Previsão de partidas de futebol usando modelos dinâmicos

Oswaldo Gomes de Souza Junior
Instituto de Matemática
Universidade Federal do Rio de Janeiro
junior@dme.ufrj.br

Dani Gamerman
Instituto de Matemática
Universidade Federal do Rio de Janeiro
dani@im.ufrj.br

Área: EST- Estatística

Resumo

Este trabalho consiste em realizar previsões de partidas de futebol. Para isso, são utilizados modelos contendo três fatores que explicam os números de gols feitos: ataque, defesa e campo. São utilizadas informações passadas para estimar os fatores citados. Os modelos são estimados através de simulações via MCMC (Gamerman, 1997). A partir destas simulações, calculam-se as distribuições preditivas de diferentes quantidades de interesse, como número de pontos para um time ser campeão e colocação de um dado time. Métodos de comparação entre diferentes sistemas de previsão são considerados e usados para avaliar nossas previsões. A metodologia é utilizada para analisar dados dos Campeonatos Brasileiros de futebol de 2002 e 2003.

Palavras chave: distribuição preditiva, inferência Bayesiana, verossimilhança.

Abstract

This work consists in forecasting football matches. Models containing three factors are used to explain the number of goals scored: attack, defense and home match. Past information is used to estimate the cited factors. Models are estimated through MCMC-based simulations (Gamerman, 1997). From these, predictive distributions for quantities of interest, such as number of points needed to become champion and placement of a given team are calculated. Methods for comparison between different forecasting systems are considered and used to evaluate our predictions. The methodology is used to analyse data from the 2002 and 2003 Brazilian football championships.

Key words: Bayesian inference, likelihood, predictive distribution.

1. Introdução

A Estatística vem tendo um grande avanço nos últimos anos em diversas áreas, porém mais notada em termos de quantidade aplicável a problemas reais. Uma de suas diversas aplicações é no estudo de resultados ocorridos em esportes, em particular, em jogos de futebol.

Quando são realizados campeonatos de futebol, há confrontos entre os times, e são considerados vencedores aqueles que, ao final do jogo, fazem mais gols. Se há igualdade no número de gols feitos por ambas as equipes, é porque o jogo terminou empatado.

O objetivo desse estudo é analisar o comportamento dos times através de resultados anteriores e fazer uma previsão para os jogos seguintes. Ou seja, estimar o número de gols que determinada equipe fará nas próximas partidas.

Já foram desenvolvidos vários trabalhos nessa área. Em particular podemos citar os trabalhos de Glickman (1993), Glickman e Stern (1998) e Knorr-Held (2000), que utilizam modelos dinâmicos (Harrion e Stevens, 1976; Pole, West e Harrison, 1994). Portanto, nosso objetivo principal é fazer com que os resultados obtidos sejam melhores do que os já encontrados.

Na seção 2 é descrito o modelo básico utilizado, ou seja, o modelo usado em uma fase inicial. Na seção seguinte, são mostrados os resultados de uma análise preliminar. A seção 4 fala sobre o modelo em sua fase final, fazendo extensão à parte dinâmica. Em seguida, são mostrados os resultados desta parte, seguidos de uma seção contendo as previsões citadas. O artigo é finalizado com as conclusões.

2. Modelo Inicial

Suponhamos que o interesse seria prever o resultado do jogo A x B. Através dos jogos passados, vamos obter os fatores de ataque e de defesa dos dois times. O Fator Ataque representa o comportamento do ataque de determinada equipe, ou seja, quantifica o número de gols feitos pelo time. O Fator Defesa se refere ao comportamento da defesa, ou seja, dá valores ao número de gols sofridos pela equipe. Além desses dois fatores, é usado também o Fator Campo. Esse fator é importante, pois através dele, podemos verificar se o time joga melhor em casa ou fora de casa. Dizem que para a maioria das equipes é melhor jogar em casa, mas pode ser que determinado time venha obtendo melhores resultados jogando fora de casa, durante determinado campeonato. Então é necessário colocar essa informação no modelo. Enfim, o modelo que será utilizado nesse estudo é o seguinte:

$$\begin{aligned} NGF_A &\sim Poiss(\lambda_A) \\ \log \lambda_A &= At_A - De_B + Ca_A \end{aligned} \tag{2.1}$$

onde

NGF_A representa o número de gols feitos pela equipe A;

At_A representa o Fator Ataque de A;

De_B representa o Fator Defesa de B;

Ca_A representa o Fator Campo de A;

Da mesma forma, para NGF_B, temos:

$$\begin{aligned} NGF_B &\sim Poiss(\lambda_A) \\ \log \lambda_B &= At_B - De_A \end{aligned} \tag{2.2}$$

3. Análise Preliminar

Esse estudo será feito com dados do Campeonato Brasileiro de 2002. Como análise preliminar, foi testado se a distribuição de Poisson ajustaria bem o número de gols no campeonato em questão. As figuras 1 e 2 mostram essa comparação. Em preto, temos os dados reais e em vermelho a verdadeira distribuição de Poisson. Podemos notar um ajuste muito bom tanto nos gols fora de casa, quanto nos gols marcados fora de casa.

Figura 1

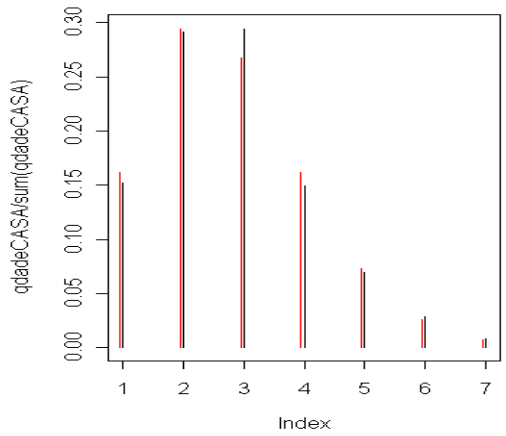
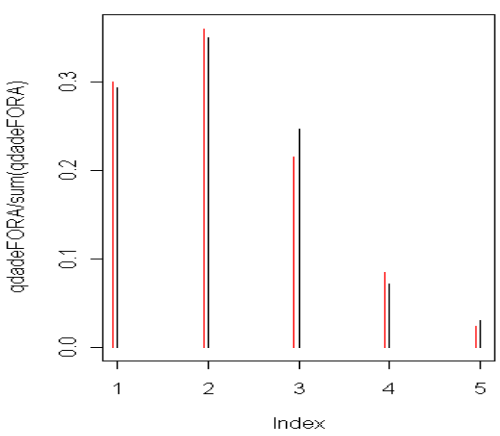


Figura 2



Além da comparação visual, foi feito também o teste de Bondade de Ajuste. E nesse teste foi aceita a hipótese de se ajustar os gols por uma Poisson com parâmetro λ, onde λ é a média de número de gols. Com os dados do Campeonato Brasileiro de 2002 em mãos, foram analisadas as 22 primeiras rodadas para ser feita a previsão para as 3 últimas da primeira fase.

Após fazer uma Análise de Regressão com os dados em questão, foram obtidos todos os fatores para os times do campeonato. Abaixo, temos esses fatores para os times do Rio de Janeiro:

Times	Fator Ataque	Fator Defesa	Fator Campo
Botafogo	-0.14	0.07	0.09
Flamengo	0.09	0.15	0.30
Fluminense	-0.03	0.24	0.61
Vasco	0.17	-0.01	-0.03

Podemos ver através da tabela acima que, até a 22ª rodada, o melhor ataque, entre os cariocas, era o do time do Vasco, já que possui o maior *Fator Ataque*. Da mesma forma, vemos que o Fluminense possui a pior defesa, uma vez que possui o maior *Fator Defesa*, ou seja, sofreu mais gols entre os times do Rio. Quanto ao *Fator Campo*, podemos dizer que o time que mais sente a diferença entre jogar em casa e jogar fora de casa é o Fluminense. Obteve melhores resultados jogando em casa.

4. Modelo Dinâmico

Até aqui, foi usado apenas o modelo estático, ou seja, onde não se permitia variar os fatores. A partir de agora, vamos dar um outro tratamento ao modelo, permitindo, dessa forma, uma variação dos fatores com o intuito de melhorar os resultados.

Com base no modelo descrito acima, serão analisadas as rodadas iniciais do campeonato em estudo para poder estimar resultados para as partidas seguintes. Através dos fatores avaliados, faremos a previsão.

Assim como em Knorr-Held (2000), para avaliar o Fator Ataque do time A, por exemplo, será usado um modelo dinâmico, de acordo com a seguinte equação

$$At_A^{i+1} = At_A^i + \omega_{i+1}\,, \tag{4.1}$$

onde $\omega_{i+1} \sim N(0, \sigma^2)$.

O mesmo raciocínio é utilizado para os *Fatores Defesa* e *Campo*.

$$De_A^{i+1} = De_A^i + \omega_{i+1} \tag{4.2}$$

$$Ca_A^{i+1} = Ca_A^i + \omega_{i+1} \tag{4.3}$$

Queremos dizer, com esse modelo, que os fatores vão sendo atualizados à medida que os jogos vão acontecendo, ou seja, ao longo das rodadas. O fator em questão, no tempo *i*, é esse fator no tempo *i-1* somado a um erro aleatório. Como não temos informação nenhuma na primeira rodada, pois estamos utilizando apenas os dados do campeonato em questão, usamos uma *priori vaga*. Esta sempre é usada no modelo dinâmico quando não se tem "informação inicial". Em nosso modelo, utilizamos a seguinte *priori vaga*:

$$At_A^1 \sim N\left(0, \sigma^2\right), \tag{4.4}$$

onde a variância σ^2 é um valor alto para poder dar chances razoáveis para todos os valores possíveis. Seguindo o mesmo raciocínio, temos:

$$De_A^1 \sim N\left(0, \xi^2\right) \tag{4.5}$$

$$Ca_A^1 \sim N\left(0, \delta^2\right) \tag{4.6}$$

Da mesma forma, seguimos raciocínio semelhante para as distribuições *a priori* do time B.

Assim, os fatores são atualizados à medida que as rodadas vão acontecendo, ou seja, os fatores variam com o tempo. Desta forma passamos a usar (4.1), (4.2) e (4.3) em (2.1) e temos a seguinte verossimilhança:

$$NGF_A^i \sim Poiss\left(\lambda_A^{i-1}\right) \tag{4.7}$$

$$\log \lambda_A^{i-1} = At_A^{i-1} - De_B^{i-1} + Ca_A^{i-1} + \varepsilon_A^{i-1}\,, \tag{4.8}$$

onde $\varepsilon_A^i \sim N\left(0, \sigma^2\right), \quad \forall i = 2, ..., n$

Da mesma forma, também temos:

$$NGF_B^i \sim Poiss\left(\lambda_B^{i-1}\right)$$

$$\log \lambda_B^{i-1} = At_B^{i-1} - De_A^{i-1} + \tau_B^{i-1},$$

onde $\tau_B^i \sim N\left(0, \sigma^2\right), \quad \forall i = 2,..., n$

5. Previsões

Considere a seguinte notação:

$$\theta = (\theta^1 , \dots , \theta^n) \qquad (5.1)$$

onde
$\theta^i = (At^i , De^i , Ca^i)$,
$At^i = (At^i_{\text{Atletico-MG}} , \dots , At^i_{\text{Vitória}})$ é o vetor com os fatores ataque para os times na rodada i,
$De^i = (De^i_{\text{Atletico-MG}} , \dots , De^i_{\text{Vitória}})$ é o vetor com os fatores defesa para os times na rodada i,
$Ca^i = (Ca^i_{\text{Atletico-MG}} , \dots , Ca^i_{\text{Vitória}})$ é o vetor com os fatores campo para os times na rodada i,
$NGF^i = (NGF^i_{\text{Atletico-MG}} , \dots , NGF^i_{\text{Vitória}})$ é o vetor com os nos. de gols dos times na rodada i,
$D^i = \{ NGF^1,\dots, NGF^i \}$ é o conjunto contendo toda a informação até a rodada i, para i=1,...,n.

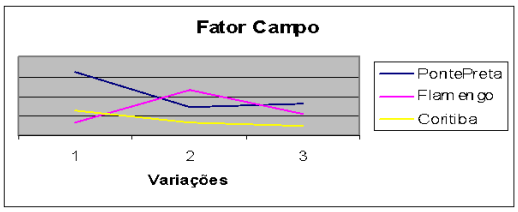
Em (5.1), temos que θ é o vetor que contém vetores com todas as *informações* dos times. Por exemplo, θ_1 é um vetor de tamanho 3 com as características (ataque, defesa e campo) do time 1. E assim para os demais times. Os times em questão são: Atlético MG, Atlético PR, Bahia, Botafogo, Corinthians, Coritiba, Cruzeiro, Figueirense, Flamengo, Fluminense, Gama, Goiás, Grêmio, Guarani, Internacional, Juventude, Palmeiras, Paraná, Paysandu, Ponte Preta, Portuguesa, Santos, São Caetano, São Paulo, Vasco e Vitória.

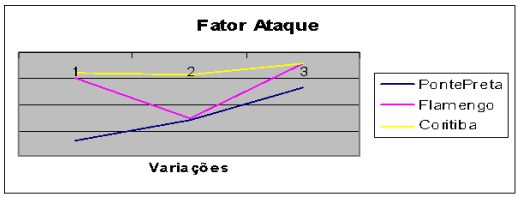
Utilizando o teorema de Bayes, a estimação dos parâmetros até a *rodada i*, será feita a partir da posteriori. Esse cálculo não é fácil e, da mesma forma que Glickman (1993), utilizamos o *WinBugs*, um pacote estatístico que usa MCMC, para extrair amostras *a posteriori,* que pode ser obtida da seguinte forma (ver DeGroot, 2002):

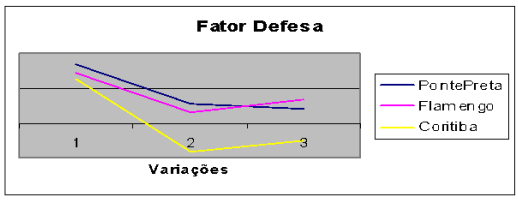
$$p\left(\theta^1,\dots,\theta^i \mid D^i\right) \propto L\left(\theta^1,\dots,\theta^i\right) p\left(\theta^1,\dots,\theta^i\right)$$

onde $L(\theta^1 , \dots , \theta^i)$ é a verossimilhança até a rodada i e $p(\theta^1 , \dots , \theta^i)$ é a priori até a rodada i.

Podemos exemplificar a utilização do modelo dinâmico com os parâmetros de 3 times: Coritiba, Flamengo e Ponte Preta. Foram feitas apenas 3 variações nas rodadas 15, 30 e 44 devido a limites computacionais.







Podemos, assim, escrever o modelo que faz previsão dos resultados, baseado na preditiva, da seguinte forma:

$$\underset{1}{p\left(NGF^{i+h} \mid D^{i}\right)} = \int \underset{2}{p\left(NGF^{i+h} \mid \theta^{i}, D^{i}\right)} \underset{3}{p\left(\theta^{i} \mid D^{i}\right)} d\theta^{i}$$

onde: $NGF^{i+h} \mid \theta^{i}, D^{i} \sim Poisson\left(\lambda^{i+h}\right)$

3 é obtido por simulação via MCMC, servindo de parâmetro para simular amostras de 2. Desta forma, automaticamente temos amostras de 1.

Essas simulações de Monte Carlo via Cadeia de Markov (MCMC) são utilizadas para resolver esse tipo de problema como dito anteriormente. Gamerman (1997) descreve o método detalhadamente.

Para analisar os resultados, poderiam ser investigados os diferentes placares (1x0, 2x1, ...), porém se tornaria complicado avaliar o modelo, uma vez que existe uma grande possibilidade de resultados. Uma maneira de contornar isso, é agregar os resultados da preditiva nos 3 eventos possíveis: vitória, empate e derrota.

Assim, com os resultados dos jogos previstos, poderemos realizar vários cálculos, em particular, achar o número de pontos que os times farão ao final do campeonato. Com a pontuação final de todos os times para todas as simulações feitas, podemos calcular as probabilidades de um time ser rebaixado, de um time se classificar pra libertadores, a pontuação mínima para um time ser campeão, dentre outros resultados de interesse.

Desta forma, podemos escrever

$NP_A^T = f\left(NGF^1, ..., NGF^T\right)$, onde NP^T_A é o número de pontos do time A na rodada final T.

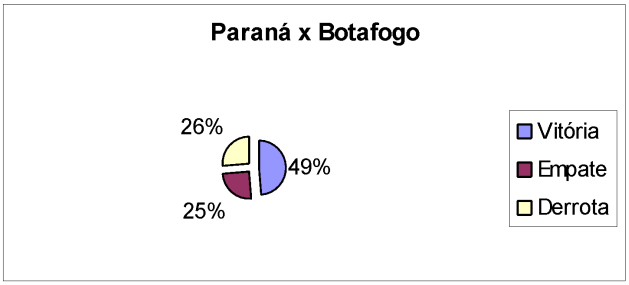
Cabe ressaltar que, qualquer função desse tipo pode ter sua distribuição aproximada por simulação.

6. Resultados

Utilizando o *WinBugs*, simulamos 1000 vezes os jogos das 3 últimas rodadas. Com os resultados obtidos, ou seja, a probabilidade preditiva, pegamos as proporções para os 3 eventos possíveis. Desta forma, estamos aproximando os resultados reais por quantidades teóricas. Abaixo, temos alguns exemplos para as previsões dos jogos em questão:

Placar Real: *Paraná 2x0 Botafogo*

A previsão encontrada para esse jogo foi a seguinte:

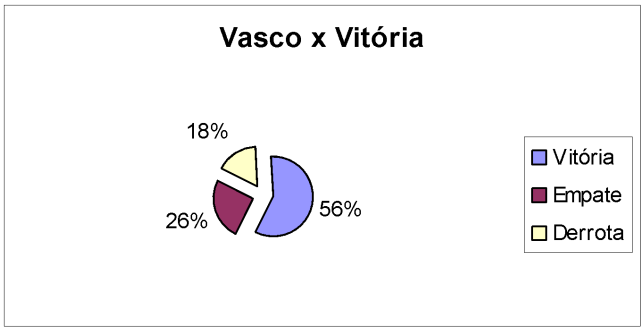


Podemos detalhar ainda mais:

Vitória		Empate		Derrota	
1x0	7.9%			0x1	3.8%
2x0	8.3%	0x0	4.9%	0x2	3.5%
2x1	9.9%	1x1	7.7%	1x2	5.9%
3x0	6.7%	2x2	6.3%	0x3	1.1%
3x1	5.8%	3x3	3.8%	1x3	3.0%
3x2	4.4%	Outros	2.3%	2x3	4.2%
Outros	6.0%			Outros	4.5%

Placar Real: *Vasco 4x1 Vitória*

A previsão encontrada para esse jogo foi a seguinte:



Podemos detalhar ainda mais:

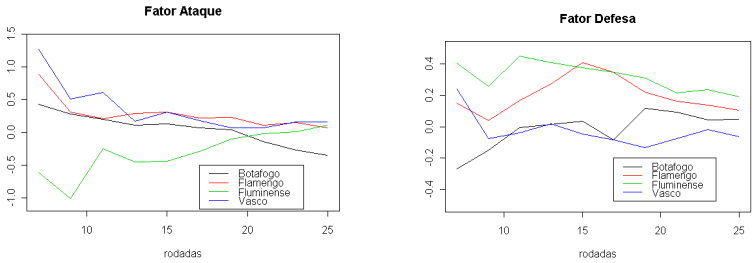
Vitória		Empate		Derrota	
1x0	13.9%			0x1	5.5%
2x0	11.3%	0x0	6.6%	0x2	3.3%
2x1	9.3%	1x1	8.4%	1x2	3.7%
3x0	5.8%	2x2	7.6%	0x3	1.1%
3x1	4.9%	3x3	0.9%	1x3	1.2%
3x2	3.5%	Outros	2.5%	2x3	1.6%
Outros	6.3%			Outros	1.6%
				Outros	2.1%

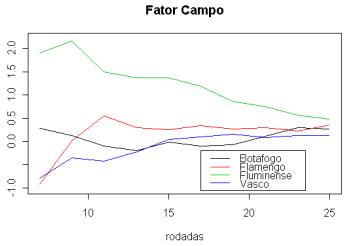
As chances de vitória, empate e derrota para os 39 últimos jogos da primeira fase do Campeonato Brasileiro de 2002 podem ser calculadas. Desses jogos, o modelo "acertou" 21 resultados, ou seja, obteve um aproveitamento de 54%. Da mesma forma, temos as previsões para a fase final:

Jogo					Vitoria	Empate	Derrota
Santos	3	x	1	SaoPaulo	44,8%	22,3%	32,9%
Fluminense	3	x	0	SaoCaetano	33,1%	24,6%	42,3%
AtleticoMG	2	x	6	Corinthians	50,7%	22,5%	26,8%
Gremio	0	x	0	Juventude	54,3%	27,6%	18,1%
SaoCaetano	2	x	0	Fluminense	73,1%	16,5%	10,4%
Corinthians	2	x	1	AtleticoMG	39,2%	21,9%	38,9%
Juventude	0	x	1	Gremio	46,3%	29,1%	24,6%
SaoPaulo	1	x	2	Santos	66,6%	13,8%	19,6%
Santos	3	x	0	Gremio	46,9%	26,4%	26,7%
Fluminense	1	x	0	Corinthians	45,6%	22,7%	31,7%
Gremio	1	x	0	Santos	45,1%	24,8%	30,1%
Corinthians	3	x	2	Fluminense	52,3%	23,9%	23,8%
Santos	2	x	0	Corinthians	47,8%	25,7%	26,5%
Corinthians	2	x	3	Santos	36,8%	25,2%	38,0%

Na fase final, o modelo "acertou" 9 dos 14 jogos, obtendo, portanto, 64% de aproveitamento.

Com esse estudo, podemos também analisar a variação dos fatores durante as rodadas. Para dar um exemplo disso, vamos verificar o comportamento desses fatores para os times do Rio:



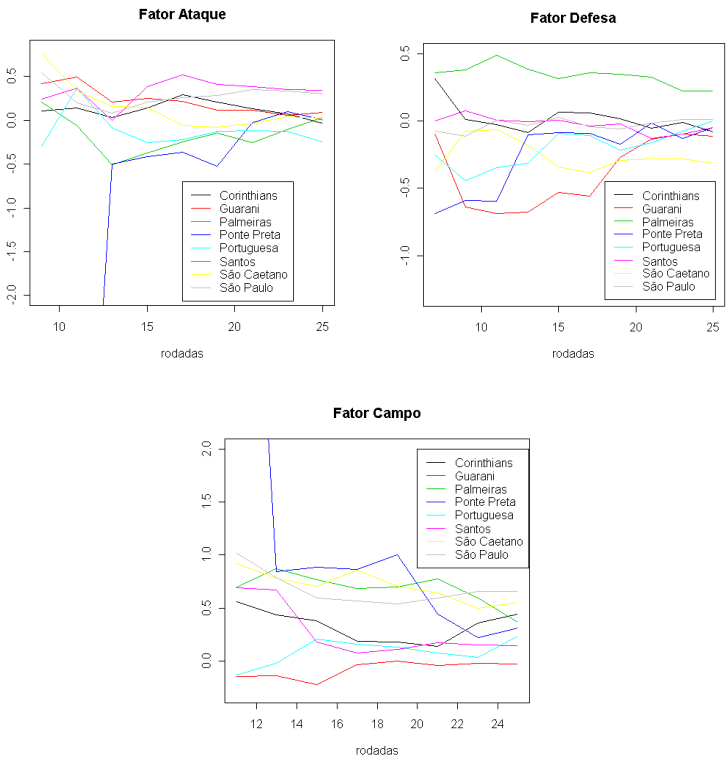


Através do gráfico do *Fator Ataque*, pode-se notar que o Fluminense melhorou seu *Fator Ataque*. Observa-se também, que, ao final da primeira fase, o Botafogo tem o pior ataque entre os cariocas, enquanto que para os outros times do Rio, esse fator fica em torno do mesmo valor.

No gráfico do *Fator Defesa*, tem-se que ao final da primeira fase o Vasco tem o menor *Fator Defesa*, ou seja, sofreu menos gols e, portanto, entre os cariocas, tem a melhor defesa. Seguindo o mesmo raciocínio, o Fluminense tem a pior defesa entre os times do Rio.

No último gráfico, *Fator Campo*, observa-se que para o Fluminense, ao longo do campeonato, faz mais diferença entre jogar em casa ou fora do que para Vasco, Flamengo ou Botafogo.

Para os times de São Paulo, temos:



Através do gráfico do Fator Ataque, podemos notar que a Ponte Preta tem o pior ataque no começo do campeonato e melhora seu desempenho no final da primeira fase. A Portuguesa tem um desempenho ruim ao longo da primeira fase e se destaca, negativamente, ao final, fator decisivo para seu rebaixamento. Notamos também, que Santos e São Paulo tem os melhores ataques na parte final da primeira fase.

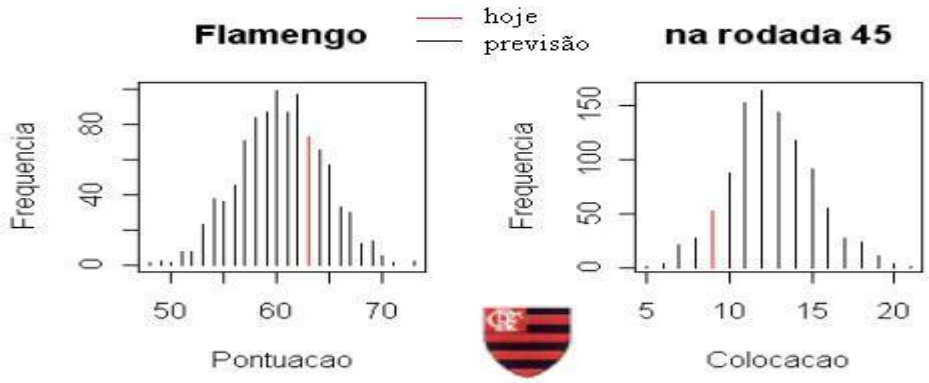
No gráfico do Fator Defesa, temos que o Palmeiras mantém a regularidade: sempre com a pior defesa, fator decisivo para seu rebaixamento. O Guarani, no meio do campeonato, pode ser considerado a equipe com a melhor defesa, posto que é perdido para o São Caetano no final da primeira fase.

No último gráfico, o do Fator Campo, observamos que para a equipe do São Paulo é melhor jogar em casa do que para os demais times. Para Portuguesa e Guarani, por exemplo, não faz diferença em jogar dentro ou fora de casa.

Gráficos similares podem ser construídos para todos os outros times que disputaram o campeonato. Podemos também, em qualquer rodada do campeonato, calcular as distribuições do número de pontos e da colocação de um dado time.

A partir de agora, utilizaremos os dados do campeonato brasileiro de 2003. Como o estudo feito para prever os resultados já foi visto anteriormente, agora será enfocada a parte final do estudo, ou seja, nos preocuparemos com as chances de rebaixamento, pontuação final,...

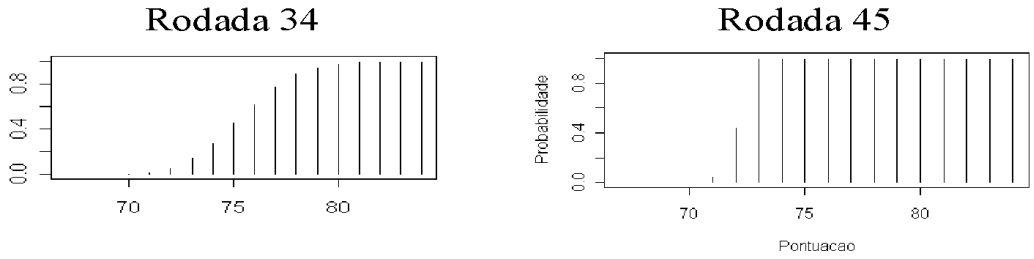
Foram cadastrados todos os resultados até a rodada 34 (considerada como sendo o dia de hoje) e então, feitas simulações para calcular as pontuações dos times no final do campeonato. Com as probabilidades calculadas, comparamos os resultados com o que aconteceu, de fato. E aqui apresentaremos, a título de ilustração, os resultados para o Flamengo:



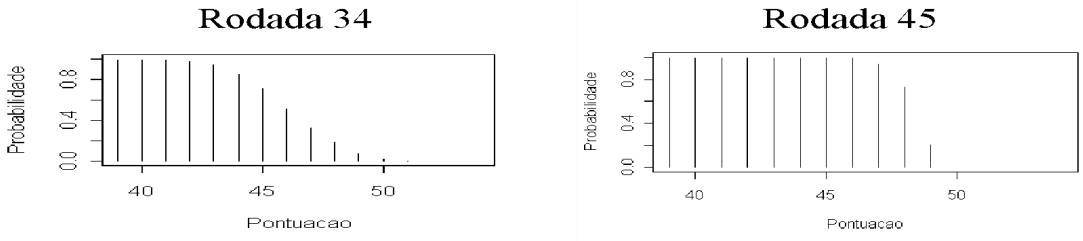
Na figura anterior, temos que a linha em vermelho é o resultado que de fato aconteceu, ou seja, o Flamengo fez 63 pontos e em preto é a distribuição de probabilidade da pontuação final deste time. O mesmo raciocínio é usado para a figura ao lado, só que com relação à posição final.

Outros resultados interessantes podem ser calculados. Por exemplo, são muito divulgados pela imprensa o mínimo de pontos que um time precisa para se classificar para a Copa Libertadores da América, campeonato sul-americano de prestígio, e o número de pontos que um time precisa para não ser rebaixado à 2ª. Divisão. Essas questões são resolvidas através do uso de distribuições preditivas.

Os gráficos abaixo mostram as chances de um time se classificar para a Libertadores com determinado número de pontos em duas rodadas distintas.



Os gráficos abaixo mostram as chances de um time ser rebaixado com determinado número de pontos em duas rodadas distintas.



Comparações de modelos podem ser feitas através das verossimilhanças preditivas, isto é, através do cálculo da probabilidade que um modelo dá para o que realmente aconteceu. No caso de previsões para vários eventos, podemos simplificar e calcular o produto das probabilidades que o modelo dá para os diferentes eventos observados. Para o campeonato em questão, foram comparadas as verossimilhanças por nós obtidas com aquelas fornecidas pelo site Chance de Gol (www.chancedegol.com.br). As verossimilhanças obtidas foram: $2{,}26 \times 10^{-17}$, para o "Chance de Gol" e $7{,}66 \times 10^{-17}$ para o nosso modelo. Isso mostra que nossas previsões foram cerca de 3 vezes mais corretas em termos probabilísticos.

7. Conclusões

É mais razoável a utilização desse modelo dinâmico, pois este se aproxima mais da realidade, uma vez que mostramos que o desempenho de cada equipe varia ao longo das rodadas.

Além disso, podemos estender esses resultados a qualquer campeonato e a partir de algumas informações cadastradas, prever resultados de jogos, pontuações e outros resultados de interesses.

Agradecimentos

- LEs – Laboratório de Estatística/UFRJ, pelo uso de seus computadores;

- CNPQ – que através da bolsa oferecida, incentivou, desde o começo, a realização do projeto;

8. Bibliografia

DeGroot, M.H. (2002) Probability and Statistics – 3a edição, Addison-Wesley.

Gamerman, D. (1997) Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Chapman & Hall, Londres.

Glickman, M. E. (1993) Paired Comparison Models With Time-Varying Parameters. Departament of Statistics, Havard University, Cambridge.

Glickman, M. E. e Stern H. S. (1998). A state-space model for National Football League scores.

Harrison, P. J. e Stevens, C. F. (1996) Bayesian forecasting (com discussão). Journal of the Royal Statistical Society, Series B, 38, 205 – 247

Knorr-Held, L. (2000) Journal of the Royal Statistical Society, Series D, The Statistician, 49, 200-225.

Pole, A., West, M. e Harrison, J. (1994) Applied Bayesian Forecasting and time series analysis. Springer, Nova York.