UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Comparação entre os Campeonatos de futebol Brasileiro e Italiano utilizando Cadeias de Markov

Autor: Eduardo Seligman
Orientador: Professor Dr. Hudson da Silva Torrent
Co-Orientador: Drª. Suzi Alves Camey

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Comparação entre os Campeonatos de futebol Brasileiro e Italiano utilizando Cadeias de Markov

Autor: Eduardo Seligman

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professor Dr. Hudson da Silva Torrent
Professor Dr. Cleber Bisognin

Porto Alegre, 21 de Novembro de 2011.

Agradecimentos

Agradeço aos meus pais Luiz Carlos e Tânia, minha irmã Fernanda e aos meus amigos pelo apoio, auxílio, compreensão e paciência prestados em todos os momentos. Também agradeço à professora Suzi que desenvolveu e acompanhou todo projeto e ao professor Hudson que me guiou durante esse processo.

Resumo

Os apreciadores de futebol e a imprensa especializada costumam discutir as diferenças entre o futebol brasileiro e europeu. O campeonato brasileiro é conhecido por sua imprevisibilidade, já os campeonatos europeus costumam ser dominados anualmente pelas equipes de maior poder financeiro. O presente estudo utilizou dados dos campeonatos brasileiro e italiano para investigar essas diferenças, utilizando, além do torneio, a qualidade das equipes e o local onde o jogo foi realizado. O estudo parte para comparações entre essas competições, através de uma modelagem utilizando cadeias de Markov, e mostra que o fator local da partida influencia mais no torneio brasileiro, enquanto que no campeonato italiano a qualidade das equipes merece maior destaque. Posteriormente, o estudo propõe modelos alternativos, testes de aderência e comparação entre esses modelos.

PALAVRAS CHAVE: Cadeias de Markov; campeonato brasileiro; campeonato italiano; futebol.

Abstract

Football fans and specialized press often discuss the differences between Brazilian and European football. The Brazilian league is known for its unpredictability, the Europeans tend to be dominated annually by better-funded teams. This study used data from the Brazilian and Italian championships to pursue these differences, using not only the tournament where the game was played, but also the quality of the teams and the local factor. Using Markov chain model, the study compares these competitions and shows that the local factor of the match influences more in the Brazilian tournament, while in the Italian championship the teams quality deserves more attention. Later, the study proposes models, tests their grips and compares these models to highlight where they most diverge.

KEYWORDS: Markov Chains; Brazilian league; league Italian; football.

Sumário

Lista de Figuras

Lista de Tabelas

1) Introdução

O futebol é o esporte mais difundido e praticado no mundo. Muitos estudos são conduzidos tentando entender melhor essa prática que tanto mobiliza diversas pessoas em diferentes continentes. Por ser um esporte famoso em todo o planeta, diferenças culturais e econômicas aparecem dependendo do local onde ele é praticado. Na Europa, grande parte dos campeonatos nacionais são amplamente e repetidamente dominados pelas equipes cujo poderio econômico é maior. No Brasil, entretanto, o campeonato nacional exibe um comportamento mais competitivo, em que a cada ano diferentes equipes acabam no topo da tabela de classificação.

Historicamente o campeonato italiano é dominado por três equipes de maior prestígio: Milan, Juventus e Internazionale. Juntas elas venceram 63 dos 107 (58,9%) campeonatos. Em contrapartida, juntando os três maiores vencedores do campeonato brasileiro (São Paulo, Flamengo e Corinthians) desde 1971, o resultado é de 37,5%. Em 2011, a CBF resolveu que a Taça Brasil, jogada nos anos 50 e 60, também seria considerada título brasileiro, aumentando o valor desse cenário para 40,7%, e mudando as três equipes mais vencedoras para Santos Palmeiras e São Paulo.

Ao longo dos anos, o futebol tornou-se uma grande fonte de estudos, isso porque cada vez mais dados estão disponíveis para análises. Oberstone (2009) utilizou regressões múltiplas para identificar as razões do sucesso das principais equipes inglesas, separando importantes dados das partidas de futebol em variáveis para essa análise. Em 2011, Oberstone expandiu sua pesquisa e analisou também o campeonato italiano e espanhol, visando encontrar as semelhanças das equipes vencedoras dentro de campo. Um estudo sobre cartões amarelos e vermelhos e sua importância foi conduzido por Anders e Rotthoff (2011), que utilizou dados do campeonato alemão para estimar seu efeito sobre a probabilidade de vitória das equipes em uma partida.

Artigos utilizando cadeias de Markov para análise esportiva tornam-se cada vez mais comuns, porém essa modelagem ainda não costuma ser muito utilizada no futebol. Newton e Asla (2009) criaram um modelo estocástico markoviano utilizando os dados do circuito profissional de tênis em 2007. Com esses dados calcularam as chances dos atletas vencerem

os pontos dependo se estavam sacando ou não, e assim estimar as chances de um jogador de tênis vencer uma partida. Krautmann et al (2010) também utilizaram cadeias de Markov na modelagem esportiva calculando o tempo de atividade de jogadores de baseball na liga profissional norte-americana.

O estudo proposto por Rump (2006) cria uma Cadeia de Markov para avaliar séries melhores de sete de *playoff* de basquete norte-americano. Nesses casos, o time de melhor campanha naquela temporada tem o direito de disputar quatro, de um total de sete partidas, em sua casa, enquanto o adversário jogará à frente de sua torcida os outros três jogos. O time favorecido sempre joga o primeiro e o sétimo jogo em seu território, entretanto, as demais cinco partidas já foram arranjadas em diferentes ordens. A ordem em que devem ser disputadas essas cinco partidas gera enorme discussão. O artigo citado propõe que diferentes formatos de organização da série podem ter um resultado comercial mais interessante para a liga que organiza o torneio (NBA), pois, dependendo da ordem, a série tem maior chance de alcançar, em média, um maior número de partidas e, consequentemente, uma maior renda para as equipes e para a liga.

Com base na ideia do estudo de Rump (2006), o banco de dados do presente trabalho foi composto pelos campeonatos brasileiro e italiano. Onde o objetivo seria comparar e destacar as principais diferenças de comportamento do campeonato brasileiro e italiano.

Utiliza-se variáveis de qualidade, a nacionalidade do campeonato e o local onde a partida foi realizada, para criar assim diferentes matrizes de transição. Estas foram utilizadas para estimar e testar as diferenças entre torneios e também dentro deles. Assim, foi criado um modelo onde o saldo de gols - número de gols do mandante menos o número de gols do visitante - seria avaliado a cada cinco minutos dentro de cada partida. Esse saldo será modelado, ao longo de cada partida, utilizando cadeias de Markov. Depois dessa primeira avaliação, diferentes modelos serão criados. Posteriormente, esses modelos terão sua aderência testada, utilizando um banco de dados composto por dados de cinco diferentes campeonatos (brasileiro, italiano, alemão, inglês e espanhol). Por fim, os tempos de absorção e recorrência dos modelos serão observados e comparados com o intuito de reforçar as diferenças dos modelos.

A próxima etapa do estudo, segundo capítulo, explica o funcionamento das cadeias de Markov, suas propriedades, estimadores, e o teste de hipóteses a ser utilizado. O terceiro capítulo detalha o banco de dados, apresenta o modelo proposto, e, a partir desse, os diferentes modelos que serão selecionados. Esses modelos serão testados e comparados com intuito comprovar suas diferenças. Finalmente será apresentada a conclusão do estudo.

2) Cadeias de Markov a Tempo Discreto

Este capítulo fundamenta a parte teorica do estudo, mostrando a técnica estatística utilizada. Primeiramente mostra-se as definições básicas das cadeias de Markov discretas e finitas. Em um segundo momento inferências sobre as cadeias de Markov serão apresentadas, como o estimador de máxima verossimilhança para as probabilidades de transição e a eliminação dos parâmetros. Por fim o teste de hipoteses para certificar que diferentes amostras pertencem a uma mesma cadeia de Markov é explicado.

2.1) Definições e Propriedades

No ano de 1907, Andrei Andreyevich Markov iniciou um estudo sobre um importante e inovador tipo de processo, onde o resultado de um experimento pode afetar o resultado do experimento seguinte. Esse processo recebeu seu nome: cadeia de Markov, aqui denotadas por CM. O fato de um processo satisfazer a propriedade de Markov significa que dado o presente, o resto do passado é irrelevante para a previsão do futuro.

O conjunto de possíveis resultados de uma cadeia de Markov chama-se espaço de estados, que definiremos aqui por $S = \{s_1,s_2,...,s_r\}$. Neste caso temos uma CM com r finito de estados. O processo começa em um estados e se move entre eles. Cada um desses movimentos é chamado de passo. Se, em um determinado momento, o processo está no estado s_i, e move-se para o estado s_j no passo seguinte, é chamada de probabilidade de transição do estado s_i para o estado s_j e é denotada por p_{ij}, e esta probabilidade não depende dos estados anteriores, e sim do estado atual.

Essas probabilidades p_{ij} são chamadas de probabilidades de transição. Outra definição importante é a de recorrência do estado, ou seja, o processo se encontrar no estado s_i e continuar nesse mesmo estado após um passo. Nesse caso a probabilidade é de p_{ii}.

Supondo que um processo tenha quatro diferentes estados, automaticamente ela terá dezesseis diferentes probabilidades de transição, doze delas de mudar de estado e mais quatro

de recorrência de cada um dos estados. Essas probabilidades formam a matriz de transição denotada por *P*. A equação (1) a seguir exemplifica essa matriz.

$$P = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \end{array} \begin{array}{c} \begin{array}{cccc} s_1 & s_2 & s_3 & s_4 \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{pmatrix} \end{array}. \quad (1)$$

Para uma CM com *r* estados e uma transição do estado *i* para o estado *j* em dois passos, ficaríamos com a equação (2):

$$p_{ij}^{(2)} = \sum_{k=1}^{r} p_{ik}\, p_{kj}\ . \quad (2)$$

No caso de um processo sair do estado s_1 e ir para o estado s_4 em dois passos, existem quatro possibilidades de o processo percorrer esse caminho. O resultado é a soma das quatro probabilidades condicionais é apresentado na equação (3).

$$p_{14}^{(2)} = p_{11}p_{14} + p_{12}p_{24} + p_{13}p_{34} + p_{14}p_{44}\,. \quad (3)$$

O estudo de cadeias de Markov se interessa pelo estado em que o processo se encontrará após um grande número de passos.

No estudo serão utilizadas cadeias de Markov absorventes, isto é, onde pelo menos um dos estados s_i, para $i = 1,2,...,r$, tem probabilidade de recorrência igual a um ($p_{ii} = 1$), ou seja, é impossível sair de tal estado, consequentemente a probabilidade de transição desse estado para qualquer outro é igual a zero. Além disso, também é necessário poder ir de qualquer estado para algum estado absorvente, não necessariamente em um passo. Os estados que não são absorventes nessas cadeias são conhecidos como estados transientes.

A matriz canônica é a forma de separar a matriz de transição. Esta representação tem como objetivo estudar os tempos de absorção e recorrência do processo. A equação (4) a seguir apresenta a representação canônica da matriz de transição. Nesta representação, apenas a ordem das colunas e das linhas da matriz foi alterada, separando os estados transientes, inserindo estes nas de cima e a esquerda, e absorventes, a baixo e a direita.

$$P = \left(\begin{array}{c|c} Q & R \\ \hline 0 & I \end{array}\right). \tag{4}$$

O quadrante denotado por *Q* será composto pelas probabilidades de transição do processo entre os estados transientes. No segundo quadrante, a sua direita, denotado por *R*, ficam as probabilidades de o processo ser absorvido. A baixo deste, denotado por *I*, as probabilidades de transição dos estados absorventes, onde sempre terá apenas o valor de sua recorrência igual a um na diagonal principal e o restante dos valores iguais a zero, formando uma matriz identidade. Por último, denotado por *0*, a sua esquerda sempre será formada apenas com zeros, pois a probabilidade do processo sair de um estado absorvente é nula.

Utilizando a matriz *Q* serão calculados os tempos médios de absorção do processo, $N = (I - Q)^{-1}$, onde *I* é uma matriz identidade de mesmas dimensões a matriz *N*. Assim, cada linha de *N* representa o estado transiente que começou o processo e mostra o tempo médio que o processo fica em cada um dos estados transientes (colunas).

O próximo passo é calcular o tempo de absorção, ou seja, o tempo que o processo demora a ser absorvido, considerando o estado onde começou. Para isso, foi multiplicado um vetor coluna *c*, onde todos os valores são iguais a um, a matriz *N*. O resultado será uma coluna com os tempos médios totais de absorção do processo para qualquer estado transiente que ele comece.

Finalmente, será utilizada a matriz *R* para calcular a probabilidade de absorção do processo para cada um dos estados absorventes, levando em consideração o estado transiente onde o processo iniciou. O resultado é a matriz $B = NR$, que terá nas linhas os estados

transientes e nas colunas os estados absorventes, assim dado que o processo começou em algum dos estados transientes influenciará na probabilidade de ele acabar em cada um dos estados absorventes.

2.2) **Inferência em Cadeias de Markov**

Nesta seção será apresentado o estimador de máxima verossimilhança para as probabilidades de transição de cadeias de Markov e sequencialmente a edliminação dos parâmetros. Por ultimo será apresentado o teste de hipóteses de Anderson e Goodman (1955).

2.2.1) **Estimador de Máxima Verossimilhança para as probabilidades de Transição de cadeias de Markov**

Considerando o caso de uma cadeia de Markov X_1^{∞},com *r* estados. A matriz de transição *P* é desconhecida sem restrições impostas, porém pretende-se estimá-la através dos dados. Os parâmetros a serem inferidos são as entradas p_{ij}, que para serem estimados a CM precisa ser estacionária, da matriz *P* de dimensão r^2, definidas na equação (5) a seguir:

$$p_{ij} = \Pr(X_{t+1} = j \mid X_t = i)\,. \tag{5}$$

Baseados em uma realização (ou caminho) $x_1^n \equiv (x_1, x_2, ..., x_n)$ da CM, isto é, de uma realização da variável aleatória $X_1^n \equiv (X_1, X_2, ..., X_n)$, a probabilidade dessa realização é dada por:

$$\Pr(X_1^n = x_1^n) = \Pr(X_1 = x_1)\prod_{t=2}^{n} \Pr(X_t = x_t \mid X_1^{t-1} = x_1^{t-1})\,. \tag{6}$$

$$\Pr(X_1 = x_1)\prod_{t=2}^{n} \Pr(X_t = x_t \mid X_{t-1} = x_{t-1})\,. \tag{7}$$

Na equação (6) usa-se a definição de probabilidade condicional, e na equação (7) usa-se, a propriedade markoviana, onde o futuro é independente do passado, dado o presente.

Reescrevendo a equação (7) em termos de probabilidade de transição, p_{ij}, equação (8), para conseguir a verossimilhança de certa matriz de transição. Podemos rescrever a equação

(7) em termos das probabilidades de transição de um passo. Assim a função das entradas da matriz de transição *P* as quais queremos estimar:

$$L(p) = \Pr(X_1 = x_1) \prod_{t=2}^{n} p_{x_{t-1} x_t} \ . \tag{8}$$

Definindo $N_{ij} \equiv$ numero de vezes *i* é seguido pelo j em X_1^n e podemos reescrever a equação (8) da seguinte forma:

$$L(p) = \Pr(X_1 = x_1) \prod_{i=1}^{k} \prod_{j=1}^{k} p_{ij}^{\ n_{ij}} \ . \tag{9}$$

Onde *P* = (..) é um vetor de probabilidades de transição.

Maximizar a função de verossimilhança é p mesmo que maximizar o logaritimo da função de verossimilhança, Ou seja ,

$$\ell(p) = \log L(p) = \log \Pr(X_1 = x_1) + \log \sum_{i,j} n_{ij} \log p_{ij} \ . \tag{10}$$

Derivando a equação (10) em relação à p_{ij}, para $i,j \in \{0,1,...,m\}$ temos,

$$\frac{\partial \ell}{\partial p_{ij}} = \frac{n_{ij}}{p_{ij}} \ . \tag{11}$$

Igualando a equação (11) a zero temos,

$$\frac{n_{ij}}{p_{ij}} = 0 \ . \tag{12}$$

Conclui-se assim que as probabilidades de transição estimadas deveriam ser iguais a infinto.

O procedimento falhou porque não foi possível comprovar a aderência dos parâmetros. Eles não podem trocar-se arbitrariamente, pois a probabilidade de fazer a transição de um estado tem que somar 1, ou seja, para cada i $\in \{0,1,...,m\}$, temos que:

$$\sum_j p_{ij} = 1. \tag{13}$$

Isso significa que o número de graus de liberdade para uma matriz de transição não é m^2 e sim $m(m - 1)$.

2.2.2) **Eliminando parâmetros**

Foi escolhida arbitrariamente uma probabilidade de transição para expressar em termos das outras. Diga-se a probabilidade de ir para 1, partindo de algum $i \in \{0,1,...,m\}$ para cada i, $p_{i1} = 1 - \sum_{j=2}^{m} p_{ij}$. Agora quando obtidas as derivadas da verossimilhança (14), tira-se $\partial / \partial p_{i1}$, e os outros termos mudam:

$$\frac{\partial \ell(p)}{\partial p_{ij}} = \frac{n_{ij}}{p_{ij}} - \frac{n_{i1}}{p_{i1}}. \tag{14}$$

Igualando a equação (14) acima a zero temos,

$$\frac{n_{ij}}{\hat{p}_{ij}} = \frac{n_{i1}}{\hat{p}_{i1}}. \tag{15}$$

Ou seja,

$$\frac{n_{ij}}{n_{i1}} = \frac{\hat{p}_{ij}}{\hat{p}_{i1}}. \tag{16}$$

Já que isso serve para todos $j \neq 1$, conclui-se que $\hat{p}_{ij} \propto n_{ij}$, e na verdade:

$$\hat{p}_{ij} = \frac{n_{ij}}{\sum_{j=1}^{m} n_{ij}}. \qquad (17)$$

Claramente, a escolha de p_{i1} como a probabilidade de transição para eliminar foi arbitrária e de qualquer maneira o mesmo resultado seria obtido.

Pode-se mostrar que o estimador de Máxima Verossimilhança definido acima é consistente e assintoticamente normal sob hipóteses bastante razoáveis. Para detalhes ver Vargas (2011)

2.3) **Teste da hipótese que diferentes amostras são da mesma cadeia de Markov**

Anderson e Goodman (1955) propuseram diversos testes de hipóteses para cadeias de Markov. Neste estudo, utilizamos o teste de hipótese proposto para certificar se duas amostras são de uma mesma cadeia de Markov de uma certa ordem. O teste propõe utilizar *s* amostras $(s \geq 2)$ e *s* é finito. Como serão feitas apenas comparações duas a duas, foi utilizado somente o teste com duas amostras *(s = 2).*

Sendo $\hat{p}_{ij\ldots kl}^{(h)} = n_{ij\ldots kl}^{(h)} / n_{ij\ldots k}^{*(h)}$ a probabilidade estimada para as transições do CM, para cada amostra *h (h = 1,2,...,s).* Foi obtida a equação (18):

$$\chi_{ij\ldots k}^{2} = \sum_{l} C_{ij\ldots k} \left(\hat{p}_{ij\ldots kl}^{(1)} - \hat{p}_{ij\ldots kl}^{(2)} \right)^2 / \hat{p}_{ij\ldots kl}^{(\bullet)}. \qquad (18)$$

Onde $\hat{p}_{ij\ldots kl}^{(\bullet)}$ é a probabilidade estimada de transição onde as duas amostras são somadas e $C_{ij\ldots k}^{-1} = \left(1/n_{ij\ldots k}^{*(1)}\right) + \left(1/n_{ij\ldots k}^{*(2)}\right)$. Finalmente, a distribuição de $\sum_{i,j,\ldots,k} \chi_{ij\ldots k}^{2}$ é uma qui-quadrado com $m^r(m-1)$ graus de liberdade, provada por Anderson e Goodman (1955)

O teste propõe que o número de graus de liberdade seja $m^r(m-1)$, pois sua idéia é testar todos os valores na matriz de transição. Porém, como a matriz proposta nesse estudo utiliza apenas transições de, no máximo dois estados, muitas de suas transições são iguais a zero, assim uma redução dos graus de liberdade foi pensada, e através do artigo proposto por

Billingsley P. (1961) foi utilizado o número de graus de liberdade igual $d - m$, onde $d = \sum_i d_i$ é o número de entradas positivas na matriz de transição. Como são dez os estados no presente estudo, o total então é de cem entradas na matriz, e retirando as entradas que não serão utilizadas no teste, entradas nulas onde as transições não são possíveis no modelo proposto, o total fica igual a 39, assim o número de graus de liberdade utilizado para o teste será igual a 29.

3) Aplicação

Dados de dois diferentes campeonatos (brasileiro e italiano) foram coletados e o objetivo será procurar diferenças entre suas matrizes de transição. Cada partida será um evento independente e seu comportamento será estudado de forma isolada. A variável em estudo é o saldo de gols das equipes dentro de cada uma das partidas. Um segundo banco de dados criado, para os teste de aderência, também é mostrado, assim como a variável de qualidade que utilizou-se para as comparações entre as matrizes de transição.

Num primeiro momento serão apresentados os dados, o que foi coletado para criar o banco utilizado e como ele foi estruturado. Na sequência, será apresentado o modelo proposto, onde utilizando o conteúdo das cadeias de Markov mostrado anteriormente, será montando um modelo para estudar as partidas de futebol e seu comportamento. Em seguida será aplicado o teste proposto por Anderson e Goodman (1955) que compara matrizes de transição, e verifica se as duas amostras (matrizes) testadas são originárias da mesma cadeia. Essa comparação tentará provar diferenças ao serem separadas pelos países e pela qualidade das equipes. Os resultados desses testes serão utilizados para criação de modelos de estimação. Posteriormente, serão comparados aos dados reais, com objetivo de testar se eles realmente estimam bem os resultados das partidas. Modelos de estimação e previsão serão propostos e avaliados. Assim como os tempos de absorção e recorrência dos processos estudados.

3.1) **Dados**

Na base de dados composta pelos campeonatos nacionais brasileiro e italiano, apenas a primeira divisão de cada um deles foi utilizada. Os dados coletados apresentam os minutos exatos e os atletas que marcaram os gols em cada uma das partidas selecionadas. Além disso, para cada uma das partidas da amostra, a posição de ambas as equipes ao final do campeonato do ano em que a partida ocorreu também compõe o banco de dados. Essa posição ao final do campeonato posteriormente será utilizada para criação de uma estratificação para diferenciar a qualidade das equipes.

O campeonato brasileiro foi selecionado a partir do estabelecimento do sistema de pontos corridos em 2003. Esse sistema consiste em um torneio em que todos os times se enfrentam duas vezes, uma vez como mandante de campo e outra como visitante. Ao final de todas as rodadas, o time com maior número de pontos é considerado campeão. Os pontos são obtidos através de vitórias que valem três pontos, ou empates que somam um ponto. As derrotas não contam pontos para a equipe. O número de vitórias obtidas pela equipe, o saldo de gols (diferença entre o número de gols feitos e o número de gols sofridos), os gols feitos, o confronto direto, o número de cartões vermelhos e o número de cartões amarelos são exemplo dos critérios de desempate, caso os times obtenham a mesma pontuação no campeonato. Os critérios de desempate são aplicados nos campeonatos para classificar as equipes.

3.1.1) Campeonato Brasileiro

O sistema de pontos corridos foi estabelecido no Brasil em 2003. Portanto os dados para o presente estudo, foram a totalidade das partidas realizadas desde o campeonato de 2003 até o campeonato de 2010. Assim, de um total de 3477 partidas, 3463 foram utilizadas nas análises. Essas 14 exclusões serão explicadas a seguir.

No ano de 2003, o campeonato brasileiro era formado por 24 times, esse número foi reduzido no campeonato de 2005, onde participaram 22 equipes e no ano seguinte esse número foi novamente reduzido, restando apenas 20 times a partir de 2006. A maioria dos campeonatos nacionais no mundo utiliza as mesmas 20 equipes.

O ano de 2005 foi um ano conturbado na história do campeonato brasileiro, nesse ano houve manipulações de resultados, e as onze partidas apitadas pelo árbitro Edilson Pereira da Carvalho foram anuladas no dia 2 de Outubro de 2005 pelo Superior Tribunal de Justiça Desportiva, decidindo assim que elas fossem jogadas novamente. Para o trabalho foram consideradas apenas as partidas repetidas. Isto é, excluindo os jogos originais. Como a influência dos resultados foi dentro do campo, tendo como consequência a alteração do resultado final, a decisão foi utilizar os jogos em que, aparentemente, não houve influência da arbitragem.

Nos casos em que a justiça tomou decisões somente baseada em problemas jurídicos, foi considerado o resultado de campo da partida. Em 2005, o Brasiliense abriu os portões amparados pela justiça comum, e foi punido pela Superior Tribunal de Justiça Desportiva com a perda dos pontos da partida. O jogo originalmente havia sido 2x2, porém a justiça atribui o resultado de 1x0 para o Vasco da Gama. Nesse caso, o jogo foi considerado 2x2 e a tabela final do campeonato foi modificada. Essa decisão foi tomada porque nesses casos o resultado da partida, em campo, parece pertinente e teoricamente ele demonstra a qualidade real do time.

Na temporada de 2010, o Grêmio Prudente perdeu três pontos por escalar um jogador sem condições legais. Já em 2003, a Ponte Preta e o Paysandu perderam pontos respectivamente de duas e quatro diferentes partidas, em ambos os casos devido a atletas inscritos irregularmente. Sendo assim, seus adversários nas partidas receberam os três pontos. Em 2004, a equipe do São Caetano teve 24 pontos retirados por escalar um jogador sem condições médicas, e que veio a falecer em campo. Nesse caso, pode-se ver que o São Caetano, um time de boa qualidade no campeonato, teria sido considerado um time ruim. Ao invés de estar no topo da tabela, devido aos seus bons resultados, ficaria na parte de baixo dela devido a uma decisão da justiça. A idéia do trabalho não é discutir as questões de justiça, mas sim, utilizar a qualidade do time para análises, e no caso essa decisão da justiça afetaria na decisão da qualidade da equipe paulista. Por esta razão, em todos os casos acima citados foram mantidos os resultados de campo e as tabelas finais dos campeonatos foram adaptadas.

Foram excluídos três jogos da análise, sendo eles: Coritiba 1x4 São Paulo na vigésima quinta rodada de 2005, Portuguesa 5x5 Figueirense na primeira rodada de 2008 e Sport Recife 4x2 Flamengo na quinta rodada de 2009. Os três casos mostraram ser discrepantes, pois do total da amostra coletada (6893), apenas nesses três jogos foi observado uma equipe aumentar a vantagem em três gols num mesmo intervalo de cinco minutos. Ou seja, situação representa menos de 0,05% dos resultados.A razão da exclusão dessas três situações será apresentada na Seção 3.2, na apresentação do modelo.

3.1.2) Campeonato Italiano

Na Itália o sistema de pontos corridos é muito mais antigo, foi estabelecido em 1929. Neste estudo, os campeonatos selecionados começam na temporada 1999/00 e seguem até 2010/11. Resultando assim num total de 3430 partidas. Na primeira temporada estudada a liga italiana era formada por 18 times. A temporada de 2004/05 marca a mudança para 20 equipes.

Assim, como no Brasil, o campeonato italiano foi marcado por escândalos, e duas de suas temporadas, a de 2004/05 e a de 2005/06, foram desconsideradas neste trabalho. Em ambos os campeonatos foram descobertas corrupções, onde os times manipulavam os resultados. A Juventus que havia sido considerada campeã italiana nas duas temporadas teve seus dois títulos retirados. Essa relação clara entre o resultado final do campeonato e a influência da arbitragem nas partidas com certeza exibiria uma tendência nas análises, já que boa parte das partidas teve seu resultado favorecendo certas equipes.

Para o Campeonato Italiano o mesmo critério foi considerado nos casos de decisão da justiça. Na temporada seguinte aos escândalos (2006/07) SS Lazio, AC Milan, AC Fiorentina e Reggina Calcio tiveram suas pontuações reduzidas, começaram os campeonatos com pontuações negativas como punição por terem participado dos escândalos. A Juventus foi punida com o rebaixamento e perda de dezoito pontos, assim não podendo participar da Série A nessa temporada. Essas pontuações negativas foram desconsideradas pelo mesmo critério apresentado no caso do campeonato brasileiro.

Nessa mesma temporada, o Siena teve um ponto retirado devido a problemas com pagamento de previdência social. Na temporada 2010/11 o Bologna teve sua pontuação final reduzida em três pontos devido a problemas com impostos. Novamente as tabelas foram adaptadas.

Finalmente na décima segunda rodada do campeonato italiano de 2002/03 o resultado da partida Calcio Como contra Udinese foi alterado. O jogo originalmente teve o resultado de 0x1, porém foi interrompido por problemas com a torcida aos 81 minutos. A justiça decidiu considerar o jogo 0x2 em favor da Udinese. Para as análises foi utilizado o resultado original. O resultado final da amostra selecionada está demonstrado na Figura 1.

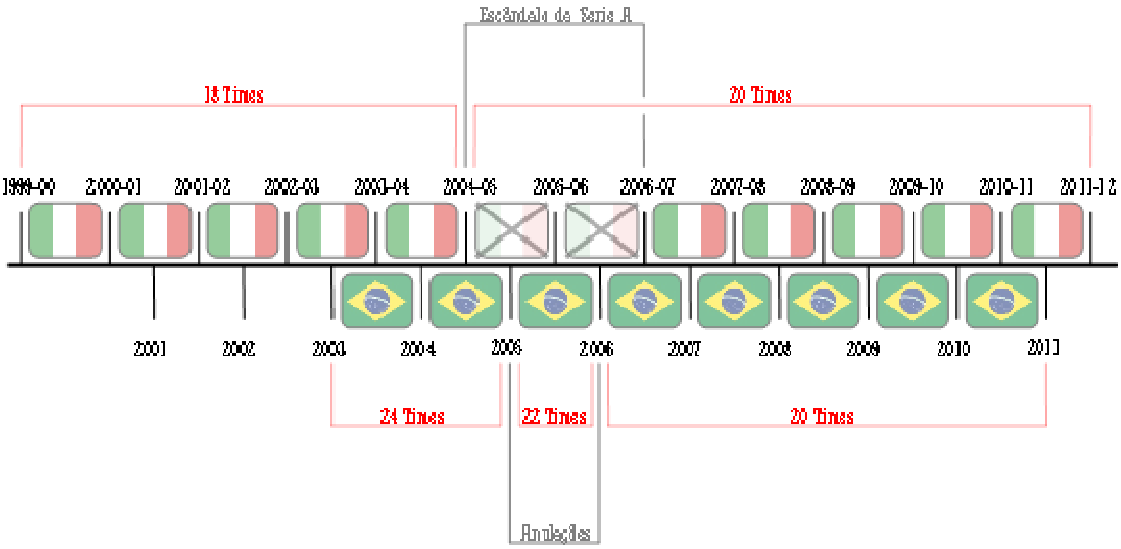

Figura 1: Campeonatos utilizados no Modelo

3.1.3) Dados selecionados para o Teste de Aderência

O campeonato brasileiro teve o sistema de pontos corridos estabelecido no ano de 2003, assim os dados anteriores a esta temporada não parecem pertinentes ao estudo. Como o comportamento da competição parece muito distinto dos campeonatos mais famosos do planeta, devido ao equilíbrio entre as equipes que participam, seria complicado testar o modelo do Brasil em ligas de outros países. Como a base de dados para a criação do modelo engloba as temporadas de 2003 a 2010, tem-se disponível apenas a temporada do ano de 2011 para a realização desse teste.

Para testar os modelos da Itália, foram selecionadas as cinco temporadas anteriores da liga italiana, isso porque antes desses cinco torneios a vitória valia dois pontos, fato que mudou na temporada 1994-95, passando assim a valer três pontos. Essa situação distorce os dados porque infla o número de empates. Um empate vale um ponto e muitas vezes o comportamento das equipes se mostrava diferente durante as partidas, que se satisfaziam com um empate quando jogavam fora de casa.

Além dessas cinco temporadas do campeonato italiano, também foram selecionadas dez temporadas dos campeonatos inglês, espanhol e alemão. Isso porque essas competições

tendem a demonstrar um comportamento parecido com o italiano, onde a diferença de qualidade dos times é muito grande e o torneio é amplamente dominado por poucas equipes de maior potencial financeiro. A Figura 2, ilustra a amostra selecionada para o teste de aderência.

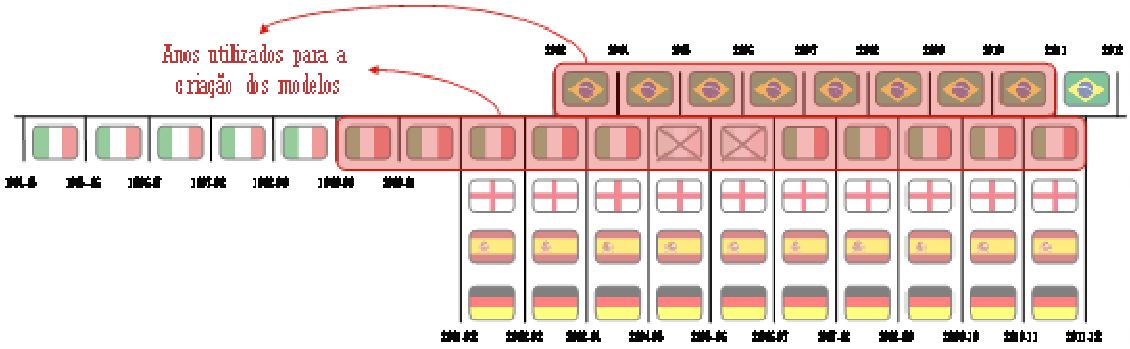

Figura 2: Campeonatos utilizados para o teste de aderência dos modelos

3.1.4) Variável de Qualidade das Equipes

Uma variável para medir a qualidade de cada uma das equipes também foi adicionada ao banco. A cada partida realizada foi pesquisada a posição das duas equipes ao final daquela específica temporada. Assim, tanto o time mandante como o time visitante tem uma medida que avalia sua qualidade no decorrer daquela temporada. Um ano independe do outro, um time que obteve uma boa colocação num ano, não tem nenhuma relação com seu desempenho no ano anterior.

Posteriormente essa variável foi estratificada, o campeonato foi divido em três subgrupos, os times de melhor qualidade sendo o grupo dos times bons (B). Um segundo grupo era composto pelos times que terminaram o campeonato na metade da tabela: os times médios (M). O último grupo era o dos times ruins (R), formado pelas equipes mais fracas. Como nem todos os campeonatos estudados podem ser divididos em três grupos de mesmo tamanho, boa parte deles têm vinte participantes, então foi dada preferência por colocar mais times no grupo de força inferior. Por exemplo, o grupo dos times bons teria seis times, e os

outros dois grupos ficariam com sete equipes. Assim, deixando sempre o melhor grupo com um número menor, ou igual aos outros dois grupos.

3.2) **Modelo Proposto**

Uma partida de futebol tem noventa minutos de jogo e mais os acréscimos do primeiro e segundo tempo. O objetivo do trabalho era analisar os gols da partida, mais especificamente a diferença do placar. A fim de se transformar tempo em uma variável fixa, cada partida foi dividida em intervalos de cinco minutos. Além disso, considera-se os minutos finais de desconto dos primeiro e segundo tempos como mais dois intervalos, totalizando assim, dez intervalos em cada tempo, ou seja, 20 intervalos por partida. Foi adicionado mais um intervalo ao final de cada jogo para que as partidas decididas nos últimos minutos também fossem absorvidas.

Uma das variáveis da base de dados é o minuto exato de cada gol ocorrido. Ou seja, para cada intervalo sabe-se o número de gols marcados tanto pelo time mandante como visitante. A partir daí, foi estudado o placar da partida em cada um desses intervalos. Para o estudo foi selecionada a variável diferença de gols. Onde cada uma das diferenças de gols seria um estado. Essa variável agrega as informações de diferentes placares, o que é conveniente, pois se fossem utilizados todos os placares obtidos, o resultado seria uma matriz de transição muito grande, com muitos estados e com pouca informação nesses estados.

Essa diferença de gols é avaliada em cada um dos intervalos, onde é subtraído o número de gols do visitante do número de gols do mandante. Quando a diferença for positiva significa que o time dono da casa está na frente. Em caso de placar negativo significa que o visitante se encontra a frente no placar.

O passo seguinte foi observar a freqüência da variável da placar. Foram avaliadas 6893 partidas, cada uma delas com 21 intervalos, 20 da partida mais o intervalo utilizado para absorção do processo, totalizando assim 144753 saldos. A Figura 3 ilustra a distribuição de frequência do saldo de gols. Nesta distribuição foi observado que a maior diferença encontrada tanto a favor como contra o time mandante foi de sete gols (série original). Porém, em uma observação mais minuciosa ficou claro que as frequências a partir de uma diferença

de três gols, tanto positiva como negativa eram baixas, e que a decisão tomada foi juntar esses valores (série utilizada). Totalizando assim sete estados transientes.

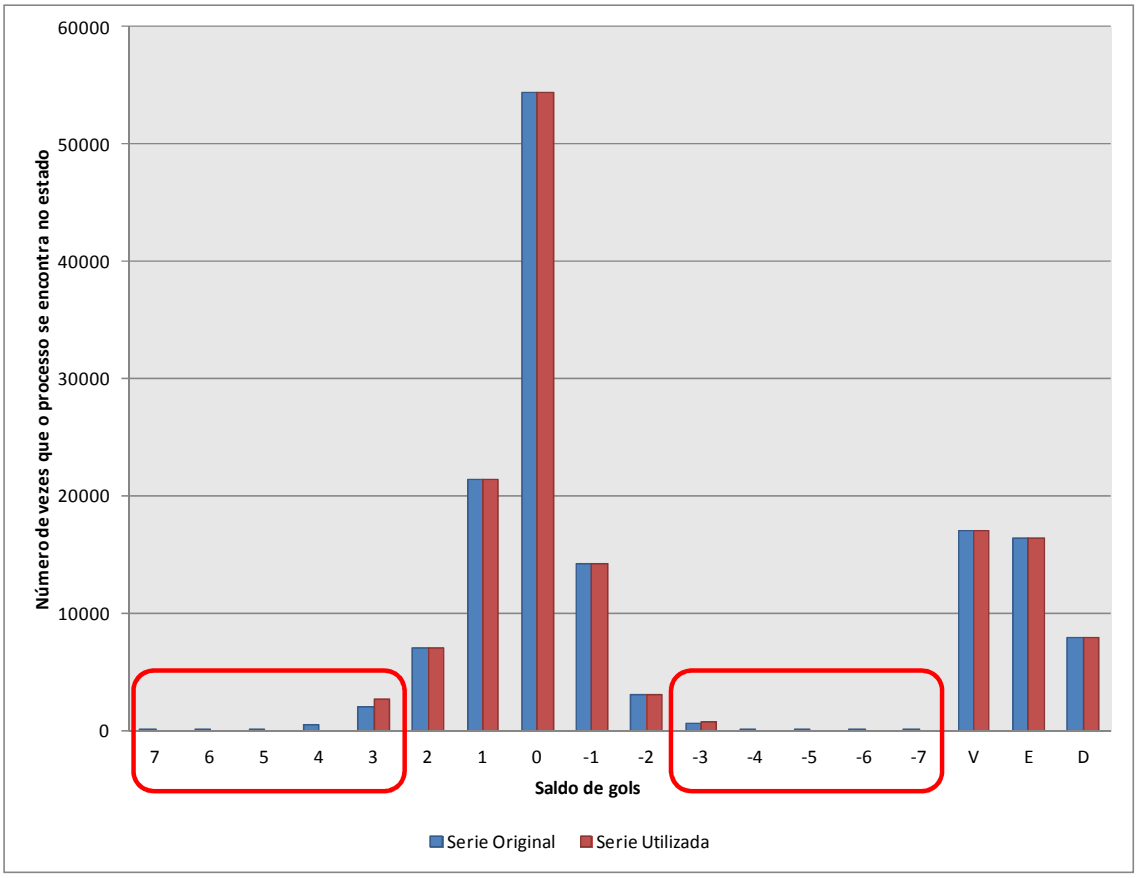

Figura 3: Seleção do número de estados

Esses dados foram passados para uma matriz de transição. Onde foi observado, que além da recorrência dos estados, quando o jogo continua com o mesmo placar no intervalo seguinte, deveriam ser estudadas as transições para mais ou menos um ou dois gols. Em todos os 6893 jogos apenas três deles apresentaram uma transição em que a vantagem foi aumentada por uma das equipes em três gols em apenas um passo. A decisão de excluir esses dados foi tomada porque, além de raros, esses casos não aparentam ter significância prática para o estudo.

Três estados absorventes foram criados com o objetivo de obter estimativas de tempos de recorrência e absorção do processo. Depois de uma das duas equipes marcar o último gol da partida, o intervalo seguinte é transformado em vitória (V), empate (E) ou derrota (D). Onde o resultado final é referente ao mandante do jogo, é considerada vitória quando o time

local vence a partida. Ou seja, assim que a partida é definida o processo é absorvido. A Figura 4 apresenta esquematicamente as possíveis transições dessa CM.

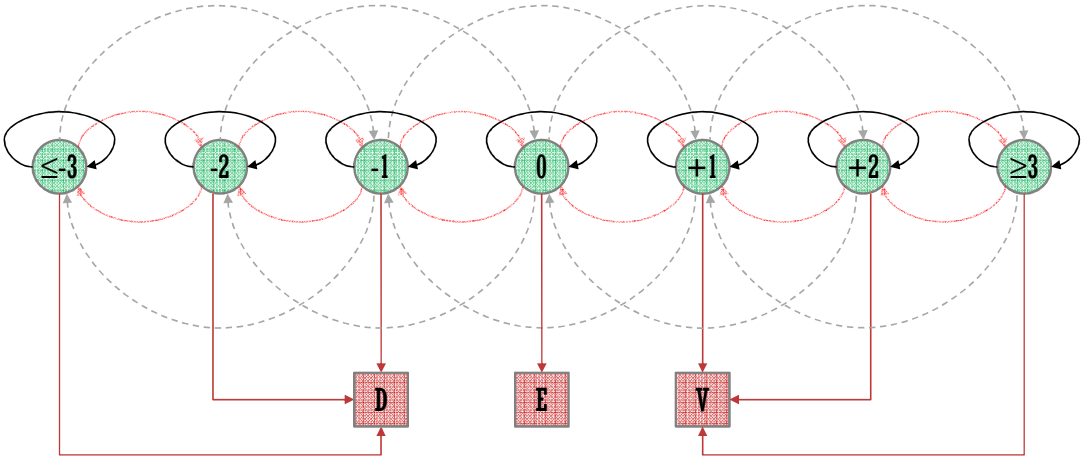

Figura 4: Modelo selecionado

3.3) **Estimação e Testes**

Com modelo acima proposto foram criadas matrizes de transição com o objetivo de comparar diferentes campeonatos e times de diferentes qualidades. Deste momento em diante diversas comparações serão mostradas com o objetivo de ilustrar as diferenças entre o campeonato brasileiro e italiano.

3.3.1) Comparação das CM: Brasil x Itália

A primeira comparação é separar os campeonatos dos dois países. Porém, quando é colocada apenas essa variável – país - e é desconsiderada a qualidade das equipes, não consegue-se observar diferenças. Nessa comparação o valor p obtido foi 0,9999, portanto não se pode afirmar que as duas amostras vieram de diferentes cadeias.

Na Figura 5, cada bolha representa uma transição no placar realizado no intervalo de 5 minutos de uma partida. Por exemplo, uma bolha na intersecção no ponto (2,0) significa que o time da casa ganhava por uma diferença de 2 gols e cinco minutos depois o jogo estava empatado. Já o tamanho das bolhas representa o número de vezes que essa transição

aconteceu. Razão pela qual se observa que a maior bolha representa o empate entre os dois times antes do final da partida, uma vez que todos os jogos começam empatados. Ainda em relação a Figura 5 se pode ver que em quase todas as transições e recorrências os campeonatos se sobrepõem, o que reforça ainda mais o resultado do teste de hipótese.

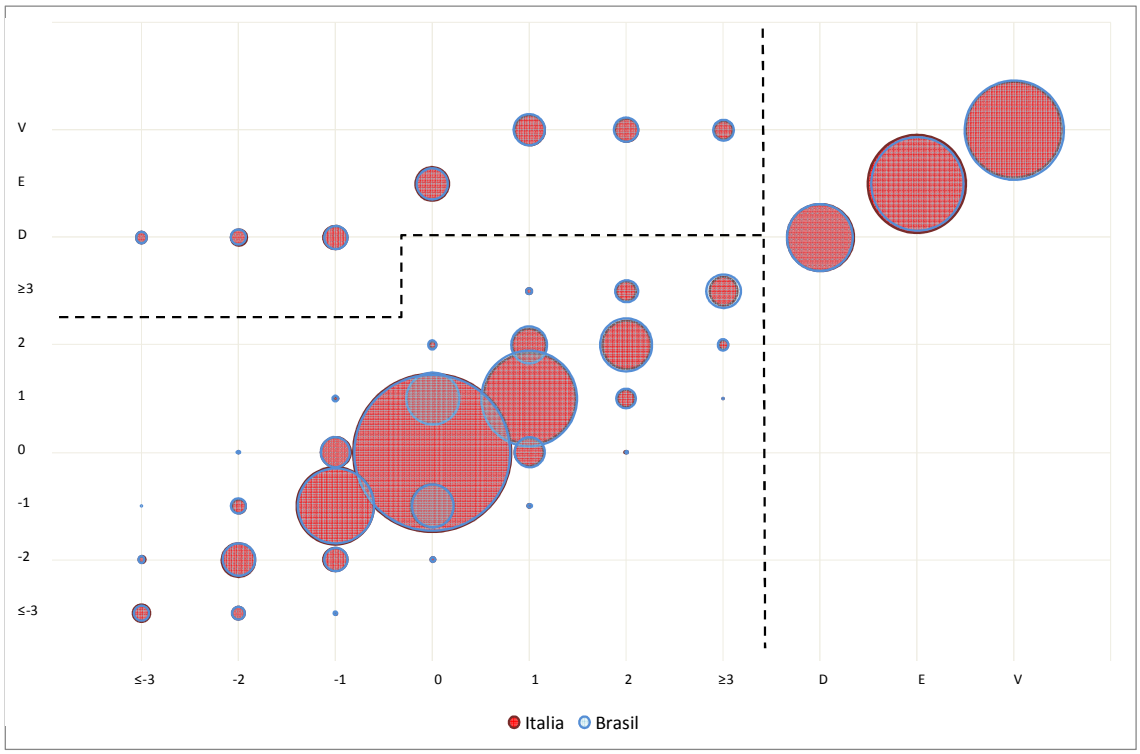

Figura 5: Comparação entre as matrizes de transição do campeonato brasileiro x campeonato italiano

3.3.2) Comparação das CM: Total_BR x Total_RB

Uma nova abordagem será tentada agora. Será utilizada a variável de qualidade, e ignorada a nacionalidade do campeonato. Para isso serão selecionadas as duas matrizes de transição que teoricamente apresentariam maiores diferenças possíveis. A primeira delas (Total_BR) sendo a matriz onde o time da casa é bom (B), ou seja, da melhor qualidade possível, enfrentando uma equipe da pior qualidade possível, uma equipe ruim (R). Esta matriz será comparada com a sua matriz oposta, onde a equipe fraca recebe uma equipe de forte (Total_RB) em seu estádio.

Desta vez, a Figura 6 demonstra uma diferença muito grande entre as duas matrizes: o tamanho das bolhas. Por isso, é evidentemente superior na parte em que os estados são

positivos (bolhas vermelhas), quando observado um time de melhor qualidade jogando em casa e vencendo. O contrário também é válido, pode-se observar que o time de pior qualidade costuma estar atrás mais vezes, mesmo estando dentro de sua casa.

Então, primeiramente, será analisada a parte inferior esquerda do gráfico. A diagonal central significa a probabilidade do placar ser mantido ao passar um intervalo de tempo. Logo, fica evidente que as bolhas vermelhas são maiores, mantendo um placar positivo, que está ilustrado na parte superior do gráfico. Já as bolhas azuis são maiores na parte inferior do gráfico, o que significa que quando um time de qualidade inferior está atrás no placar em sua casa, ele tem mais chance de continuar perdendo o jogo com essa diferença no placar do que uma equipe de nível superior jogando em casa.

As outras quatro diagonais mostram as transições, ao observar as bolhas vermelhas nota-se que elas sempre são maiores na diagonal acima da diagonal central. Isso significa que a transição é sempre a favor mandante de melhor qualidade. Em qualquer um dos placares, a chance de ele marcar é sempre superior do que a de sofrer um ou dois gols. Esse raciocínio vale de maneira oposta para o mandante de pior qualidade, independente do placar a chance de ele conceder um gol é sempre maior do que sua chance de marcar um gol.

Na diagonal da parte direita do gráfico, observa-se o mesmo raciocínio que a diagonal central do gráfico, até porque ela faz parte dessa diagonal central. Onde as bolhas representam o tempo que os processos se mantiveram nos estados absorventes. Obviamente, a matriz Total_BR tem probabilidade muito superior de manter a vitória, e muito inferior de se manter perdendo a partida ao ser comparada com a matriz Total_RB. Nessa segunda matriz, vê-se que o jogo permanece mais tempo sendo um derrota ou empate do que vitória para o mandante de qualidade inferior.

A parte superior esquerda do gráfico mostra as transições de cada um dos placares para os estados absorventes. Após o intervalo em que foi marcado o último gol da partida, o processo é absorvido no intervalo seguinte. Nota-se que a absorção das bolhas vermelhas para a vitória é sempre superior, independentemente do placar, a absorção de empates e derrotas. Já no caso da matriz Total_RB, observa-se uma superioridade na absorção do empate.

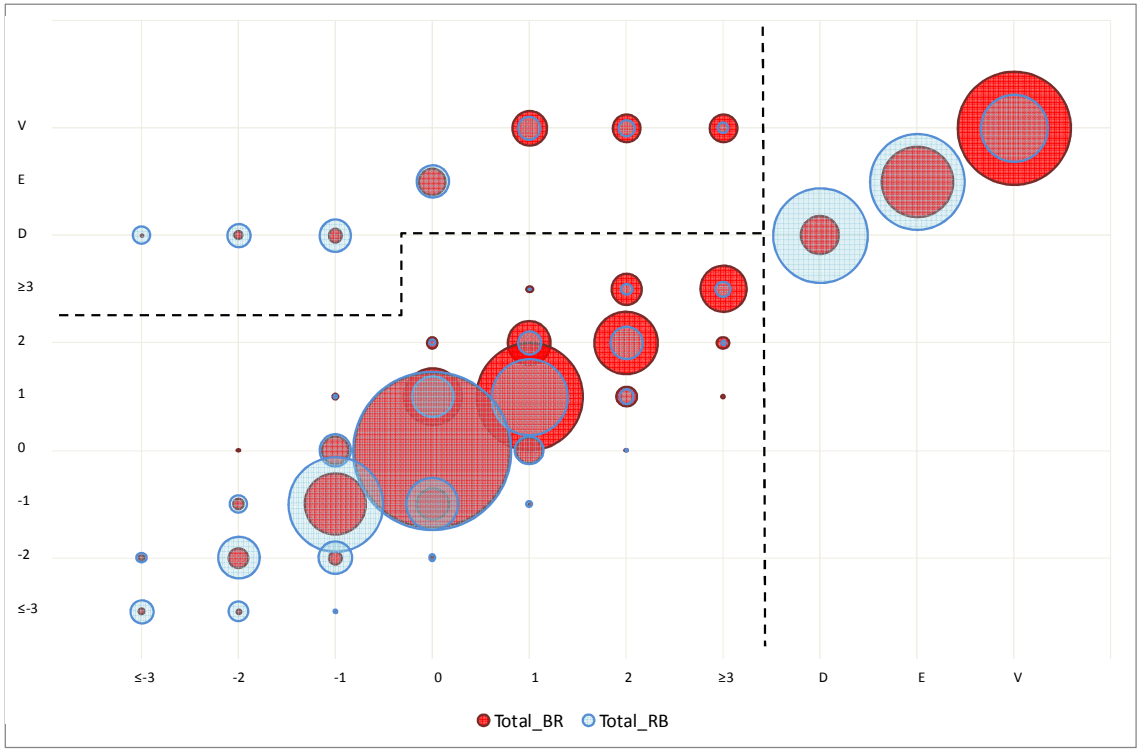

Figura 6: Comparação entre as matrizes de transição Total_BR x Total_RB

Concluí-se com esse gráfico que um time de melhor qualidade jogando em casa, contra uma equipe de qualidade ruim, tem um desempenho muito superior a um time de qualidade ruim recebendo uma equipe de boa qualidade.

Após a análise gráfica, o teste de hipóteses foi aplicado para testar se as duas amostras são da mesma cadeia. Porém, o teste não teve poder para rejeitar a hipótese que elas são da mesma cadeia (p = 0,1055). Isso pode ter ocorrido porque ao fazer a estratificação o tamanho da amostra diminui muito e acabou ficando muito pequeno para provar essa hipótese.

3.3.3) Comparação das CM: Total_melhor_casa x Total_melhor_fora

A ideia para corrigir o problema do tamanho da amostra foi juntar as matrizes onde o time da casa é de força superior ao adversário e vice-versa. Assim, a primeira matriz ficará com os resultados de partidas de time bons contra ruins, bons contra médios, e médios contra ruins. Porém, a segunda matriz será montada com o raciocínio contrário, onde o time superior sempre jogará fora de casa, ruins contra bons, ruins contra médios e médios contra bons.

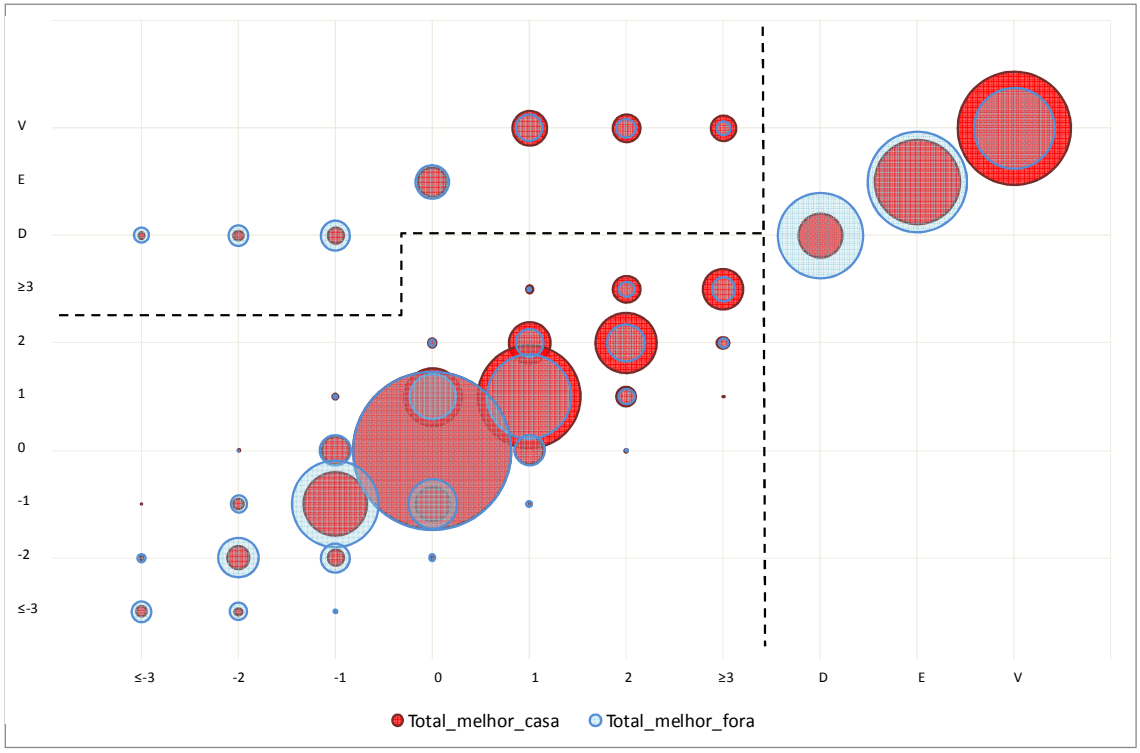

Figura 7: Comparação entre as matrizes de transição Total_melhor_casa x Total_melhor_fora

A Figura 7 mostra as mesmas tendências do gráfico apresentado na subseção anterior. Onde a matriz Total_melhor_casa mostra um desempenho superior, mantendo os resultados quando os mesmo o favorecem e tendo transições maiores independentemente do placar.

Porém, dessa vez a aplicação do teste de hipóteses provou essa diferença, rejeitando a hipótese nula (p = 0,0014). Provando assim que as duas matrizes não são da mesma cadeia.

A ideia nesse momento foi separar os países e testar se esse comportamento é igual em ambos, ou se essa diferença ocorre devido a apenas um dos dois campeonatos.

3.3.4) Comparação das CM: Italia_mehor_casa x Italia_melhor_fora

A Figura 8 novamente mostra que a qualidade do time aparentemente faz diferença no campeonato italiano. O teste de hipóteses confirma a ideia (p = 0,0456) rejeitando a hipótese nula. Assim, afirma-se que as duas amostras não são da mesma cadeia. Isso prova que a qualidade do time é importante na Itália. Ou, que uma equipe de melhor qualidade além de ser

favorecido em sua casa, consegue também trazer bons resultados de partidas jogadas fora de seu estádio.

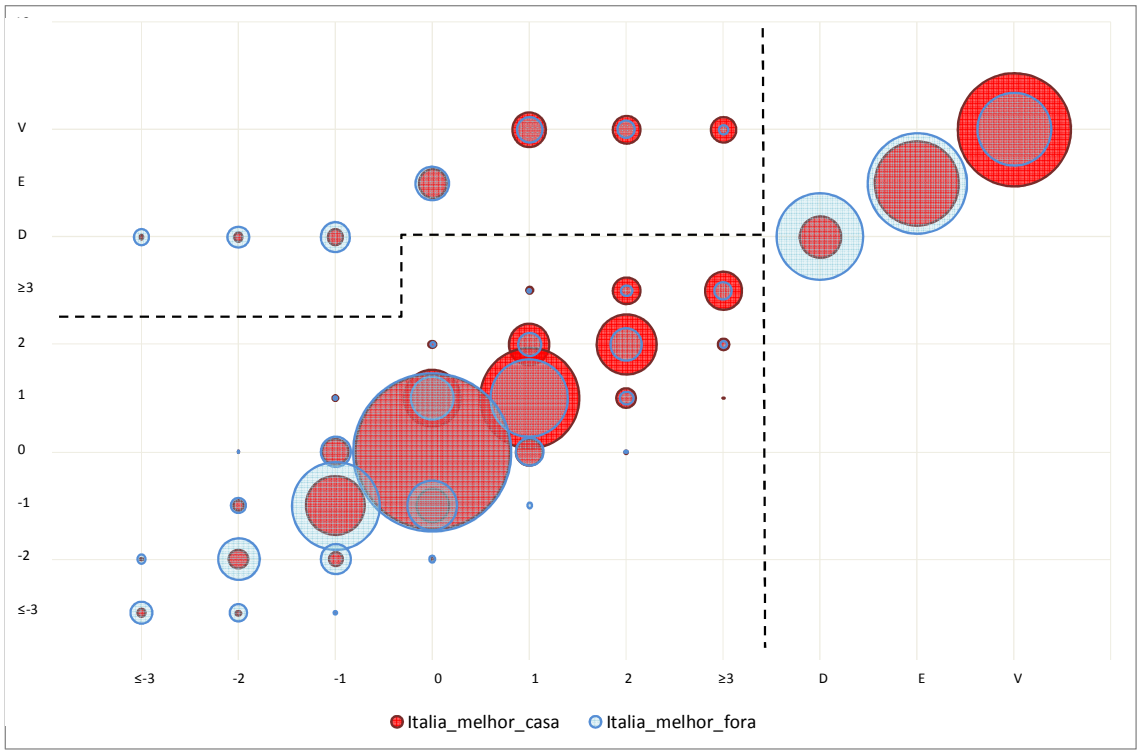

Figura 8: Comparação entre as matrizes de transição Itália_melhor_casa x Itália_melhor_fora

3.3.5) Comparação das CM: Brasil_melhor_casa x Brasil_melhor_fora

Parte-se agora para o último teste, onde se testará se existe diferença na comparação de times de melhor qualidade jogando em casa ou fora de casa no campeonato brasileiro.

Primeiramente observa-se a Figura 9, em que fica claro que ao contrário do comportamento do torneio italiano, o campeonato brasileiro não mostra diferenças muito grandes nessa comparação. Aparentemente o fator qualidade do time não influencia tanto o resultado quanto na Itália.

O teste de hipóteses foi aplicado e seu resultado não teve poder suficiente para rejeitar a hipótese nula. Ou seja, não se pode afirmar que as duas amostras apresentadas não são da mesma cadeia (p-valor 0,9090). Isso significa que no campeonato brasileiro independentemente da qualidade das equipes que se enfrentam o que prepondera é o fator

casa, isso fica provado quando separa-se os times por qualidade e não descobre-se diferença nessa separação.

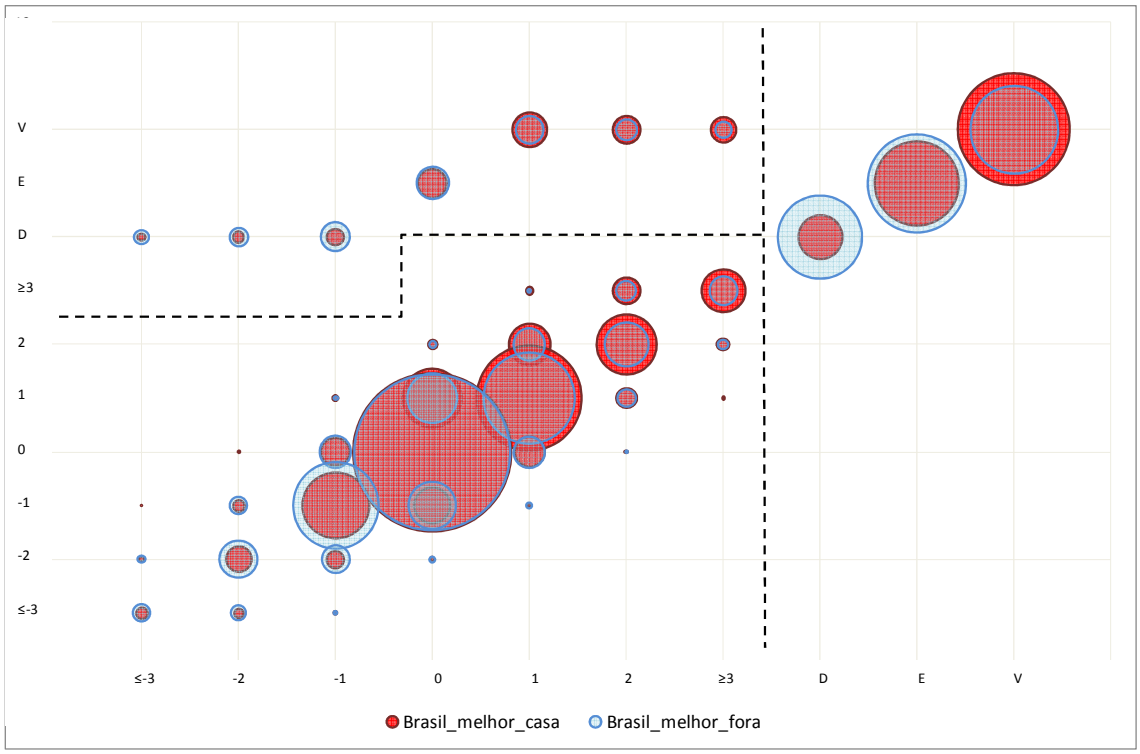

Figura 9: Comparação entre as matrizes de transição Brasil_melhor_casa x Brasil_melhor_fora

Esses testes servem para demonstrar que na Itália a diferença de qualidade dos times é muito maior do que no Brasil. Os times italianos de qualidade superior costumam vencer mesmo estando fora de seu estádio. No Brasil, esses mesmos times de maior qualidade não conseguem vencer tanto fora de suas casas, assim emparelhando o campeonato.

Ao observar o aproveitamento da equipe campeã de cada um dos anos avaliados na amostra, observa-se outra prova de que o campeonato brasileiro é mais disputado. Enquanto os oito campeões brasileiros tiveram uma média de 65,5% de aproveitamento, os campeões italianos tiveram uma média de 74,2%. Mais um argumento que reforça a ideia de que uma equipe de melhor qualidade no Brasil tem mais dificuldade de conquistar pontos do que na Itália.

Outra comparação que pode ser feita é que na amostra observada foram apresentados oito campeonatos brasileiros, nela apenas um time ganhou mais de uma vez o campeonato e

seis diferentes equipes foram campeãs. Na Itália foram dez campeonatos e apenas cinco diferentes campeões. Essa diferença não parece tão grande, provavelmente devido à pequena amostra que se tem para o Brasil. Porém se for pensada na totalidade dos campeonatos italianos o cenário muda drasticamente. É sabido que o campeonato é dominado por três diferentes equipes (Milan, Juventus e Internazionale) que juntas venceram 63 dos 107 (58,9%) campeonatos. Enquanto nessa mesma proporção, juntando os três maiores campeões do campeonato brasileiro desde 1971, seria de 37,5%. Mesmo com a modificação da CBF que considerou os títulos da Taça Brasil (década de 50 e 60) o valor brasileiro continua muito inferior 40,7%. Isso reforça a idéia de que o campeonato brasileiro é mais disputado, pelo fato das equipes terem nível mais parecido e tornando a disputa mais democrática, onde diferentes equipes conseguem conquistar o título ou ficar no final do ano no topo da tabela.

3.3.6) Total_igual, Italia_igual e Brasil_igual

As matrizes que foram compostas com times de mesma força também foram testadas. Tanto conjuntamente como separadamente pelos dois países. Essas matrizes foram comparadas com as amostras compostas pelos casos onde melhor time estava em casa ou fora.

A matriz total dos times de mesma força comparada a matriz total onde a melhor equipe estava jogando em casa teve a hipótese nula aceita, com o valor-p de 0,8731. O mesmo ocorreu ao ser comparada com a matriz total onde a melhor equipe estava fora (p = 0,9663). No Brasil, espera-se o mesmo comportamento e ambos os testes resultaram num valor-p igual a 0,9999. A última comparação que restou foi para o campeonato italiano, onde os resultados foram semelhantes. Quando o melhor time está em casa valor-p é 0,9656, já para a melhor equipe fora valor é de 0,9910.

Como ficou claro nos testes anteriores que a liga brasileira e a italiana têm comportamentos diferenciados, a decisão de separá-las parece pertinente. Assim, separou-se a variável país para a criação do modelo e serão utilizadas as matrizes que consideram o campeonato.

No caso do campeonato brasileiro não existem diferenças entre nenhuma das três matrizes apresentadas. Então, será criado apenas um modelo para a liga do Brasil onde a qualidade da equipe será desconsiderada.

Mesmo que a matriz de times de força igual não tenha diferença significativa quando testada contra as matrizes de diferentes forças. O campeonato italiano mostrou uma diferença significativa quando introduzida a variável da qualidade. Intuitivamente parece errado deixar esses casos em um mesmo modelo. Por isso, as separaremos em três diferentes modelos.

3.4) **Teste de Aderência**

Para uma melhor avalição da pertinência dos modelos utilizou-se o banco de dados citado anteriomente composto pelos campeonatos brasileiro, italiano, inglês, alemão e espanhol. Os resultados dessas temporadas foram utilizados para testar a aderência dos modelos propostos no estudo. A cor vermelha representa as temporadas aderiram ao modelo, em azul as temporadas que não aderiram.

3.4.1) Modelo 1 – Brasil

A Figura 10 mostra um comportamento adequado, e juntamente com o teste qui-quadrado de aderência (p = 0,3334) pode-se comprovar que o modelo parece pertinente. O acerto do modelo para os três resultados é bem similar aos dados das 32 primeiras rodadas do campeonato de 2011.

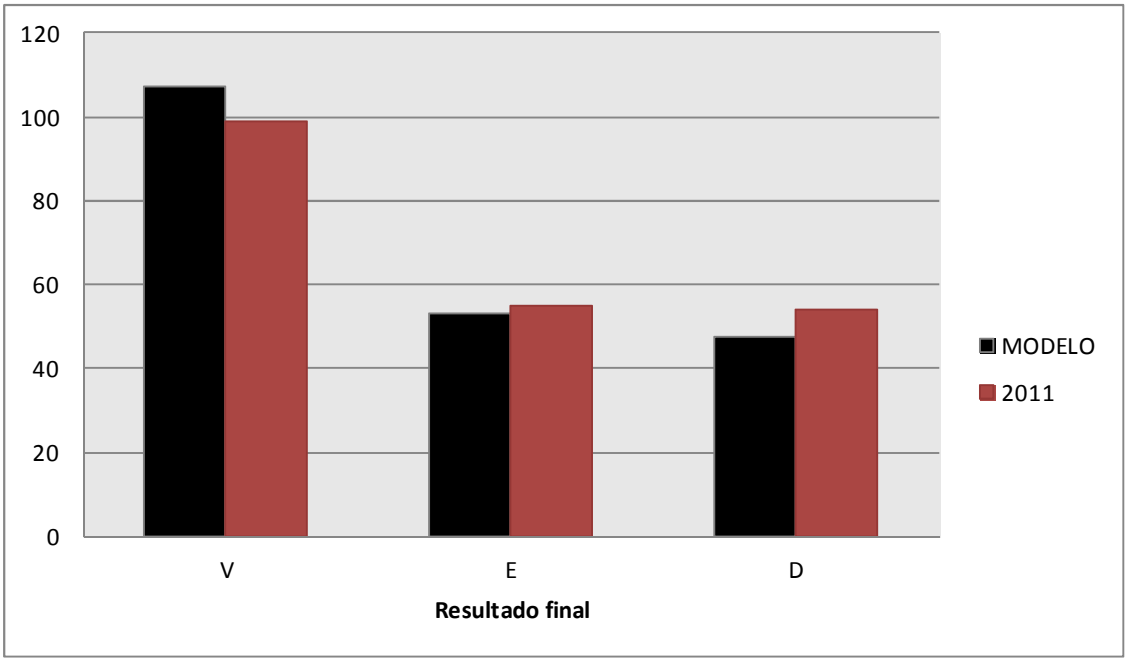

Figura 10: Teste de aderência do Modelo 1 – Brasil

3.4.2) Modelo 2 – Italia_melhor_casa

O primeiro modelo italiano, ilustrado na Figura 11 mostrou um ótimo acerto para as cinco temporadas desse mesmo campeonato. Testando esse mesmo padrão para os dez anos do torneio inglês vemos uma boa aderência, onde nove das dez competições apresentadas mostraram aderir ao padrão proposto. Um comportamento semelhante foi encontrado no campeonato alemão, onde o modelo teve bom acerto em oito dos dez campeonatos e, finalmente a competição espanhola que nos mostrou um comportamento tão eficiente, onde metade dos torneios aderiram ao modelo.

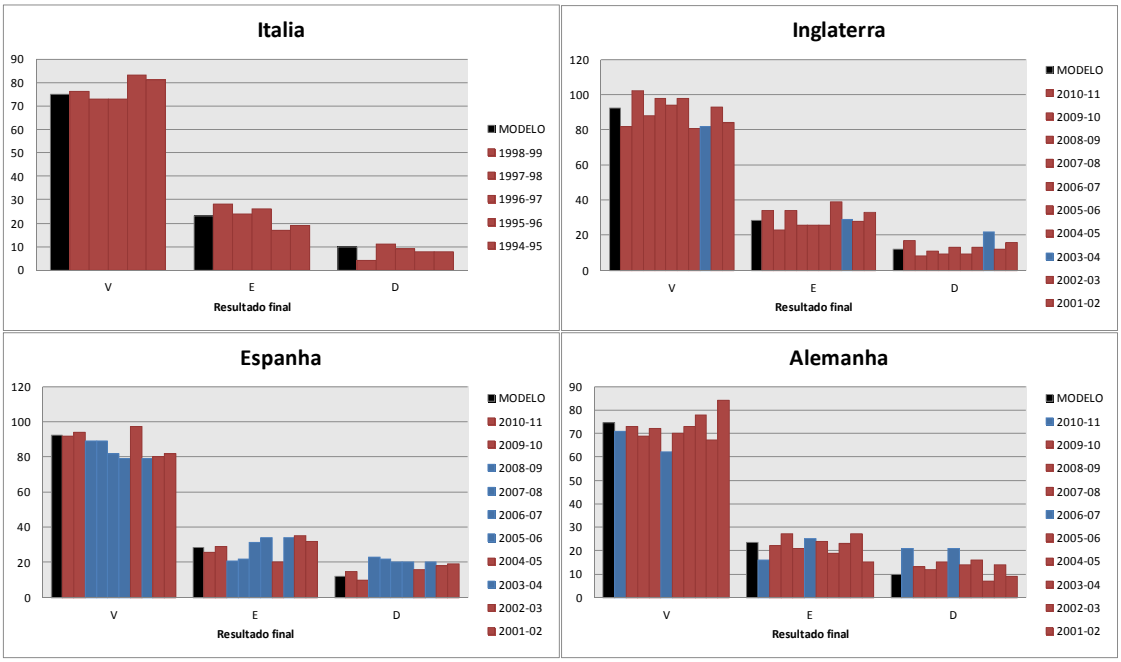

Figura 11: Teste de aderência do Modelo 2 – Italia_melhor_casa

3.4.3) Modelo 3 - Italia_igual

A matriz de times de mesma força do campeonato da Itália também apresentou um bom desempenho nos cinco anos de torneio italiano estudados, apenas uma matriz em uma competição não aderiu ao padrão, a Figura 12 representa esse modelo. O desempenho do modelo pareceu pertinente para os campeonatos inglês e alemão, onde o os resultados esperados não acertaram em duas temporadas para cada um dos países. Novamente o campeonato espanhol não obteve um resultado tão bom, apenas metade das vezes o modelo pareceu se adaptar bem aos dados coletados.

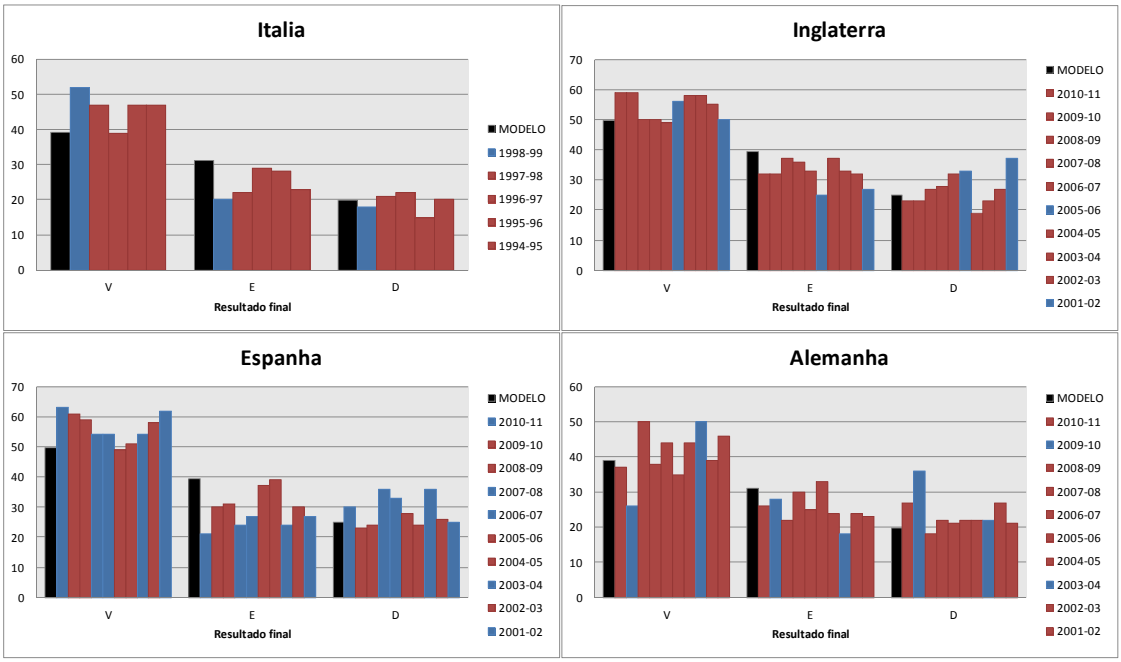

Figura 12: Teste de aderência do Modelo 2 – Italia_igual

3.4.4) Modelo 4 - Italia_melhor_fora

O último modelo, Figura 13, estudado é o formado pela matriz onde as equipes de melhor qualidade se encontram fora de casa enfrentado equipes de pior qualidade. O teste não foi tão satisfatório quanto os outros modelos para os resultados do campeonato italiano, onde dois dos cinco anos estudados não obtiveram aderência ao modelo proposto. O modelo parece satisfatório para o campeonato inglês onde o acerto foi de sete dos dez campeonatos estudados. No campeonato alemão e espanhol o desempenho foi ainda melhor onde às dez temporadas tiveram aderência ao modelo.

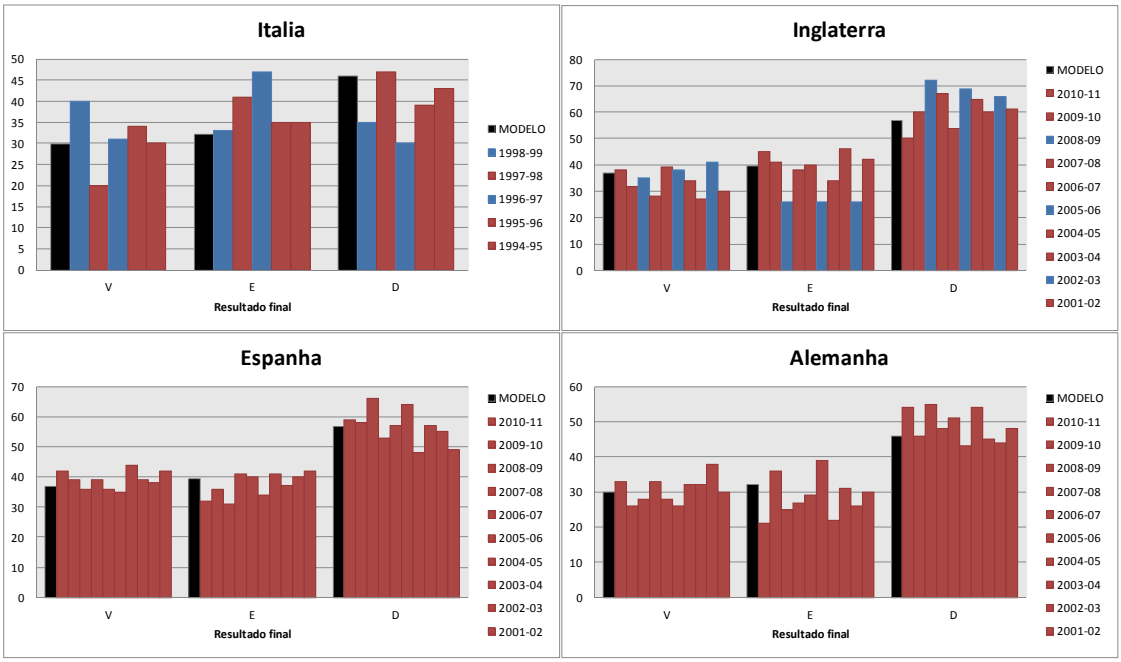

Figura 13: Teste de aderência do Modelo 2 – Italia_melhor_fora

3.5) **Tempos de recorrência e absorção**

O estudo parte agora para os tempos de absorção e recorrência médios, das matrizes propostas. Cada um dos quatro modelos será estudado separadamente e os resultados obtidos neles serão comparados com o intuito de novamente mostrar as diferenças das matrizes de transição selecionadas.

Como é sabido, o modelo propõe que cada partida tenha vinte intervalos de tempo até o jogo acabar. As tabelas abaixo mostram quanto tempo, em média, o processo ficará em cada um dos estados, dado que ele começou em um dos estados apresentados na parte vertical da tabela, até ele chegar ao seu resultado final e ser absorvido.

3.5.1) Modelo 1 – Brasil

O primeiro exemplo será novamente o modelo brasileiro. Onde observa-se que a partir do momento em que o jogo está com três ou mais gols de diferença, tanto a favor como contra o mandante, ele costuma se resolver rapidamente, pois em média, nos próximos 4,7 intervalos, algo em torno de 25 minutos o jogo estará decidido. Isto

porque equipes provavelmente demoram mais tempo para chegar a placares elásticos como esses, e quando chegam muitas vezes acabam por diminuir o ritmo da partida, parando de marcar gols e fazendo com que o processo seja absorvido.

Pode-se observar também que o tempo médio de absorção favorece o mandante, já que o jogo acaba por se decidir mais rápido para placares onde o time visitante está perdendo.

Na Tabela 1, nota-se que quando partida está empatada (saldo inicial = 0) ela tende a ficar em média aproximadamente 40 (5 minutos x 7,779 intervalos de tempo) minutos nesse placar. Logo, observa-se também na linha central (saldo inicial = 0) um deslocamento para o lado esquerdo, devido ao mandante ter vantagem, e obter maior número de vitórias, fazendo assim o processo ficar mais tempo (em relação aos momentos onde o visitante está ganhando) com o mandante a frente do placar.

Tabela 1: Tempos de recorrência da CM dos resultados do Modelo 1 - Brasil

Saldo	*Saldo após 5 minutos*							
Inicial	**≥3**	**2**	**1**	**0**	**-1**	**-2**	**≤-3**	**Total**
≥3	3,430	0,753	0,333	0,158	0,043	0,009	0,002	**4,729**
2	1,400	3,535	1,476	0,704	0,190	0,040	0,010	**7,356**
1	0,728	1,746	5,401	2,480	0,672	0,142	0,035	**11,204**
0	0,454	1,091	3,234	7,779	2,016	0,428	0,105	**15,107**
-1	0,238	0,573	1,703	3,906	5,367	1,099	0,269	**13,155**
-2	0,097	0,233	0,691	1,591	2,065	3,384	0,779	**8,839**
≤-3	0,026	0,063	0,189	0,434	0,567	0,827	2,604	**4,710**

A Tabela 2 a ser apresentada mostra a probabilidade de absorção, dado que o processo se encontra em um dos sete estados, saldo de gols naquele momento, e acabará em um dos estados absorventes. Por exemplo, se uma partida se encontra com uma diferença de três gols, ou mais (saldo inicial ≥ 3), a favor do mandante a chance de esse jogo ser absorvido pelo estado vitória (V), ou seja, acabar com vitória da equipe dona da casa é de 98,1%. A linha central da tabela apresenta os valores quando a partida está empatada. Assim a probabilidade da equipe mandante vencer no campeonato brasileiro quando a partida inicia (ou está empatada) é de 51,5%, de empatar é de 25,6% e os outros 22,9% são da equipe da casa perder.

Nesta tabela, também pode-se observar que a partir do momento que a diferença é a favor da equipe dona da casa, a chance do jogo ser absorvido pelo estado da vitória dessa equipe aumenta. Logo, suas probabilidades são superiores que àquelas de absorção para as mesmas diferenças a favor do visitante. Indicando que quando a equipe mandante consegue uma diferença de gols, suas chances de vencer a partida são bem superiores aos momentos em que o visitante obtém essa diferença.

Tabela 2: Probabilidades de absorção do Modelo 1 – Brasil

Saldo	*Resultado final*		
Inicial	**V**	**E**	**D**
≥3	0,990	0,005	0,005
2	0,955	0,023	0,022
1	0,842	0,082	0,076
0	0,515	0,256	0,229
-1	0,271	0,129	0,601
-2	0,110	0,052	0,838
≤-3	0,030	0,014	0,956

3.5.2) Modelo 2 – Italia_melhor_casa

Observa-se o comportamento do modelo onde as equipes de melhor qualidade jogam em seu território, Tabela 3, pode-se reparar que o jogo demora mais tempo para ser absorvido quando a equipe inferior está à frente, ou seja, uma equipe de melhor qualidade consegue prolongar mais o jogo, e brigar mais tempo pela vitória.

Vale destacar que o tempo de absorção quando a equipe superior está perdendo por um gol de diferença (saldo inicial = -1) é maior que quando o placar está igual (saldo inicial = 0), o que demonstra uma dificuldade maior do visitante decidir a partida. Mesmo abrindo um gol de diferença a favor do visitante, o tempo total para o jogo ser absorvido continua em torno de quinze intervalos de tempo.

Outra grande diferença é a comparação entre os tempos de absorção quando o placar está com diferença de três ou mais gols (saldo inicial ≥ 3). O tempo para o jogo se decidir quando o mandante está vencendo é muito menor que quando o visitante está ganhando por este placar. Esta tabela mostra uma clara vantagem do mandante, e mostra que ele se impõe devido a sua qualidade superior.

Tabela 3: Tempos de recorrência da CM dos resultados do Modelo 2 – Italia_melhor_casa

Saldo	*Saldo após 5 minutos*							
Inicial	**≥3**	**2**	**1**	**0**	**-1**	**-2**	**≤-3**	**Total**
≥3	3,056	0,681	0,224	0,088	0,014	0,002	0,000	**4,064**
2	1,365	3,617	1,100	0,436	0,071	0,008	0,001	**6,599**
1	0,797	1,992	5,011	1,777	0,287	0,034	0,005	**9,904**
0	0,599	1,501	3,671	7,921	1,280	0,154	0,023	**15,149**
-1	0,414	1,038	2,541	5,250	5,273	0,624	0,095	**15,235**
-2	0,225	0,563	1,379	2,854	2,763	3,159	0,431	**11,373**
≤-3	0,075	0,188	0,460	0,951	0,921	1,053	3,255	**6,902**

Novamente entende-se uma grande diferença entre o mandante e o visitante na Tabela 4. A partir do momento em que o mandante abre o (saldo inicial = 1), a chance de vencer o jogo é superior a 93,1%. Ainda a partir do momento em que o mandante abre dois gols (saldo inicial = 2) no placar sua chance é de 98,3%. A linha central (saldo inicial = 0) da tabela mostra uma grande vantagem da equipe mesmo quando não existe diferença, lembrando que as partidas sempre começam nesse estado.

Mesmo quando o time visitante abre o placar (saldo inicial = -1), a vantagem continua sendo do mandante, mostrando que a qualidade superior da equipe dona da casa pode acarretar na virada do resultado. Além disso, ressalta-se que a probabilidade dessa virada é maior que do jogo acabar apenas empatado.

Nenhum dos outros modelos mostrou tamanho favorecimento como o Modelo 2. Isso fica muito claro porque a equipe além de estar em seu próprio território tem qualidade superior ao adversário.

Tabela 4: Probabilidades de absorção do Modelo 2 – Italia_melhor_casa

Saldo	*Resultado final*		
Inicial	**V**	**E**	**D**
≥3	0,997	0,002	0,001
2	0,983	0,012	0,005
1	0,931	0,048	0,021
0	0,692	0,216	0,092
-1	0,479	0,143	0,378
-2	0,260	0,078	0,663
≤-3	0,087	0,026	0,888

3.5.3) Modelo 3 – Italia_igual

O terceiro modelo, mostrado na Tabela 5, novamente favorece a equipe dona da casa. Os tempos de absorção são menores a partir do momento em que o mandante está ganhando, e as partidas demoram, em média, mais tempo para se resolver quando o visitante está na frente do placar.

Tabela 5: Tempos de recorrência da CM dos resultados do Modelo 3 - Italia_igual

Saldo	*Saldo após 5 minutos*							
Inicial	**≥3**	**2**	**1**	**0**	**-1**	**-2**	**≤-3**	**Total**
≥3	2,848	0,860	0,430	0,216	0,062	0,013	0,003	**4,432**
2	0,859	3,481	1,740	0,875	0,252	0,054	0,013	**7,273**
1	0,394	1,493	5,724	2,807	0,808	0,173	0,042	**11,440**
0	0,220	0,835	3,075	7,795	2,187	0,469	0,114	**14,694**
-1	0,124	0,471	1,735	4,304	5,697	1,175	0,285	**13,791**
-2	0,054	0,204	0,751	1,864	2,412	3,742	0,888	**9,914**
≤-3	0,019	0,071	0,263	0,652	0,844	1,310	3,211	**6,370**

Como foi mostrado nos modelos anteriores, o domínio continua sendo do mandante, aqui a qualidade é supostamente igual, porém o fator casa continua favorecendo o mandante, e isso se comprova quando observamos a linha central da Tabela 6 (saldo inicial = 0), onde a diferença é nula e mostra as probabilidades no início da partida. Novamente se observa que as probabilidades de absorção favorecem o mandante, dado que a partir do momento em que ele está vencendo a chance do processo ser absorvido é maior do que que a chance do processo ser absorvido em resultados com diferença a favor do visitante.

Tabela 6: Probabilidades de absorção do Modelo 3 – Italia_igual

Saldo	*Resultado final*		
Inicial	**V**	**E**	**D**
≥3	0,984	0,010	0,006
2	0,936	0,039	0,025
1	0,794	0,125	0,081
0	0,434	0,346	0,220
-1	0,245	0,191	0,564
-2	0,106	0,083	0,811
≤-3	0,037	0,029	0,934

3.5.4) Modelo 4 – Italia_melhor_fora

Este é o modelo que apresenta um desempenho diferenciado. A Tabela 7, que mostra o número de intervalos que a partida demora a ser absorvida, mostra um comportamento antagônico aos outros três modelos, onde o tempo que a partida demora a ser absorvida é menor quando favorece o visitante. Isso se justifica porque neste modelo a equipe de melhor qualidade se encontra fora de casa, e sua superioridade supera o fator casa, tornando nesses casos o visitante como sfavorito.

Tabela 7: Tempos de recorrência da CM dos resultados do Modelo 4- Italia_melhor_fora

Saldo	*Saldo após 5 minutos*							
Inicial	**≥3**	**2**	**1**	**0**	**-1**	**-2**	**≤-3**	**Total**
≥3	3,487	1,131	0,628	0,455	0,162	0,043	0,013	**5,919**
2	0,778	3,112	1,727	1,250	0,446	0,118	0,035	**7,466**
1	0,281	1,054	5,393	3,569	1,278	0,339	0,101	**12,016**
0	0,117	0,440	2,172	8,261	2,824	0,750	0,224	**14,788**
-1	0,044	0,167	0,824	3,048	5,289	1,357	0,406	**11,135**
-2	0,012	0,045	0,220	0,816	1,378	3,279	0,946	**6,696**
≤-3	0,003	0,011	0,055	0,204	0,345	0,820	2,833	**4,270**

A Tabela 8 mostra as probabilidades de absorção. Ademais, pela primeira vez na comparação entre os placares onde a diferença é positiva e negativa pode-se observar maiores probabilidades favorecendo o visitante.

A linha central (saldo inicial = 0) dessa vez favorece o time que está fora de casa, novamente mostrando a superioridade do visitante devido a sua maior qualidade desde o início da partida.

Tabela 8: Probabilidades de absorção do Modelo 4- Italia_melhor_fora

Saldo	*Resultado final*		
Inicial	**V**	**E**	**D**
≥3	0,959	0,016	0,024
2	0,888	0,045	0,067
1	0,679	0,128	0,192
0	0,277	0,297	0,426
-1	0,105	0,110	0,785
-2	0,028	0,029	0,943
≤-3	0,007	0,007	0,986

3.6) **Comparações entre os Modelos**

Os tempos de absorção já foram comentados nas tabelas de cada modelo especificamente. Porém, nesta seção o objetivo é compará-los de forma mais clara. Observa-se que o tempo de absorção dos quatro modelos é muito parecido quando a diferença de gols é nula. Isso porque independentemente da qualidade das equipes, de onde o jogo ocorre e do campeonato, a partida, obviamente, sempre começa em 0x0.

O modelo brasileiro não mostra nenhum padrão quando comparado aos outros modelos, porém os três modelos italianos apresentam diferenças bem visíveis nos gráficos.

Quando o placar favorece o mandante, o jogo costuma se decidir mais rapidamente no modelo em que o dono da casa tem melhor qualidade, o modelo com equipes de mesma qualidade vem na sequência, e por último o modelo que a equipe de pior qualidade está fora de casa. Esse comportamento se repete de maneira contrária em placares com diferença a favor do visitante.

Uma justificativa para isso é que quanto menor o tempo para a absorção do processo, mais perto ele está de ser decido, mais fácil de esse jogo ter seu resultado final definido. Por isso, o menor tempo de absorção está quando a equipe de melhor qualidade está em casa e vencendo por três ou mais gols de diferença, ou seja, o cenário onde todas as variáveis favorecem uma equipe. Esses tempos vão aumentando quanto mais perto eles chegam da diferença nula, isso porque uma diferença nula inclui os placares que apenas um gol pode mudar o resultado da partida, assim deixando o resultado mais distante de ser absorvido.

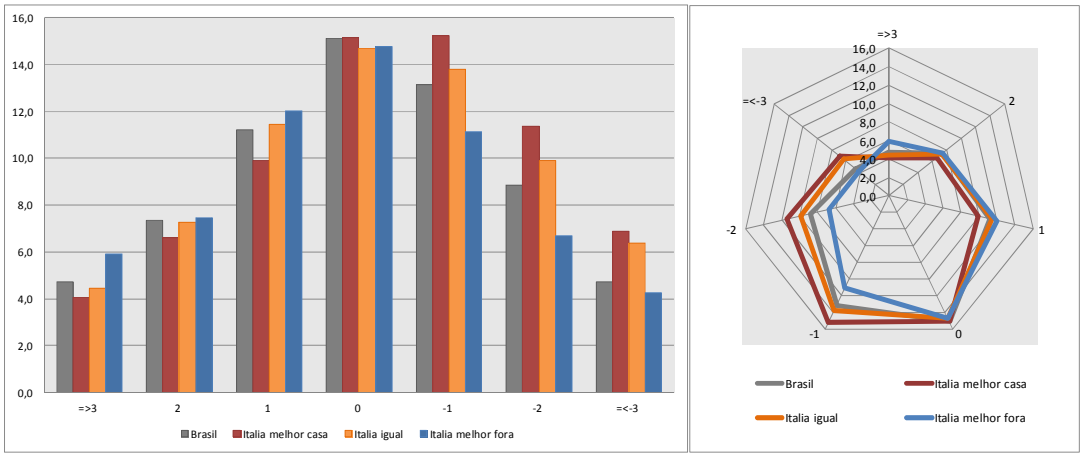

Figura 14: Comparação dos tempos de absorção dos modelos

Os gráficos da Figura 14 apresentam comparações entre as probabilidades do jogo ser absorvido (por qualquer um dos três possíveis resultados) considerando o atual estado em que o processo se apresenta, mais especificamente, na diferença de gols naquele momento.

3.6.1) Vitórias dos Mandantes

Uma primeira comparação é nas probabilidades de absorção de vitórias do mandante. Obviamente o desempenho do modelo Itália_melhor_casa é muito superior aos outros, esses são os casos que a melhor equipe se encontra em casa. O modelo brasileiro mostrou as segundas maiores probabilidades de absorção. Em seguida, o modelo italiano de equipes de mesma força, e por último, com um desempenho muito inferior, o modelo onde a equipe de melhor qualidade se encontra fora. A Figura 15 demonstra que a superioridade das equipes é visível, pois independente do placar em que a partida se encontra, a probabilidade de uma equipe de qualidade superior confirmar a vitória em um jogo é sempre superior aos outros times obterem o mesmo resultado.

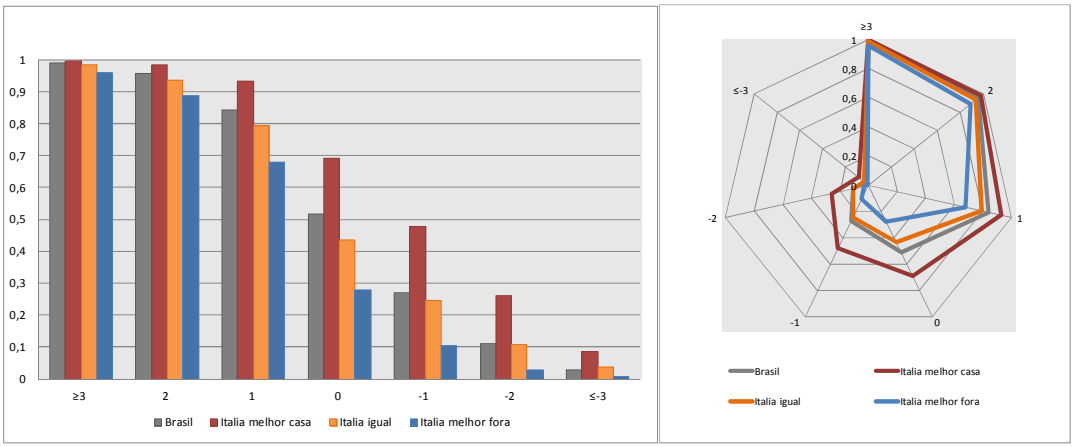

Figura 15: Comparação das probabilidades de vitória dos modelos

3.6.2) Empates

No caso dos empates, vemos que independente do modelo, a maior probabilidade de absorção é sempre quando o jogo não tem diferença de gols, isso porque a única diferença de gols que caracteriza o empate é a diferença nula. A Figura 16 mostra um decaimento quando mais distante do centro, a justificativa é porque quanto maior a diferença mais longe a equipe que está sendo derrotada está de recuperar o placar e empatar a partida.

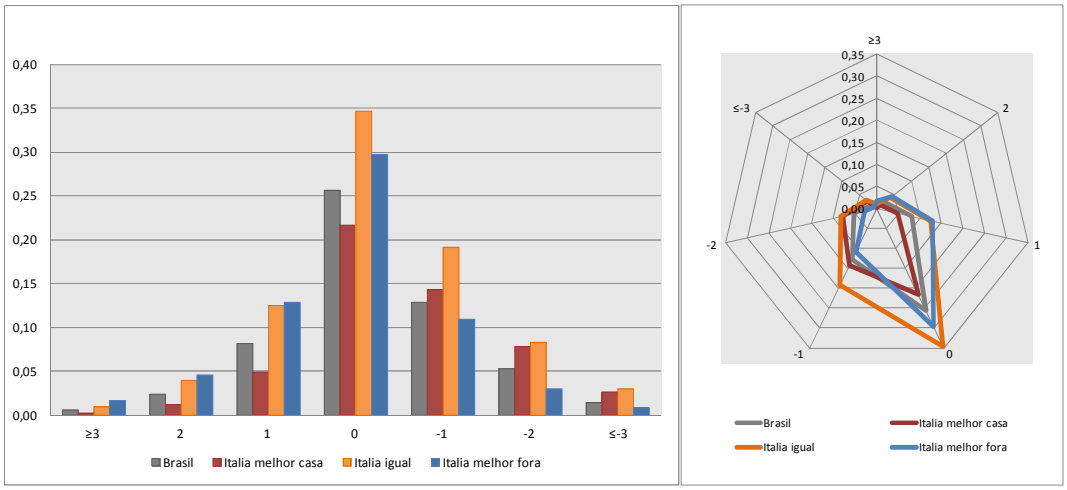

Figura 16: Comparação das probabilidades de empate dos modelos

3.6.3) Derrotas dos Mandantes

A última comparação, na Figura 17, prova novamente uma grande diferença entre os modelos apresentados. Onde a probabilidade de absorção de derrota do mandante é sempre superior nas situações onde o time de melhor qualidade está vencendo a partida. Mesmo ele jogando fora de casa, consegue ter uma grande vantagem devido a sua melhor qualidade. O desempenho do modelo Italia_melhor_fora é superior em todo o gráfico, destacando a situação onde o jogo está empatado (inclui o início da partida) e já nesse momento a probabilidade absorção desse modelo se destaca tamanha sua superioridade em relação aos outros modelos.

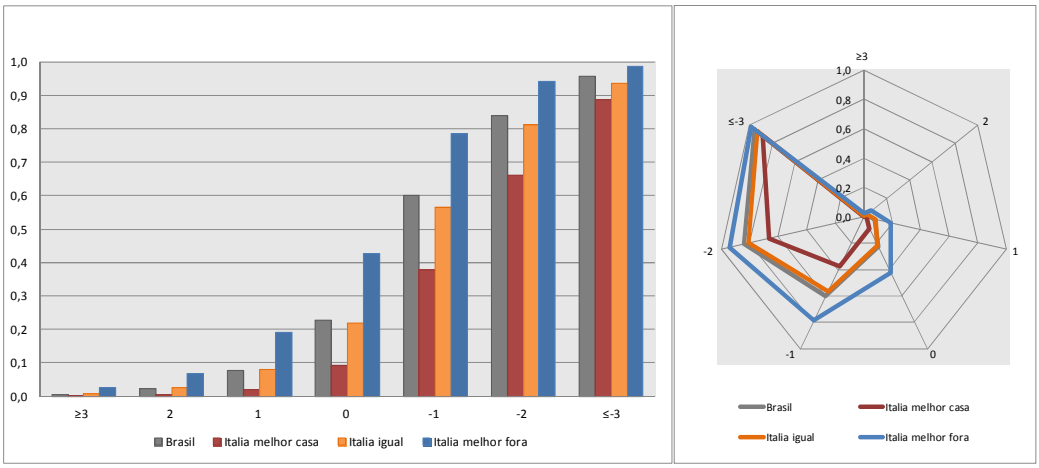

Figura 17: Comparação das probabilidades de derrotas dos modelos

A seleção dos quatro modelos pareceu pertinente. Os gráficos e teste provaram que o campeonato brasileiro e italiano têm diferenças significativas, e juntar suas informações parecia errado, pois seus comportamentos são muito distintos. O mesmo raciocínio valeu para a separação dos modelos italianos dentro de seu próprio campeonato, ficou muito claro que a força das equipes influencia muito mais o resultado do jogo que apenas a o fator casa.

Os testes apresentados mostraram um bom desempenho do modelo para a liga brasileira, assim como os testes realizados para o modelo italiano, onde outros campeonatos europeus de comportamento parecido também foram testados.

Finalmente, os resultados das probabilidades de absorção e recorrência dos estados foram apresentados reforçando a ideia de que os três modelos italianos têm comportamentos muito distintos, onde a qualidade das equipes sobrepõem o local da partida. Além disso, reforça a tese de que eram necessários modelos diferentes.

Em resumo, a análise proposta mostrou a diferença que o campeonato brasileiro e italiano tem, onde o segundo mostra uma diferença muito maior quando inserida a variável qualidade. Isto porque a estratificação em três grupos de diferentes qualidades mostrou que no campeonato nacional da Itália, a diferença entre esses grupos é muito grande, enquanto no toneio brasileiro essa diferença não é tão aparente. Esse fato levou a comparações dentro de cada um dos campeonatos nacionais, o que resultou na criação de quatro modelos. Três deles dentro do torneio italiano onde ficou mais clara ainda a importância da variável qualidade. Ainda, um modelo brasileiro geral que não utilizou a variável qualidade, pois nesse país o fator casa prepondera.

4) Conclusão

O presente estudo utilizou dados dos campeonatos brasileiro e italiano para investigar as diferenças entre esses campeonatos. Foram utilizandas, além do torneio, a qualidade das equipes e o local onde o jogo foi realizado. O estudo efetuou comparações entre essas competições, através de uma modelagem utilizando cadeias de Markov, e mostrou que o fator local da partida influencia mais no torneio brasileiro, enquanto que no campeonato italiano, a qualidade das equipes merece maior destaque.

A availiação desses modelos foi satisfatória, os modelos propostos utilizando o campeonato italiano mostraram um bom desempenho ao serem avaliados tanto em campeonatos de outros anos, quanto em outros torneios com características similares campeonatos alemão, inglês e espanhol. O modelo brasileiro também se mostrou satisfatório ao ser testado com a temporada do ano de 2011. Finalmente, os modelos tiveram seus tempos de absorção e recorrência avalidados, confirmando a ideia que entre eles existe uma grande diferença. Onde o domínio da equipe de melhor qualidade dentro de sua casa foi muito amplo no campeonato italiano, e mesmo jogando fora de casa os times de melhor qualidade obtiveram um desempenho muito positivo, até mesmo superando o fator casa tamanha sua superioridade.

O estudo buscou mostrar essa diferença, algo aparentemente subjetivo, entre os campeonatos através de um modelo estatístico para provar que ela realmente existe, ainda que diferentes campeonatos devam ser avaliados muitas vezes separadamente para melhor previsões e modelagens.

5) Referências Bibliográficas

ANDERS, A. e ROTTHOFF, K. W., Yellow Cards: Do They Matter?, Journal of Quantitative Analysis in Sports: Vol. 7 (1). 2011.

ANDERSON T.W. e GOODMAN L. A., Statistical Inference About Markov Chains, Ann. Math. Statist. Vol. 28, p.p 89-109. 1956.

BILLINGSLEY P., Statiscal Methods in Markov Chains, Ann. Math. Statist. Vol 32 (1), p.p 12-40. 1961.

GRINSTEAD C. e SNELL L., Introduction to Probability, Cap. 11, p.p 405-470. 1997.

GUTTORP P., Stochastic Modeling of Scientific Data, Chapman & Hall. 1995.

KRAUTMANN, A. C., CIECKA, J. E. e SKOOG, G. R., A Markov Process Model of the Number of Years Spent in Major League Baseball, Journal of Quantitative Analysis in Sports: Vol. 6 (4), 2010.

NEWTON, P. K. e ASLAM, K., Monte Carlo Tennis: A Stochastic Markov Chain Model, Journal of Quantitative Analysis in Sports: Vol. 5 (3), 2009.

OBERSTONE, J., Differentiating the Top English Premier League Football Clubs from the Rest of the Pack: Identifying the Keys to Success, Journal of Quantitative Analysis in Sports: Vol. 5 (3). 2009.

OBERSTONE, J., Comparing Team Performance of the English Premier League, Serie A, and La Liga for the 2008-2009 Season, Journal of Quantitative Analysis in Sports: Vol. 7 (1). 2011.

RUMP, C. M., The Effects of Home-Away Sequencing on the Length of Best-of-Seven Game Playoff Series, Journal of Quantitative Analysis in Sports: Vol. 2 (1) . 2006.

RUMP, C. M., Data Clustering for Fitting Parameters of a Markov Chain Model of Multi-Game Playoff Series, Journal of Quantitative Analysis in Sports: Vol. 4 (1). 2008.

VARGAS, R. N., Inferência Estocástica e Modelos de Mistura de Distribuições. 2011.

FOOTBALISTIC. Disponível em: <http://www.footbalistic.com> Acesso em: 15 abr. 2011.

THE REC.SPORT.SOCCER STATISTICS FOUNDATION. Disponível em: <http://www.RSSF.com> Acesso em: 27 mar. 2011.

ZEROZERO.PT. Disponível em: <http://www.zerozero.pt> Acesso em: 29 mar. 2011.