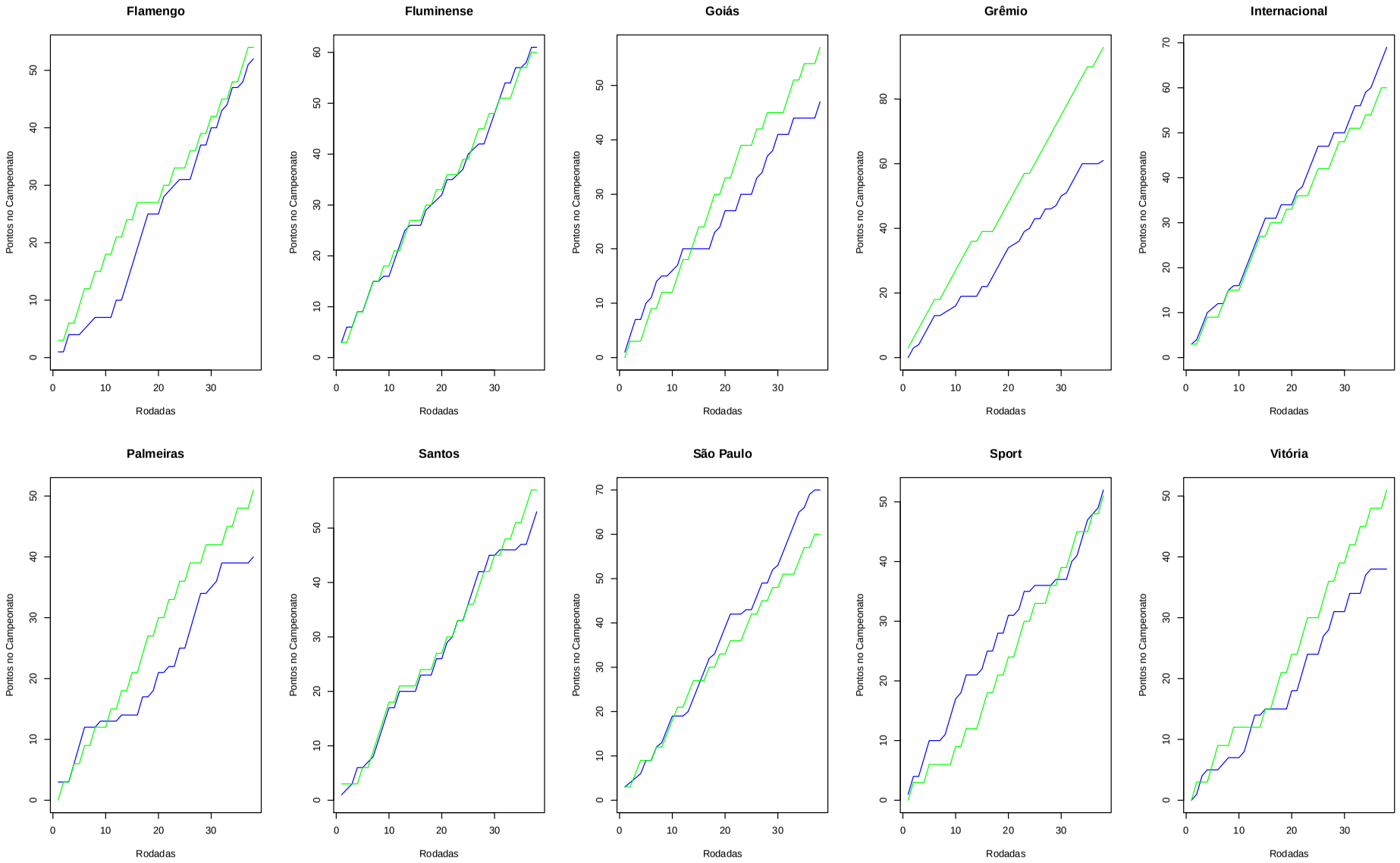

Figura 4: Validação da predição da posteriori: a linha azul representa os pontos acumulados observados no Campeonato Brasileiro de 2014 e a linha verde o predito pelo modelo - PARTE 2.

4 Conclusão

O objetivo desse trabalho era replicar o ajuste do modelo Bayesiano Hierarquico feito em [1] nos jogos do Campeonato Brasileiro de 2014. As análises iniciais feitas com as ditribuições *a priori* propostas no artigo não geraram bons resultados, e por isso, uma nova distribuição *a priori* foi utilizada.

Após o ajuste do modelo, foi calculado como seria o resultado do campeonato a partir dos resultados preditos para os 380 jogos. Como pode ser visto nas Tabelas e Figuras mostradas na Seção 3, os resultados foram satisfatórios, pois existe uma correlaçao positiva(0.84) entre as posições observadas e as obtidas a partir das predições do modelo. Além disso, comparando os pontos acumulados ao longo das rodadas observados e preditos pelo modelo, observa-se que para a maior parte das equipes, esses valores foram próximos.

Inicialmente, foi pensado utilizar o WinBUGS/OPENBUGS. Entretanto, por dificuldades encontradas essa ideia foi abondada e somente o R foi utilizado.

Referências

[1] G. Baio and M. Blangiardo. Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, 37(2):253–264, 2010.

[2] D. Karlis and I. Ntzoufras. Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393, 2003.

Anexo

Tabela 5: Frequência de gols marcados por cada time como visitantes

Equipe	Gols Marcados						
	0	1	2	3	4	5	6
Atlético-MG	4	7	4	3	1	0	0
Atlético-PR	4	7	7	1	0	0	0
Bahia	8	9	1	1	0	0	0
Botafogo	5	8	5	0	0	0	1
Chapecoense	6	7	3	2	0	1	0
Corinthians	1	9	6	2	0	1	0
Coritiba	6	6	3	4	0	0	0
Criciúma	7	8	1	3	0	0	0
Cruzeiro	1	3	8	5	1	1	0
Figueirense	4	10	2	2	1	0	0
Flamengo	4	8	2	3	2	0	0
Fluminense	2	7	3	3	2	2	0
Goiás	7	3	4	3	1	0	1
Grêmio	4	8	6	0	1	0	0
Internacional	1	5	9	2	2	0	0
Palmeiras	5	9	4	0	1	0	0
Santos	6	4	6	3	0	0	0
São Paulo	1	7	8	3	0	0	0
Sport	4	8	6	1	0	0	0
Vitória	7	3	7	2	0	0	0

Tabela 6: Frequência de gols marcados por cada time como visitantes

Equipe	Gols Marcados				
	0	1	2	3	4
Atlético-MG	5	8	3	3	0
Atlético-PR	6	8	4	1	0
Bahia	6	10	2	1	0
Botafogo	13	5	1	0	0
Chapecoense	10	5	3	0	1
Corinthians	8	7	3	0	1
Coritiba	9	3	6	1	0
Criciúma	11	7	1	0	0
Cruzeiro	5	7	4	3	0
Figueirense	9	8	1	1	0
Flamengo	8	5	6	0	0
Fluminense	5	9	3	2	0
Goiás	13	4	2	0	0
Grêmio	11	5	2	1	0
Internacional	6	10	3	0	0
Palmeiras	8	9	2	0	0
Santos	8	6	4	1	0
São Paulo	3	7	8	0	1
Sport	9	7	3	0	0
Vitória	9	7	2	1	0