UM ARCABOUÇO COMPUTACIONAL PARA CARACTERIZAR SORTE E HABILIDADE EM LIGAS ESPORTIVAS

RAQUEL YURI DA SILVEIRA AOKI

UM ARCABOUÇO COMPUTACIONAL PARA CARACTERIZAR SORTE E HABILIDADE EM LIGAS ESPORTIVAS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Renato Martins Asunção
Coorientador: Pedro Olmo

Belo Horizonte
Abril de 2017

.

Ficha catalográfica elaborada pela Biblioteca do ICEx - UFMG

Aoki, Raquel Yuri da Silveira.

A638a Um arcabouço computacional para caracterizar sorte e habilidade em ligas esportivas./ Raquel Yuri da Silveira Aoki. – Belo Horizonte, 2017.
xxii, 64 f.: il.; 29 cm.

Dissertação (mestrado) - Universidade Federal de Minas Gerais – Departamento de Ciência da Computação.

Orientador: Renato Martins Assunção
Coorientador: Pedro Olmo Stancioli Vaz de Melo

1. Computação – Teses. 2. Teoria bayesiana de decisão estatística. 3. Modelos gráficos Probabilísticos. 4. Análise esportiva. I. Orientador. II. Coorientador. III. Título.

CDU 519.6*63(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Um arcabouço computacional para caracterizar sorte e habilidade em ligas esportivas

RAQUEL YURI DA SILVEIRA AOKI

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. RENATO MARTINS ASSUNÇÃO - Orientador
Departamento de Ciência da Computação - UFMG

PROF. PEDRO OLMO STANCIOLI VAZ DE MELO - Coorientador
Departamento de Ciência da Computação - UFMG

PROF. GUSTAVO ENRIQUE DE A. P. ALVES BATISTA
Instituto de Ciências Matemáticas e de Computação - USP

PROFA. JUSSARA MARQUES DE ALMEIDA GONÇALVES
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 26 de abril de 2017.

Dedico este trabalho a minha mãe e minha tia, pois sem o apoio delas nada disso seria possível.

Agradecimentos

Nenhuma conquista é individual. Uma conquista é sempre resultado de um conjunto de pessoas e seus esforços. Desta forma, direta ou indiretamente, muitas pessoas contribuíram para que a obtenção deste grau de mestre se tornasse realidade e quero deixar aqui registrado meu profundo agradecimento.

Agradeço primeiramente minha mãe Elza e minha tia Terezinha, que mesmo a distância e diante de muitas limitações sempre acreditaram em mim e me apoiaram em minhas escolhas. Sem esse suporte, nada disso seria possível. Obrigada por tudo.

Agradeço ao João Pedro Schneider pela paciência, companheirismo, amizade e suporte. Mesmo nos dias em que nada parecia dar certo, você estava presente para me ouvir e acreditar em mim, mesmo quando nem eu acreditava mais. Obrigada de coração.

Agradeço aos meus orientadores, Renato M. Assunção e Pedro O. S. Vaz de Melo, por todo o conhecimento compartilhando, paciência, apoio e suporte em todos os momentos. Obrigada por terem apostado mim quando os procurei 2 anos atrás e espero ter cumprido com as expectativas de vocês.

Agradeço aos meus amigos Gabriel Oliveira, Gustavo Savini, Rebeca Leão, Juliana Rocha e Daniel Fonseca que mesmo a distância ou em um almoço corrido tornaram meus anos na UFMG mais agradáveis. Aos amigos que fiz em Belo Horizonte Wagner Rodrigues, Victor Bastos, Nildo Júnior e tantos outros, obrigada pelas boas conversas na cantina do ICEX ou no ônibus interno. Obrigada a todos por terem me ouvido sempre que precisei de um ombro amigo. Agradeço aos amigos feitos durante o mestrado Fabricio Rodrigues, Vaux Gomes, Raphael Campos, Adriano Lages, Huggo Ferreira e tantos outros que frequentaram a sala 3015 (*in*

memorian) e compartilharam conhecimento, dúvidas e questões de provas anteriores. Nossos debates sobre assuntos irrelevantes, mas muito interessantes, era muitas vezes a melhor parte do meu dia.

Agradeço ao pessoal da Pró-Reitoria de Graduação da UFMG, em especial a Carolina Penna e o Tales Railton, por todo o suporte durante o tempo em que trabalhei na Reitoria e fiz o mestrado. Sem o apoio de vocês essa conquista teria sido muito mais difícil de ser obtida.

Por fim, um profundo e especial agradecimento a todos os meus mestres. Essa conquista de hoje é a soma dos esforços de muitos professores que encontrei durante todos os meus anos de escola pública. Professores muitas vezes mal remunerados e que trabalhavam em péssimas condições, mas que mesmo assim davam seu melhor e me estimularam a correr atrás dos meus sonhos. Também agradeço aos meus professores da UFMG que me ensinaram sobre essa profissão que poucos conhecem, a Estatística (ou, modernamente falando, a Ciência dos Dados). Hoje eu não consigo me imaginar trabalhando em outra área e amo profundamente minha profissão.

"...while the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician."

(Sir Arthur Conan Doyle, Sherlock Holmes)

Resumo

Prever o resultado de eventos esportivos é uma tarefa muito desafiadora. Este trabalho quantifica essa dificuldade através de um coeficiente que mede a distância entre o resultado final observado em ligas esportivas e o idealizado em competições completamente balanceadas em termos de habilidade. Este coeficiente indica a presença relativa de sorte e habilidade no campeonato. Foram coletados e analisados todos os jogos de 198 ligas esportivas, compostas de 1503 temporadas, oriundas de 84 países diferentes em 4 esportes: basquete, futebol, voleibol e handebol. Foi medida a competitividade por país e esporte. Também foram identificadas em cada temporada quais equipes deveriam ser removidas para que a liga ficasse completamente aleatória. Surpreendentemente, não é necessária a remoção de muitas equipes. Outra contribuição deste trabalho um modelo gráfico probabilístico cujo objetivo é aprender sobre as habilidades das equipes e decompor o peso relativo da sorte e da habilidade em cada partida. O componente da habilidade foi separado em variáveis associadas às características da equipe. O modelo também permite estimar como 0.36 a probabilidade do pior time, o chamado *underdog*, vencer uma partida na liga americana de basquete NBA. Como mostrado na primeira parte deste trabalho, a sorte está substancialmente presente mesmo nos campeonatos mais competitivos, o que parcialmente explica porque modelos sofisticados e complexos dificilmente conseguem ter resultados melhores que modelos mais simples na tarefa de prever resultados esportivos.

Palavras-chave: Modelos Gráficos Probabilísticos, Análise Esportiva, Estatística Bayesiana.

Abstract

Predicting the outcome of sports events is a hard task. We quantify this difficulty with a coefficient that measures the distance between the observed final results of sports leagues and idealized perfectly balanced competitions in terms of skill. This indicates the relative presence of luck and skill. We collected and analyzed all games from 198 sports leagues comprising 1503 seasons from 84 countries of 4 different sports: basketball, soccer, volleyball and handball. We measured the competitiveness by countries and sports. We also identify in each season which teams, if removed from its league, result in a completely random tournament. Surprisingly, not many of them are needed. As another contribution of this paper, we propose a probabilistic graphical model to learn about the teams' skills and to decompose the relative weights of luck and skill in each game. We break down the skill component into factors associated with the teams' characteristics. The model also allows to estimate as 0.36 the probability that an underdog team wins in the NBA league, with a home advantage adding 0.09 to this probability. As shown in the first part of the paper, luck is substantially present even in the most competitive championships, which partially explains why sophisticated and complex feature-based models hardly beat simple models in the task of forecasting sports' outcomes.

Keywords: Probabilistic Graphical Model, Sports Analytics, Bayesian Statistics.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Esportes são extremamente surpreendentes. Muitas vezes uma disputa parece estar com seu resultado definido e que não há nada que a outra equipe possa fazer para virar a partida e vencer. Entretanto, de alguma forma, a equipe que está perdendo vira o jogo de uma maneira inacreditável. Essas reviravoltas no esporte às vezes parecem ocorrer devido a eventos de pura sorte ocorridos durante a partida, mas em outras situações parecem ser devido à habilidade das equipes que estão na competição. Sorte ou habilidade, assistir a uma sólida vitória se desfazer no ar traz a sensação de que o mundo é imprevisível e incontrolável.

Existem alguns estudos teóricos e empíricos mostrando que a imprevisibilidade como um todo não pode ser evitada. Ainda assim, previsões precisas são o *holy grail* avidamente procurado no mercado financeiro [Chen & Du, 2009], em esportes [Chen & Joachims, 2016], política [Tumasjan et al., 2010] e em prêmios de artes e entretenimento [Haughton et al., 2015]. Essa influência da sorte não torna a tarefa de fazer previsões impossível, mas muito difícil. No basquete, por exemplo, apesar da grande quantidade de dinheiro envolvido, não existem algoritmos capazes de produzir predições acuradas e existem algumas evidências que eles nunca encontrarão tais algoritmos [Martin et al., 2016].

Competições esportivas são compostas por uma mistura de sorte e habilidade [Owen, 2013; Fort & Maxcy, 2003; Zimbalist, 2002] e aparecem como uma área atrativa de estudo por três razões. A primeira delas é que ligas esportivas formam um sistema relativamente isolado, com poucas e restritas influências externas além

de ser replicado sobre o tempo aproximadamente nas mesmas condições e sob as mesmas regras. A segunda razão é a grande quantidade de dados disponíveis, o que torna possível aprender seus padrões estatísticos. A terceira razão é sua popularidade, o que sustenta um comércio multibilionário que inclui televisão, propagandas e um enorme mercado de apostas. Todos os envolvidos em esportes, empresas, torcedores e jogadores iriam beneficiar-se de um melhor entendimento da imprevisibilidade nos esportes.

Essa mistura entre a sorte e a habilidade presente nos campeonatos esportivos é responsável por grande parte da atração que as pessoas sentem pelo esporte, como mostra Chan et al. [2009]. Enquanto um jogador individual pode encontrar diversão em um jogo totalmente definido pela habilidade dos competidores, como xadrez, ou jogos definidos puramente através da sorte, como loteria, para a grande maioria das pessoas que torcem por algum esporte, disputas que misturam sorte e habilidade soam mais interessantes. A explicação para essa preferência é simples: se o jogo é totalmente definido pelas habilidades das equipes, o resultado seria totalmente previsível e potencialmente chato [Fort & Quirk, 2011]. Da mesma forma, uma competição definida puramente através da sorte terá o interesse de uma audiência menor e consequentemente irá gerar menos lucro [Owen, 2013]. As finais de campeonatos decididos aleatoriamente, por exemplo, não seriam merecedoras de tanta atenção porque suas posições finais não serão resultado do mérito das equipes. Dessa forma, fortes emoções não serão despertadas nos torcedores [Khanin, 2000].

1.1 Contribuições

Diante deste contexto esportivo, a primeira contribuição que será apresentada nessa dissertação é o *coeficiente de habilidade*, denotado por ϕ, que mede em uma temporada esportiva quanto da distribuição final dos pontos observada é devido a habilidade das equipes. Além disso, o coeficiente ϕ mede o que esperar de um campeonato totalmente aleatório que considera somente circunstâncias contextuais em que a partida está ocorrendo. Quanto mais alto for o valor de ϕ, mais distante a liga está de uma competição cujo resultado é determinado aleatoriamente. Uma exigência para o cálculo de tal coeficiente é que a liga siga um formato de pontos-

corridos, em que todas as equipes do campeonato jogam uma contra as outras pelo menos duas vezes: uma como mandante e outra como visitante. Essa exigência existe devido a forma com que o campeonato aleatório é construído.

A partir do coeficiente ϕ foram derivadas duas técnicas para caracterizar o papel da habilidade em ligas esportivas. Na primeira delas foi proposto um teste de significância para o coeficiente ϕ que mostra, por exemplo, que a maioria das ligas de certos esportes, como handebol, não são diferentes de uma ordenação aleatória das equipes. A segunda técnica desenvolvida, que é baseada no coeficiente ϕ, identifica quais equipes são significativamente mais ou menos habilidosas que as demais equipes do campeonato. Através dessa técnica é possível mostrar que as ligas em que a competição é mais intensa e o suporte financeiro maior, como por exemplo a liga espanhola de futebol *Primera División*, a remoção de apenas 3 times de um conjunto de 30 equipes deixa a competição completamente aleatória. Também é mostrado através do coeficiente ϕ a presença de um raro fenômeno em algumas ligas presentes na base de dados. Nessas ligas, a pontuação de todas as equipes no final do campeonato foi exageradamente parecida, sendo tão extrema que é quase impossível que ela tenha sido gerada por uma competição totalmente aleatória, muito menos uma competição em que diferentes habilidades estão presentes.

Como anteriormente citado, o valor do coeficiente ϕ aumenta com o peso do componente habilidade nas competições, mas essa relação é desconhecida e não linear. Além disso, esse valor não responde à questão mais relevante: quanto da variação observada da distribuição final dos pontos é devido a sorte ou a habilidade? Um alto valor de ϕ pode indicar a presença de algumas poucas equipes *outliers* com habilidade extrema ou também pode indicar uma distribuição de habilidade sem *outliers* mas com uma excessiva curtose. Para tentar estimar as habilidades das equipes, a segunda contribuição dessa dissertação é um modelo gráfico probabilístico que estima as habilidades das equipes de uma temporada/liga baseando-se em características das equipes, dos jogadores e nos resultados finais ao longo do campeonato. Foi assumido um modelo linear generalizado com efeito aleatório, o que permite a inclusão de variáveis para explicar as diferenças entre as habilidades das equipes. Um resultado adicional é a análise de coeficientes associada com as variáveis sugeridas, o que pode contribuir para a construção de equipes mais competitivas.

Por fim, é apresentada uma caracterização da presença de sorte e habilidade nas ligas esportivas usando o coeficiente proposto ϕ. Foram analisados todos os jogos de 198 ligas esportivas de 4 esportes: basquete, futebol, handebol e voleibol. No total, 1503 temporadas de 84 países diferentes foram analisadas. A partir dessas análises, foi possível determinar qual esporte é mais provável de ter uma liga cujos resultados são totalmente determinados ao acaso. Além disso, foi mostrado para cada esporte qual é o percentual esperado de times que precisam ser removidos de uma temporada/liga para fazer com que essa temporada/liga seja decidida puramente pela sorte.

Capítulo 2

Trabalhos Relacionados

Dentro da dinâmica esportiva em que este trabalho está inserido, é muito importante compreender a influência da sorte na predição de resultados esportivos. Desta forma, a Seção 2.1 mostra alguns trabalhos sobre a previsão de resultados esportivos e uma breve descrição sobre suas técnicas e resultados. Também serão mostrados trabalhos que estimaram a habilidade das equipes de um campeonato ou investigaram formas de melhorar a performance das equipes. Essa caracterização de ligas esportivas é mostrada na Seção 2.2, em que serão descritas abordagens já utilizadas para prever o rank final de um campeonato e estudos sobre movimentos táticos.

2.1 Previsão de Resultados Esportivos

Um dos objetivos desta dissertação é ilustrar a dificuldade em prever resultados esportivos devido a presença massiva da sorte em alguns esportes ou campeonatos. Portanto, é de extrema importância compreender como trabalhos anteriores lidaram com a sorte e as abordagens por eles utilizadas. Alguns destes trabalhos tentam predizer qual equipe irá marcar o próximo ponto na partida [Vračar et al., 2016; Peel & Clauset, 2015], enquanto outros trabalhos focam na tarefa de predizer quem irá vencer a partida [Gabel & Redner, 2012; Merritt & Clauset, 2014; Ben-Naim et al., 2006; Chen & Joachims, 2016; Miljković et al., 2010].

Considerando a tarefa de prever qual equipe marcará o próximo ponto, Vračar

et al. [2016] propôs um engenhoso modelo baseado em processos de Markov acoplados com uma regressão logística multinomial para predizer cada ponto consecutivo de uma partida de basquete. Nesse trabalho, foram incluídas um grande número de variáveis explicativas que caracterizam a evolução da partida, como por exemplo, o tempo da partida, a diferença de pontos no momento estudado e as características da equipe adversária.

Gabel & Redner [2012] e Merritt & Clauset [2014] trabalharam com modelos estocásticos extremamente simples mas que ajustam-se aos dados empíricos muito bem. O comportamento padrão utilizado é que os eventos ocorrem aleatoriamente, de acordo com um processo de Poisson homogêneo com uma taxa específica para cada esporte. Dado um evento, seus pontos são atribuídos a uma das equipes através do lançamento de uma moeda viciada específica para cada esporte (um processo de Bernoulli). De acordo com Peel & Clauset [2015], os modelos mais bem sucedidos normalmente calculam uma probabilidade em torno de 0,65 para uma determinada equipe marcar o próximo ponto de acordo com características do modelo. Para prever qual equipe será a próxima a marcar, Peel & Clauset [2015] estudaram a influência da vantagem de pontos do vencedor (se uma força restaurativa está presente) e tentaram identificar a tendência do último time a marcar continuar a marcar (anti-persistência). Por fim, eles fazem uma extrapolação das sequências de pontos para determinar qual equipe irá vencer a partida. Eles testaram suas hipóteses em três diferentes esportes: futebol americano, hóquei e basquete, e obtiveram uma acurácia de aproximadamente 80% na predição do vencedor da partida.

Para prever o resultado de partidas esportivas, em Ben-Naim et al. [2006] foi ajustado um modelo teórico simples para dados empíricos com o objetivo de estimar a probabilidade q da pior equipe vencer. Embora simples, este modelo ajusta-se muito bem aos dados. Através desse trabalho os autores concluíram que a probabilidade da pior equipe vencer no futebol e basebol é $q \approx 0.45$, e para o basquete e o futebol americano é $q \approx 0.35$. Esses números dão uma ideia do nível de aleatoriedade presente em cada esporte. Em Miljković et al. [2010] os autores modelaram o problema de predizer o resultado de uma partida como uma tarefa de classificação e utilizaram *Naive Bayes* para fazer as predições. A base de dados utilizada foi uma temporada da NBA e eles conseguiram predizer

corretamente 67% dos resultados. Modelos gráficos probabilísticos também já foram utilizados para predizer o resultado de uma partida esportiva. Chen & Joachims [2016] apresentaram um cenário probabilístico para predizer o resultado de partidas considerando características em que a partida ocorre, como o tempo e o tipo de quadra de tênis.

De forma geral, prever qual equipe marcará o próximo ponto é uma tarefa muito complexa devido a aleatoriedade presente no processo e é mais difícil que prever o resultado de uma partida. No basquete, por exemplo, detalhes no passe da bola, uma tentativa rápida de jogada ensaiada, um bloqueio no último instante são exemplos de situações que influenciam em quem irá marcar o próximo ponto, ocorrendo a todo momento ao longo do jogo e são quase impossíveis de serem modeladas. Entretanto, prever o resultado da partida é uma tarefa um pouco mais simples, embora possa apresentar resultados igualmente surpreendentes. Ao longo de uma longa sequência de pontos, como ocorre no basquete, é esperado que a equipe mais habilidosa marque mais pontos e vença a partida. Portanto, embora uma parcela dos pontos disputados na partida sejam distribuídos aleatoriamente entre as duas equipes, grande parte dos pontos são ganhos devido as características das equipes.

2.2 Caracterização de Ligas Esportivas

Predizer o ranking final de uma temporada de uma liga esportiva qualquer é uma tarefa diferente da predição do resultado de partidas, mas é igualmente difícil de obter uma boa performance. Além dos fatores presentes na partida que podem influenciar esta tarefa, existem também efeitos devido ao design da competição que devem ser considerados [Chan et al., 2009; Ben-Naim et al., 2013, 2007].

O grande interesse da indústria e da academia em esportes tem estimulado o aparecimento de muitos artigos sobre os movimentos táticos durante os jogos de diferentes esportes. Wang et al. [2015] descobriu o melhor padrão de táticas em partidas de futebol através da mineração de dados históricos de partidas de futebol. Em Brooks et al. [2016] foi proposto um sistema de ranqueamento de jogadores de futebol baseado no valor dos passes feitos ao longo das partidas em uma temporada/liga. Técnicas de mineração de dados foram utilizadas por Van Haaren

et al. [2016] para descobrir padrões relacionados a aspectos espaciais e temporais dos jogos de voleibol para orientar o desempenho dos jogadores. Vaz de Melo et al. [2012] propôs um modelo de rede de conexões entre os jogadores e extraiu variáveis que foram usadas para modelar o desempenho das equipes. A vantagem obtida por ser a equipe mandante na NBA foi estudada em Ribeiro et al. [2016]. Neste trabalho os autores observaram, por exemplo, que a taxa de pontuação da equipe *Cleveland Cavaliers* aumenta em aproximadamente 0.16 pontos por minuto quando ele joga em casa, enquanto a taxa de pontuação do *New Jersey Nets* aumenta somente aproximadamente 0.04 pontos por minuto quando disputa uma partida como mandante. Todas essas análises citadas podem ser utilizadas para auxiliar técnicos e dirigentes esportivos no treinamento e na formação de equipes mais competitivas.

Predizer o rank final de uma competição ou qual será a equipe vencedora depende do formato em que o campeonato é disputado e do esporte. Em Chetrite et al. [2015], o número de potenciais vencedores em competições é estudado considerando a forma com que suas habilidades estão distribuídas. Um ranqueamento para as equipes da NBA foi desenvolvido por Pelechrinis et al. [2016]. Neste trabalho foi utilizado o algoritmo *PageRank* para criar uma rede entre as equipes e a acurácia obtida foi de aproximadamente 67%. Tarlow et al. [2014] desenvolveram um modelo gráfico probabilístico para prever as habilidades de equipes da liga *NCAA* de futebol americano. Note que todos esses trabalhos consideram que existe uma diferença entre as habilidades das equipes que disputam o campeonato e procuram formas de estimar essa habilidade para obter um ranqueamento das equipes, estimar o possível campeão ou um conjunto de possíveis equipes vencedoras. Entretanto, nem sempre existe uma diferença significativa entre as habilidades das equipes. Spiegelhalter [2007] estudou como medir a sorte e a habilidade em uma temporada/liga de futebol através da comparação da variância amostral com a variância esperada quando todas as equipes possuem a mesma habilidade. Diferente deste trabalho, nesta dissertação é proposto um coeficiente que permite comparar a influência da sorte entre temporadas e esportes.

Esta dissertação gerou duas publicações: um artigo no Simpósio Brasileiro de Banco de Dados (SBBD) e um artigo submetido para o *Knowledge Discovery in Databases* (KDD). Em Aoki et al. [2016] apresentado no SBBD, propõe-se um

coeficiente para medir a influência da sorte e da habilidade em resultados de ligas de futebol. A partir desse trabalho e de Spiegelhalter [2007] é possível ter uma ideia se o resultado de um campeonato é puramente aleatório, devido a grande similaridade entre as habilidades das equipes, ou se o resultado do campeonato de fato reflete as diferentes habilidades das equipes. No artigo submetido para o KDD é feita uma extensão deste coeficiente para outros esportes e é apresentado um modelo para estimar as habilidades das equipes.

Capítulo 3

Separando a sorte da habilidade

Neste Capítulo 3 será descrita a metodologia utilizada para separar a sorte da habilidade e para estimar a habilidade quando esta possuir influência nos resultados da temporada/liga. A Seção 3.1 descreve como foi obtido o coeficiente proposto ϕ que mede a influência da sorte em um campeonato. Um modelo para estimar as habilidades das equipes de um campeonato é proposto na Seção 3.2.

3.1 Um coeficiente de medida de sorte e habilidade

O objetivo desta seção é definir uma métrica que permita avaliar se as equipes de uma temporada/liga possuem diferentes habilidades ou se o rank final de um campeonato é determinado aleatoriamente. Entretanto, faz-se a ressalva de que os campeonatos determinados aleatoriamente que são considerados neste trabalho dependem do contexto em que o jogo ocorre. Em jogos de futebol, por exemplo, a equipe mandante usualmente possui uma chance maior de vencer a partida. Isso se deve principalmente à presença de uma torcida favorável, e não necessariamente a uma habilidade intrínseca que o time talvez possua. Uma análise empírica da quantidade de vitórias do time da casa, do visitante e de empates em nove temporadas sucessivas do principal campeonato brasileiro de futebol, denominado Série A, mostrou que a probabilidade da equipe visitante vencer é metade da proba-

bilidade do time da casa vencer (0.25 contra 0.50, respectivamente), com uma pequena variação ao longo do tempo. Essa análise empírica foi realizada com base nos dados coletados para o estudo do coeficiente de habilidade aqui proposto. Outras variáveis contextuais podem influenciar uma partida. Em Chen & Joachims [2016] foram consideradas variáveis contextuais que influenciam o desempenho dos jogadores de tênis, como o tipo de campo da disputa: se é *indoor* ou *outdoor*, qual tipo de quadra está sendo utilizada (grama sintética, cimento, saibro ou relva). O coeficiente desenvolvido nesta dissertação considera somente o contexto mandante/visitante pela falta de informações adicionais, mas pode ser facilmente estendido se mais informações sobre o contexto em que os jogos ocorram estiverem disponíveis.

Os campeonatos considerados nesta etapa do trabalho seguem o modelo de campeonatos de pontos corridos. Essa limitação é imposta pela forma em que é criado o modelo aleatório. Neste tipo de campeonato cada equipe joga com todas as demais equipes da liga pelo menos duas vezes: metade como mandante e a outra metade como visitante. Além disso, todas as equipes disputam o mesmo número de partidas.

Seja X_h uma variável aleatória que representa os pontos ganhos por uma equipe quando joga uma partida do campeonato como mandante. Similarmente, seja X_a outra variável aleatória associada aos pontos ganhos quando a equipe joga uma partida como visitante. A distribuição de probabilidade de X_h e X_a é específica para cada esporte. No basquete, por exemplo, não existem empates e cada partida resulta em 1 ponto para um dos dois times, como mostra a Tabela 3.1. Consequentemente, X_h e X_a são simplesmente variáveis aleatórias Bernoulli. No caso do futebol e do handebol, a situação é diferente: a equipe ganha 3 ou 1 ou 0 pontos dependendo do número de gols da equipe ser maior, igual ou menor que o da equipe adversária, respectivamente (veja Tabela 3.2). No voleibol, como pode ser visto na Tabela 3.3, a pontuação depende da quantidade de sets marcados por cada equipe na partida. Caso a vitória seja por 2 ou mais sets de diferença (3×0 ou 3×1), a equipe vitoriosa ganha 3 pontos e a outra 0 pontos. Mas se a vitória for por somente 1 set de diferença (3×2), a equipe vitoriosa ganha 2 pontos e a equipe derrotada ganha 1 ponto.

Considerando como exemplo o cenário dos jogos de handebol e futebol mos-

Tabela 3.1. Resultados Possíveis - Basquete

Probabilidade	Pontos		Resultados
	Mandante	Visitante	
P_h	1	0	Mandante ganha
P_a	0	1	Visitante ganha

Tabela 3.2. Resultados Possíveis - Handebol e Futebol

Probabilidade	Pontos		Resultados
	Mandante	Visitante	
P_h	3	0	Mandante ganha
P_t	1	1	Empate
P_a	0	3	Visitante ganha

Tabela 3.3. Resultados Possíveis - Voleibol

Probabilidade	Pontos		Resultados
	Mandante	Visitante	
P_{h1}	3	0	3×0 ou 3×1
P_{h2}	2	1	3×2
P_{a1}	1	2	2×3
P_{a2}	0	3	0×3 ou 1×3

trado na Tabela 3.2, seja P_h, P_t e P_a as probabilidades da equipe mandante vencer, empatar e perder uma partida, respectivamente. Essas probabilidades são números não-negativos e $P_h + P_t + P_a = 1$. Consequentemente, a variável aleatória X_h possui distribuição multinomial com três possíveis valores que são 3, 1 e 0 com probabilidades P_h, P_t e P_a, respectivamente. Como resultado, seu valor esperado $E(X_h)$ é igual a $\mu_{X_h} = P_h \times 3 + P_t \times 1 + P_a \times 0$ e a variância é $\sigma^2_{X_h} = P_h \times 3^2 + P_t \times 1^2 + P_a \times 0^2 - \mu^2_{X_h}$. Similarmente, X_a tem média $\mu_{X_a} = P_a \times 3 + P_t \times 1 + P_h \times 0$ e variância $\sigma^2_{X_a} = P_a \times 3^2 + P_t \times 1^2 + P_h \times 0^2 - \mu^2_{X_a}$.

Essas probabilidades são calculadas para todas as temporadas de todas as ligas estudadas. A Figura 3.1 mostra para cada esporte uma liga como exemplo. Dessa forma é possível observar o comportamento dessas probabilidades. No basquete, exemplificado pela liga americana NBA, observa-se que a probabilidade da equipe mandante vencer é aproximadamente 0.6. O campeonato brasileiro Série A foi utilizado para mostrar as probabilidades dos resultados possíveis no fute-

bol. A figura mostra que a probabilidade da equipe mandante vencer no futebol é aproximadamente 0.5, e as probabilidades de empate e da equipe visitante vencer são ambas aproximadamente 0.25. O handebol foi ilustrado com o campeonato de handebol feminino da Suíça. Diferentemente do futebol que possui o mesmo conjunto de possíveis resultados, no handebol a probabilidade de empate é muito pequena, mas a probabilidade da equipe mandante vencer ainda é de aproximadamente 0.5, isto é, não varia. Por fim, foi utilizado o campeonato sérvio de voleibol masculino para mostrar o comportamento dessas probabilidades no voleibol. O resultado mais provável no volei é o da equipe mandante vencer por 2 ou mais sets de diferença, seguido pela equipe visitante vencer por 2 ou mais sets de diferença. Embora os gráficos das outras ligas não tenham sido mostrados, essas probabilidades tendem a se manter constantes em cada esporte.

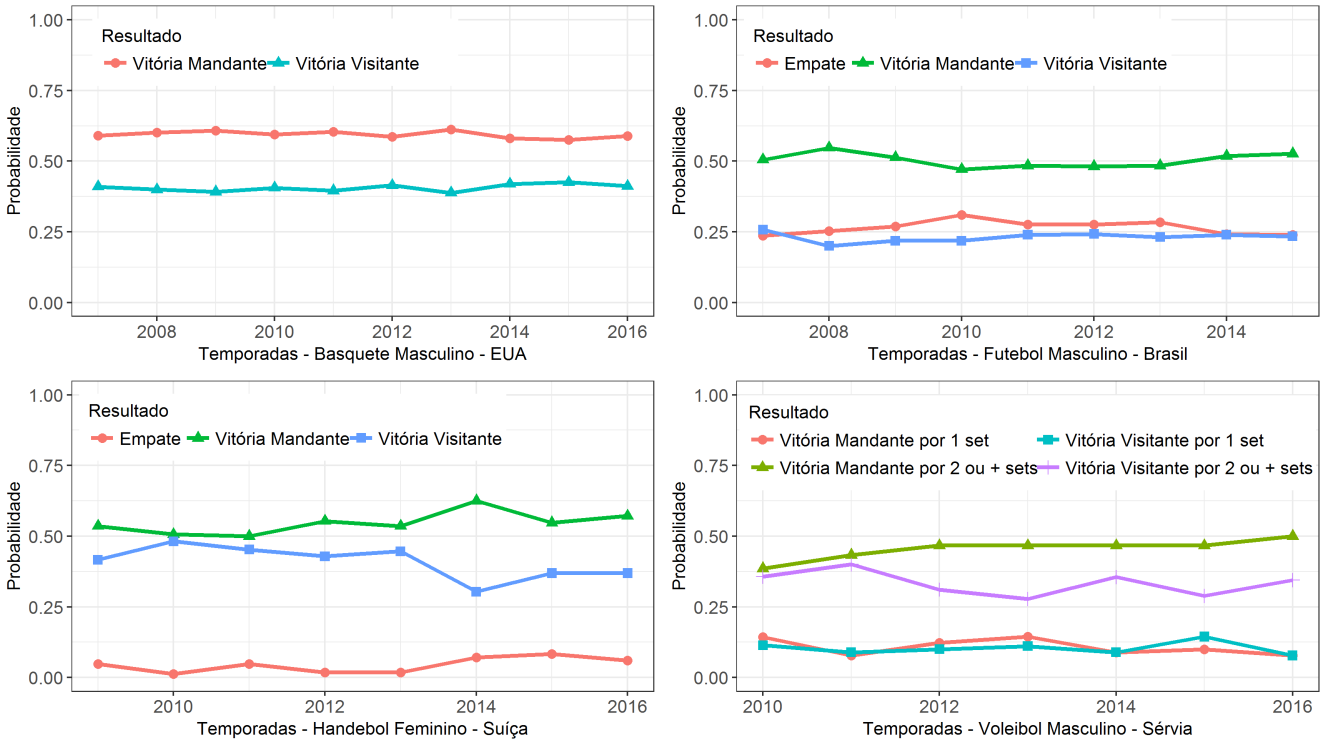

Figura 3.1. Proporção de resultados possíveis de sucessivas temporadas dos jogos de basquete, futebol, handebol e voleibol.

Em um campeonato de pontos corridos padrão composto por $k+1$ equipes,

cada um dos times joga $2k$ vezes: k partidas contra todas as demais equipes como mandante e k partidas como visitante. Seja X_{hi} e X_{ai} os pontos obtidos por uma equipe no seu i-ésimo jogo como mandante e no i-ésimo jogo fora de casa, respectivamente. Então:

$$\mathrm{Y}_{2k} = \sum_i (X_{hi} + X_{ai})$$

é uma variável aleatória que representa o total de pontos ganhos por uma equipe ao final do campeonato. Considerando novamente os resultados possíveis do futebol e do handebol como exemplo, assumir que as probabilidades P_h, P_t e P_a são as mesmas para todas as equipes de uma temporada/liga é o mesmo que assumir que a habilidade não possui um papel importante no campeonato. Qualquer pontuação final é o resultado de pura sorte, medida unicamente pelo contexto em que o jogo ocorreu. Um campeonato de pontos corridos padrão em que cada equipe disputa k jogos como mandante, k como visitante e todas as equipes envolvidas disputam a mesma quantidade de jogos garante que a distribuição das pontuação final Y_{2k} é a mesma para todas as equipes. Considerando independência estocástica entre os jogos e uma quantidade suficientemente grande de k jogos no torneio, tem-se que a distribuição final dos pontos de um campeonato aleatório segue aproximadamente uma distribuição Gaussiana:

$$Y_{2k} \sim N(\mu_{2k}, \sigma^2_{2k}) \tag{3.1}$$

em que $\mu_{2k} = k(\mu_{X_h} + \mu_{X_a})$ e $\sigma^2_{2k} = k(\sigma^2_{X_h} + \sigma^2_{X_a})$.

Em resumo, se todas as equipes possuíssem as mesmas habilidades, depois de cada equipe disputar k jogos como mandante e k jogos como visitante, a distribuição final dos pontos do campeonato Y_{2k} seria aproximadamente uma distribuição normal com média μ_{2k} e variância σ^2_{2k}. Este é o *baseline* utilizado nesse trabalho, contra o qual os resultados empíricos observados serão contrastados. Como pode ser observado, a construção da distribuição depende do esporte que está sendo analisado, do número de jogos na temporada e da quantidade de equipes. Portanto a distribuição de Y_{2k} é diferente para cada temporada/liga.

As probabilidades de ocorrência da cada evento são estimadas através da frequência simples de suas ocorrências ao longo do campeonato. No futebol, por

exemplo, a probabilidade P_h é estimado pela quantidade de jogos em que a equipe mandante venceu na temporada/liga, representado por W_h, dividida pelo número total N de jogos ocorridos no campeonato. Portanto, $P_h \hat{=} \frac{W_h}{N}$. Analogamente, as probabilidades de empates e de vitória da equipe visitante são estimadas respectivamente por $P_t \hat{=} \frac{W_t}{N}$ e $P_a \hat{=} \frac{W_a}{N}$, em que W_t e W_a representam a quantidade de empates e de vitórias da equipe visitante na temporada, respectivamente.

Considerando uma única temporada/liga de um dado esporte, é possível fazer uma comparação direta entre a variância teórica σ_{2k}^2 e variância S^2, que representa a variância observada da distribuição dos pontos finais de um campeonato. Entretanto, essa comparação não é válida para diferentes temporadas ou esportes. Para solucionar este problema e poder fazer comparações da influência da sorte nos resultados de diferentes temporadas, ligas e esportes, propõe-se o coeficiente ϕ:

$$\phi = \frac{S^2 - \sigma_{2k}^2}{S^2} \tag{3.2}$$

O coeficiente ϕ tem valores possíveis no intervalo $(-\infty, 1]$ e representa o quão distante está a variância do modelo aleatório em relação à variância observada. A interpretação dessa medida na escala é feita da seguinte maneira: valores positivos do coeficiente ϕ indicam um excesso de variabilidade na pontuação final do campeonato devido à variância adicional induzida pelas diferentes habilidades das equipes. Quanto mais próximo de 1, mais influente é o fator habilidade no campeonato. Valores do coeficiente ϕ em torno de 0 representam os campeonatos cujo fator habilidade possui somente uma pequena influência nos resultados. Valores menores que zero indicam que as pontuações das equipes possuem uma variabilidade menor que a esperada devido ao fator sorte. A presença de valores negativos parece surpreendente mas é uma possibilidade, de fato, observada nos dados. Essa situação pode ocorrer, por exemplo, devido a algum tipo de mecanismo compensatório ao longo da temporada ou algum conluio entre es equipes. Embora esta não seja uma situação comum, foram observados alguns casos com ϕ negativo na base de dados estudada.

O coeficiente ϕ é uma medida empírica baseada nos resultados da temporada. Mesmo que o modelo aleatório seja verdadeiro, é impossível obter um valor ϕ

exatamente igual a 0. Logo, é necessário obter uma medida de incerteza em torno do valor zero que viabilize a construção de um intervalo de confiança com o objetivo de medir a distância entre os dados observados e o modelo aleatório.

O intervalo de confiança para o coeficiente ϕ é construído usando simulações de Monte Carlo sob a hipótese do modelo aleatório ser o verdadeiro. No caso do futebol, as probabilidades estimadas P_h, P_t e P_a são utilizadas para gerar amostras Multinomiais que simulam o resultado de partidas entre as equipes de um campeonato. A simulação das partidas segue a mesma ordem observada durante a temporada real do campeonato. Dessa forma, são simulados os resultados dos N jogos disputados na temporada e a variância da distribuição final dos pontos desse conjunto de jogos é calculada. Repetindo esse processo de forma independente, obtêm-se um intervalo Monte Carlo com 95% de confiança em torno de ϕ para o contexto em que não existe diferença de habilidade entre as equipes, isto é, o campeonato é definido de acordo com o modelo aleatório. Esse intervalo de confiança é criado individualmente, ou seja, ele varia entre as temporadas, campeonatos e esportes.

Através desse intervalo de confiança é possível julgar um valor do coeficiente ϕ obtido a partir de dados reais. Se o valor observado de ϕ está dentro do intervalo, então ele não é significativamente diferente de 0. Isso significa que não existem evidências de que a temporada/liga desvia-se do modelo aleatório. Em contraste, um valor do coeficiente ϕ acima do limite superior do intervalo é um forte indicativo de que as habilidades das equipes é diferente, enquanto valores do coeficiente ϕ abaixo do limite inferior indicam que a temporada possui significativamente menos variabilidade que o modelo aleatório. Considerando essas temporadas em que o valor de ϕ foram significativamente diferentes de 0 e positivas, uma questão de interesse é: quantas equipes devem ser removidas da temporada/liga para que ela se torne aleatória? Para responder essa pergunta, foi desenvolvida uma segunda simulação que funciona da seguinte forma:

1. As equipes da temporada/liga são ordenadas de acordo com suas pontuações ao final do campeonato;

2. A equipe com a pontuação mais distante da média é definida como equipe X. Essa equipe X pode ser a melhor ou a pior do conjunto de times avaliados.

No caso de empates entre duas ou mais equipes:

a) entre o melhor time e o pior, X é escolhido como o melhor time;

b) entre os dois primeiros colocados ou os dois últimos colocados, X é definido usando a ordem alfabética dos nomes das equipes.

3. Após definir a equipe X, remove-se todos os jogos que envolvem esta equipe na temporada;

4. A pontuação final do campeonato é recalculado sem X e seus jogos;

5. O coeficiente ϕ é recalculado considerando esse novo subconjunto de jogos:

a) Se o novo coeficiente ϕ é significativamente diferente de 0, o algoritmo retorna em 1 com o subconjunto de jogos que exclui as partidas disputadas por X;

b) Se o novo coeficiente ϕ é significativamente igual a 0, o algoritmo é finalizado.

3.2 Estimação das habilidades

Nessa seção é apresentado um modelo gráfico probabilístico para estimar as habilidades das equipes quando seus coeficientes ϕ são positivos e significativamente diferentes de 0. Mais importante do que simplesmente aprender as diferentes habilidades de cada equipe, esse modelo será capaz de fatorar essa habilidade em variáveis explanatórias, isto é, identificar quais fatores são mais importantes para explicar as diferentes habilidades das equipes do campeonato.

Para ajustar esse modelo bayesiano, será utilizado o resultado de todas as partidas de uma temporada. O score final desses jogos serão os dados que Função de Verossimilhança irá receber. Para explicar essa pontuação final entre as duas equipes, serão utilizadas características das equipes, tais como salário médio dos jogadores e quantidade de jogadores de cada equipe. Cada uma dessas características terão um peso diferente, que juntos em um modelo log-linear definem a habilidade da equipe. Inicialmente, esses pesos serão determinados por uma distribuição *a priori*, mas através da distribuição *a posteriori*, os valores dos pesos

serão atualizados. A distribuição *a posteriori* é proporcional a informação que se conhece de antemão do problema e a informação adicionada pelos dados. Devido a modelagem adotada, será necessário utilizar um método numérico para obter uma amostra de dados dessa distribuição *a posteriori*.

Nas próxima Subseções, o modelo será construído a partir da habilidade das equipes, passando pelo modelo log-linear, distribuição de verossimilhança, distribuição *a priori* dos pesos e a inferência a partir de métodos numéricos.

3.2.1 Definição de Habilidade

Um famoso modelo de comparação entre entidades que estão em uma competição é o modelo de Bradley-Terry. Esse modelo assume que existem quantidades positivas $\alpha_1, \alpha_2, ..., \alpha_n$ que representam as habilidades das n equipes. Dessa forma, a probabilidade de uma equipe i vencer uma partida contra a equipe j é dada pelo tamanho relativo de α_i com respeito ao tamanho de α_j:

$$\pi_{ij} = P(i \text{ vence } j) = \frac{\alpha_i}{\alpha_i + \alpha_j}$$

Os valores positivos dos parâmetros α_i são tomados de uma escala irrestrita através da transformada $\alpha_i = \exp(\beta_i)$. Logo, a formulação de Bradley-Terry implica um conveniente modelo de regressão logística:

$$\log\Big(\frac{p_{ij}}{1-p_{ij}}\Big) = \beta_i - \beta_j$$

A conveniência dessa formulação é a possibilidade de expansão do modelo de Bradley-Terry para incorporar variáveis explicativas que talvez ajudem a estimar e explicar as diferenças entre as α_i's habilidades.

Seja x_i um vetor de dimensão d composto por um conjunto de características da equipe i que podem explicar a habilidade associada α_i. O modelo que será ajustado irá aprender a relevância da presença de cada uma dessas características de x_i na determinação da habilidade α_i. O primeiro passo para o ajuste do modelo é especificar a função de verossimilhança para os dados observados, como mostra a Subseção 3.2.2.

3.2.2 Função de Verossimilhança

Os dados utilizados para fazer o ajuste do modelo é o conjunto de jogos de uma temporada e a pontuação observada ao final de cada partida. Considera-se a pontuação final da partida e não somente o resultado final (vitória, derrota ou empate) pois acredita-se que a diferença de pontos marcados ao longo da partida está correlacionada com as habilidades das equipes envolvidas. Por exemplo, quando uma equipe com um grande valor de habilidade α_i joga contra outra equipe com um pequeno valor de habilidade α_j, a probabilidade da equipe i vencer deveria não somente ser grande mas a diferença de pontos marcados entre as duas equipes também deveria ser grande.

Em um torneio com K jogos e n equipes com habilidades $\alpha_1, \alpha_2, ..., \alpha_n$, seja N_k e S_k o total de pontos marcados pelas duas equipes e pela equipe mandante no k-ésimo jogo da temporada. Se cada um dos N_k pontos marcados nas partidas forem considerados resultados de uma sequência de sucessos e fracassos do ponto de vista da equipe mandante, a variável aleatória S_k terá uma distribuição binomial condicionada em N_k e nas habilidades das equipes que disputam a partida. Como o número de pontos marcados em uma partida pode ser muito grande (como em jogos de basquete), a consequência dessa abordagem é uma verossimilhança computacionalmente intratável. Para evitar esse problema, a distribuição binomial foi aproximada por uma distribuição de Poisson com média esperada $N_k \times \frac{\alpha_i}{\alpha_i+\alpha_j}$. Para considerar a influência da sorte nos resultados das partidas, foi adicionado um efeito aleatório ε_k em cada partida.

Em resumo, condicionado nos parâmetros e no efeito aleatório, o total de pontos marcados pela equipe mandante em cada partida k possui a seguinte distribuição:

$$S_k \sim \text{Poisson}\left(N_k \times \frac{\alpha_{h(k)}}{\alpha_{h(k)} + \alpha_{a(k)}} + \varepsilon_k\right) \tag{3.3}$$

em que $h(k)$ e $a(k)$ são os índices das equipes mandantes e visitantes que disputaram a k-ésima partida. A presença do efeito aleatório ε_k é essencial. Sem ele, a variabilidade dos resultados induzidos pela distribuição de Poisson seria completamente determinada pela habilidade relativa das equipes que é definida por α. Entretanto, é necessário adicionar um componente de super-dispersão, o efeito

aleatório ε_k, para dar conta da grande variação dos resultados dos jogos.

As habilidades $\alpha_1, \alpha_2, ..., \alpha_n$ são totalmente determinadas pelas habilidades intrínsecas de cada equipe. Utiliza-se a função de ligação canônica associada à distribuição de Poisson para obter um modelo log-linear.

$$\log(\alpha_i) = \mathbf{w}^T \mathbf{x}_i$$

em que $\mathbf{w}$ representa os pesos de cada uma das variáveis explicativas do modelo e $\mathbf{x}$ possui o conjunto de características da equipe i.

Seja $\mathcal{D} = \{\mathbf{x}, S_k, N_k, \forall k = 1, ..., K\}$ os dados observados de uma temporada. Os parâmetros da função de verossimilhança $\mathbf{w} \in R^d$ e $\varepsilon_1, ..., \varepsilon_K$, para todo os $k = 1, .., K$ jogos ocorridos armazenados em $\mathcal{D}$ são dados por:

$$L(\mathcal{D}|\mathbf{w}, \varepsilon_k) \approx \prod_{k=1}^{K} \left(N_k \frac{\alpha_{h(k)}}{\alpha_{h(k)} + \alpha_{a(k)}} + \varepsilon_k \right)^{S_k} \times \exp\left(-N_k \frac{\alpha_{h(k)}}{\alpha_{h(k)} + \alpha_{a(k)}} + \varepsilon_k \right) \quad (3.4)$$

3.2.3 Distribuições a Priori

A distribuição a *priori* sobre os pesos de $\mathbf{w}$ que será adotada é uma distribuição multivariada Gaussiana $p(\mathbf{w}) \sim \mathcal{N}(\mathbf{w}; \mathbf{0}, 2\mathbf{I})$. Para os efeitos aleatórios $\varepsilon_1, \ldots, \varepsilon_K$ serão utilizadas distribuições a *priori* independentes, com distribuição Gaussiana e média 0. Consequentemente, o vetor n-dimensional $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ também segue uma distribuição multivariada Gaussiana condicionada a um parâmetro de precisão: $p(\varepsilon) \sim \mathcal{N}(\varepsilon; \mathbf{0}, 3\mathbf{I})$.

A Figura 3.2 mostra a representação gráfica do modelo Bayesiano utilizado neste trabalho. Observa-se que as habilidades estimadas e os dados $\mathbf{x}$ variam entre 1 e n, em que n é a quantidade de equipes do campeonato. Já a pontuação da equipe mandante S_k varia entre a quantidade de jogos da temporada. A habilidade da equipe mandante relativa à equipe visitante é calculada para cada um dos K jogos.

3.2.4 Inferência Bayesiana via algoritmos MCMC

A inferência Bayesiana é baseada na distribuição a *posteriori* dos pesos dos coeficientes $\mathbf{w}$ e do efeito aleatório ε. A distribuição a *posteriori* é proporcional

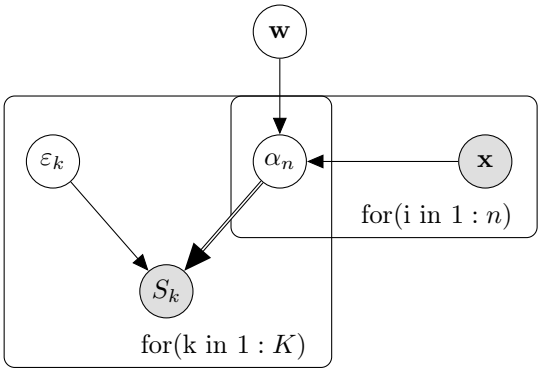

Figura 3.2. Modelo Gráfico Probabilístico do modelo Bayesiano que estima as habilidades das equipes.

ao produto entre a função de verossimilhança e a distribuição a *priori*:

$$p(\mathbf{w}, \varepsilon|\mathcal{D}) \propto \prod_{k=1}^{K} \left(N_k \frac{\alpha_{h(k)}}{\alpha_{h(k)}+\alpha_{a(k)}} + \varepsilon_k \right)^{S_k}$$

$$\times \exp\left(-N_k \frac{\alpha_{h(k)}}{\alpha_{h(k)}+\alpha_{a(k)}} + \varepsilon_k \right) \exp\left(-\frac{1}{2\sigma_w^2} \mathbf{w}^T \mathbf{w} \right) \times$$

$$\exp\left(-\frac{1}{2\sigma_\varepsilon^2} \varepsilon^T \varepsilon \right)$$

note que são eliminados todos os fatores multiplicativos que não envolvem os parâmetros desconhecidos. Além disso, após alguns testes e simulações, definiu-se $\sigma_w^2 = 2$ e $\sigma_\varepsilon^2 = 9$. O Algoritmo *Metropolis-Hastings*, que faz parte dos métodos *Markov chain Monte Carlo* (MCMC), é utilizado para obter uma amostra aleatória da distribuição a *posteriori*. A partir dessa amostra é possível fazer inferências sobre os coeficientes de $\mathbf{w}$ e o efeito aleatório. Uma abordagem sequencial deste

algoritmo é adotada, como mostra o Algoritmo 1:

Algoritmo 1: Metropolis-Hastings

Entrada: $\mathbf{w}_{inicial}$,K
Saída: $\mathbf{W}_{matriz}$

```
W_matriz = NULL
início
    w_corrente = w_inicial
    para cada i ∈ 1000 faça
        para cada j ∈ K faça
            w_proposto = RandomNormal(w_corrente, desvio padrão = 0.15)
            se w_proposto é aceito então
                w_corrente = w_proposto
            fim
        fim
        W_matriz[i] = w_corrente
    fim
fim
retorna W_matriz
```

No algoritmo de Metropolis-Hasting, para cada elemento proposto, avalia-se o critério de aceitação. Um elemento proposto é aceito com probabilidade igual ao valor mínimo do conjunto $\left\{1, \frac{\pi(\theta^*)\times q(\theta|\theta^*)}{\pi(\theta)\times q(\theta^*|\theta)}\right\}$, em que θ é o valor corrente do algoritmo, θ^* o novo valor proposto, $q(|)$ é a distribuição proposta e $\pi()$ é a distribuição da qual deseja-se obter uma amostra aleatória, no caso, a distribuição *a posteriori*. Devido ao fato da distribuição proposta ser uma Gaussiana, tem-se $q(\theta|\theta^*) = q(\theta^*|\theta)$ em consequência da simetria da distribuição Normal. Dessa forma, a probabilidade de um novo elemento ser aceito passa a ser o valor mínimo do conjunto $\left\{1, \frac{\pi(\theta^*)}{\pi(\theta)}\right\}$, que é um cálculo facilmente obtido.

3.2.5 Seleção do modelo

Vários modelos que diferiam pela inclusão ou exclusão de variáveis explicativas foram avaliados. Para selecionar o modelo que melhor se ajustava aos dados, fez-se uma seleção de modelos utilizando o *deviance information criterion* (DIC),

proposto por Spiegelhalter et al. [2002]. Esse critério é o mais utilizado para avaliar a qualidade de modelos bayesianos.

Considerando que a função de verossimilhança dos dados é $P(y|\theta)$ e o *deviance* igual a $D(\theta) = -2\log\{P(y|\theta)\}$, o DIC é definido como:

$$DIC = D(\overline{\theta}) + 2p_D \tag{3.5}$$

em que $p_D = \overline{D(\theta)} - D(\overline{\theta})$. O primeiro termo $\overline{D(\theta)}$ representa a média *a posteriori* do *deviance* e o segundo termo $D(\overline{\theta})$ representa o *deviance* da média *a posteriori* de θ. Esse critério de seleção de modelos leva em consideração a complexidade do modelo através do termo p_D e o ajuste do modelo aos dados com $D(\overline{\theta})$. A Equação 3.5 pode ser reescrita como $DIC = 2 \times \overline{D(\theta)} - D(\overline{\theta})$.

Considerando o modelo ajustado para estimar as habilidades α, o *deviance* $D(\alpha)$ será definido como:

$$D(\alpha) = -2\left\{\sum_{k=1}^{K} s_{ik} log\left(N_k \times \frac{\alpha_i}{\alpha_i + \alpha_j}\right) - \sum_{k=1}^{K}\left(N_k \times \frac{\alpha_i}{\alpha_i + \alpha_j}\right) - \sum_{k=1}^{K}\log(s_{ik}!)\right\} \tag{3.6}$$

em que K é a quantidade de jogos na temporada, s_{ik} são os pontos marcados pelo time da casa i no k-ésimo jogo da temporada, N_k é o total de pontos marcados pelas duas equipes na k-ésima partida, α_i e α_j são as habilidades estimadas da equipe mandante e da equipe visitante respectivamente.

Um modelo é preferível quando seu DIC é o menor dentre o conjunto de modelos avaliados. O valor do DIC pode ser negativo e quando isso ocorre, seu sinal deve ser levado em consideração na escolha do modelo com o menor DIC.

3.3 Programas

Para o download das bases de dados, cálculo do coeficiente ϕ, desenvolvimento do modelo e gráficos, foi utilizado o software livre R e os pacotes *ggplot2*, *RColorBrewer*, *gridExtra*, *XML*, *xtable*, *Rcpp* e *far*.

A linguagem C++ e a biblioteca *Rcpp* foram utilizadas na etapa do Metropolis-Hastings explicada na Seção 3.2.

Capítulo 4

Resultados

Esse Capítulo apresenta as duas bases de dados utilizadas, como foram coletadas e faz uma breve análise descritiva de seus conteúdos. Além disso, são mostrados os resultados empíricos do coeficiente proposto ϕ e um ajuste em dados reais do modelo gráfico probabilístico.

4.1 Base de Dados

Para o desenvolvimento desse trabalho foram coletadas duas bases de dados: uma das bases é utilizada para comparar a mistura de sorte e habilidade em diferentes esportes e ligas; a outra base é usada para ajustar o modelo gráfico probabilístico que estima as habilidades das equipes de uma temporada/liga. Essa segunda base possui informações das equipes que são usadas como variáveis explicativas.

A primeira base de dados utilizada para calcular o coeficiente ϕ abrange quatro esportes: o basquete, voleibol, futebol e handebol. Foram selecionadas para compor essa base as ligas esportivas com mais de 5 temporadas, que seguem o modelo de campeonato de pontos corridos e possui mais de 7 equipes em cada temporada. Selecionou-se as temporadas completas que seguem esses pré-requisitos ocorridas entre Janeiro de 2007 e Julho de 2016. Todas essas informações foram coletadas no site `www.betexplorer.com`. A base completa possui 1503 temporadas e 270713 jogos. As ligas estão distribuídas por 84 países da África, América, Ásia, Europa e Oceania. A Figura 4.1 mostra quais países possuem ligas representadas

neste estudo. A intensidade da cor azul aumenta de acordo com a quantidade de ligas que o país possui.

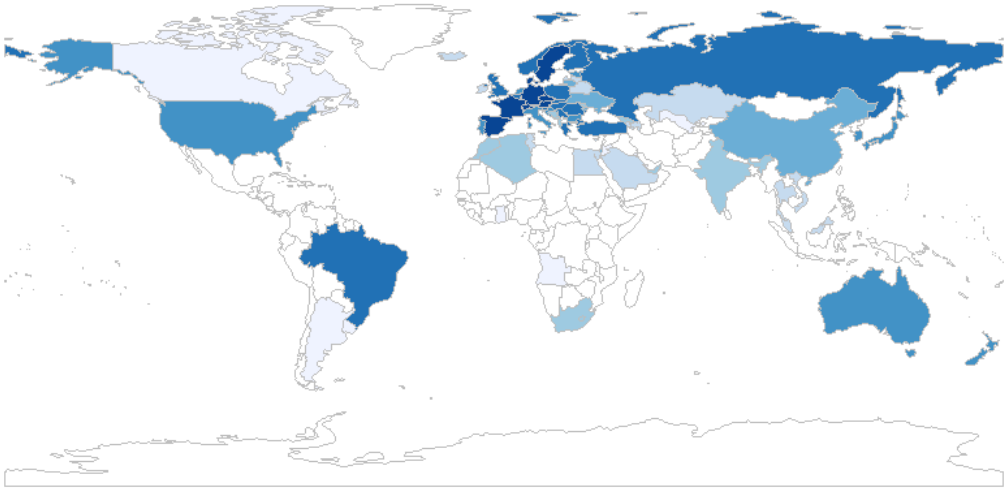

Figura 4.1. Distribuição de todas as ligas presentes no estudo pelo mundo

A Tabela 4.1 mostra a quantidade de ligas e temporadas de cada esporte presentes na base de dados. O esporte com mais ligas e temporadas no estudo foi o futebol (40.31% e 41.98%) e o esporte com menos foi o handebol (12.24% e 15.57%). Essas quantidades refletem um pouco a popularidade de cada esporte ao redor do mundo.

Tabela 4.1. Quantidade de ligas e temporadas em cada esporte

Esporte	**Ligas**		**Temporadas**	
	Freq.	**%**	**Freq.**	**%**
Handebol	25	12.63%	234	15.57%
Basquete	42	21.21%	310	20.63%
Voleibol	51	25.76%	328	21.82%
Futebol	80	40.40%	631	41.98%
Total	198	100%	1503	100%

As Figuras de 4.2 à 4.5 mostram a distribuição geográfica das ligas e temporadas de cada um dos quatro esportes da base de dados pelo mundo. A intensidade das cores dos países aumentam de acordo com a quantidade de temporadas presentes no estudo.

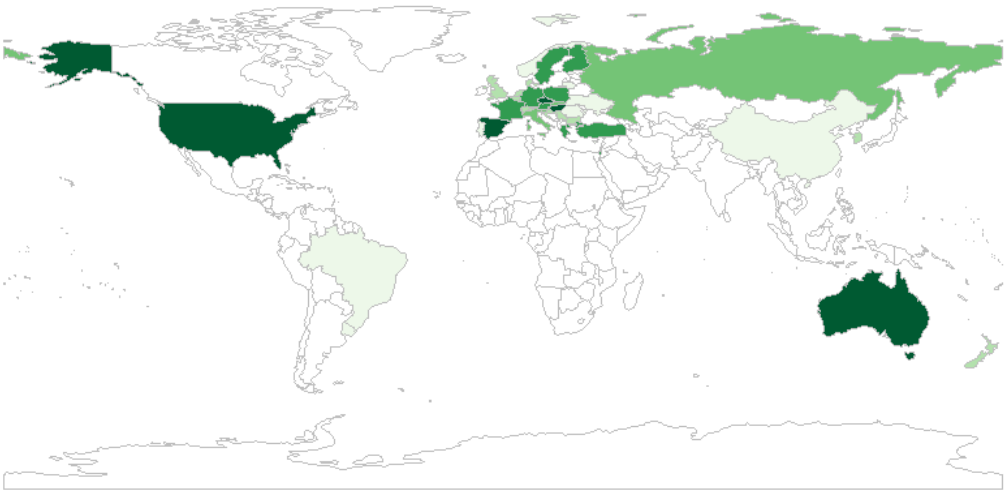

Figura 4.2. Distribuição das ligas de basquete pelo mundo

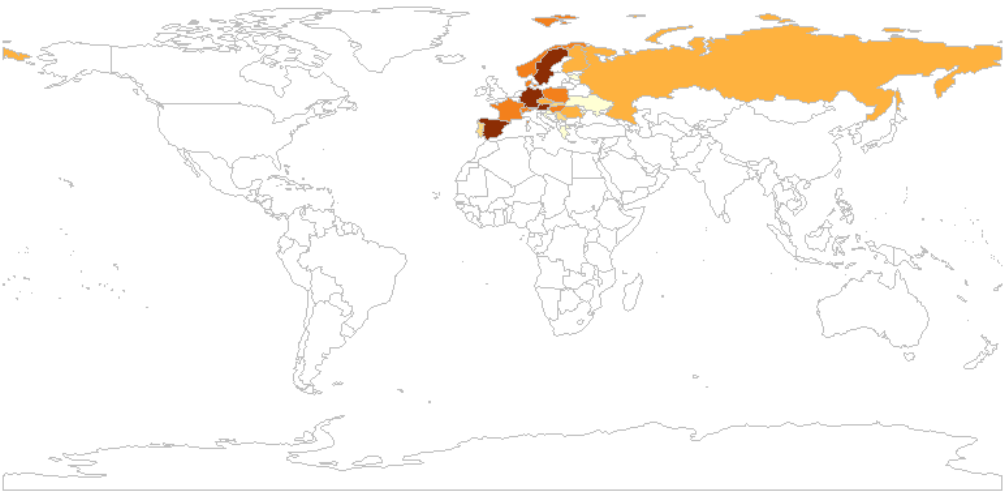

Figura 4.3. Distruibuição das ligas de handebol pelo mundo

A Figura 4.2 mostra a distribuição das ligas de basquete da base de dados pelo mundo. As ligas estão concentradas principalmente na Europa, Estados Unidos e Austrália; neste estudo não há nenhuma liga de basquete da África. A Figura 4.3 representa a distribuição das ligas de handebol presentes na base de dados pelo mundo. Nota-se que os campeonatos de handebol de pontos corridos presentes no estudo com mais de 5 temporadas e mais de 7 equipes estão restritos à Europa.

O esporte mais democrático dos quatro presentes na base de dados é o futebol,

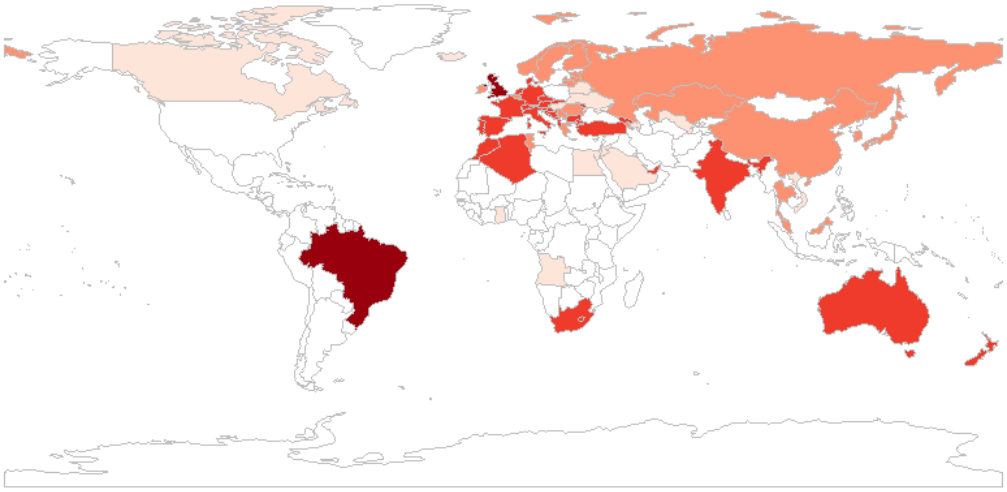

Figura 4.4. Distruibuição das ligas de fubebol pelo mundo

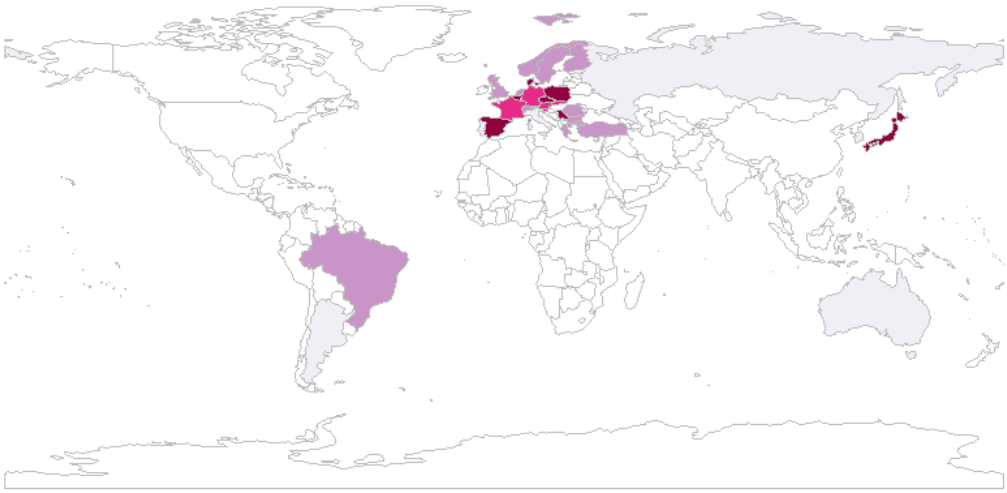

Figura 4.5. Distruibuição das ligas de voleibol pelo mundo

como mostra a Figura 4.4. O futebol está presente na África, América, Ásia, Europa e Oceania. Embora muito popular na América Latina, os campeonatos de futebol da maior parte desses países não seguem o modelo de pontos corridos e por isso não foram incluídos na base de dados deste estudo. As ligas de voleibol presentes na base de dados ocorrem principalmente na Europa, alguns países da Ásia, Oceania e América do Sul, como mostra a Figura 4.5.

Note que alguns países considerados potências em certos esportes não têm

suas ligas representadas nesse estudo. Existem três possíveis explicações: a primeira é a falta de seus dados no site em que a base de dados foi extraída, a segunda é o campeonato destes países não seguir o modelo de campeonatos de pontos corridos. A terceira possível explicação para uma liga não estar presente na base de estudos é a quantidade de temporadas disponíveis. Se a liga foi criada ou modificada e possui menos de 5 temporadas nesse novo formato ela não é incluída no estudo.

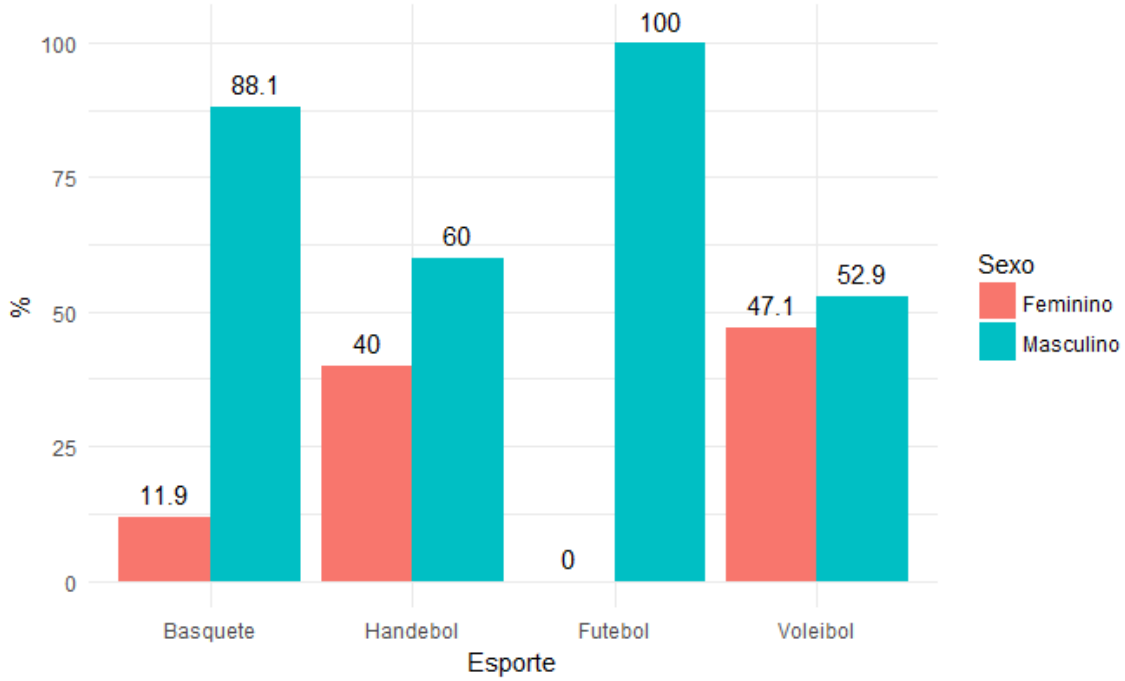

Figura 4.6. Distribuição das ligas por sexo

Existem campeonatos disputados por ambos os sexos na base de dados. Aproximadamente 80% das ligas da base coletada são disputadas por jogadores do sexo masculino e 20% por jogadores do sexo feminino. A proporção de ligas de cada sexo difere em cada esporte como mostra a Figura 4.6. No Futebol, nenhuma das ligas da base de dados é disputada por jogadores do sexo feminino e apenas 11.90% das ligas de basquete são disputadas por mulheres. No Voleibol e no handebol existe um maior equilíbrio entre a proporção de ligas disputadas por homens e mulheres, sendo o percentual de ligas disputado por mulheres igual a 40.00% e

47.06% respectivamente.

Intuitivamente, espera-se que a distribuição final dos pontos de campeonatos de pontos corridos cujas equipes tenham habilidades muito distintas possua uma variância maior que um campeonato em que as equipes possuem habilidades semelhantes. Para ilustrar esse fato, será mostrado o campeonato da Série A - Brasil em 2015 e como seria a distribuição final dos pontos simulados sob a hipótese de igualdade entre as habilidades das equipes. Nessa temporada o Corinthians foi o campeão com 81 pontos conquistados ao longo do campeonato e o último colocado foi o Joinville com apenas 31 pontos. A distribuição final dos pontos observada no campeonato é mostrada na Tabela 4.2.

Tabela 4.2. Distribuição final dos pontos da Série A - Brasil no ano de 2015

Equipe	**Pontos**
Corinthians	81
Atlético-MG	69
Grêmio	68
São Paulo	62
Internacional	60
Sport Recife	59
Santos	58
Cruzeiro	55
Palmeiras	53
Atlético-PR	51
Ponte Preta	51
Flamengo	49
Chapecoense	47
Fluminense	47
Coritiba	44
Figueirense	43
Avaí	42
Vasco	41
Goiás	38
Joinville	31

Nessa temporada da Série A a variância observada da distribuição final dos pontos foi 142.37. O modelo aleatório, em que todas as equipes possuem a mesma habilidade e existe influência apenas por ser a equipe mandante ou não, foi cons-

truído como mostrado na Seção 3.1 e sua variância foi 59.46. A Figura 4.7 mostra dois histogramas com a distribuição dos pontos observadas na temporada e do modelo aleatório. Nota-se facilmente através dos gráficos a diferença entre as variâncias dos dois tipos de campeonatos.

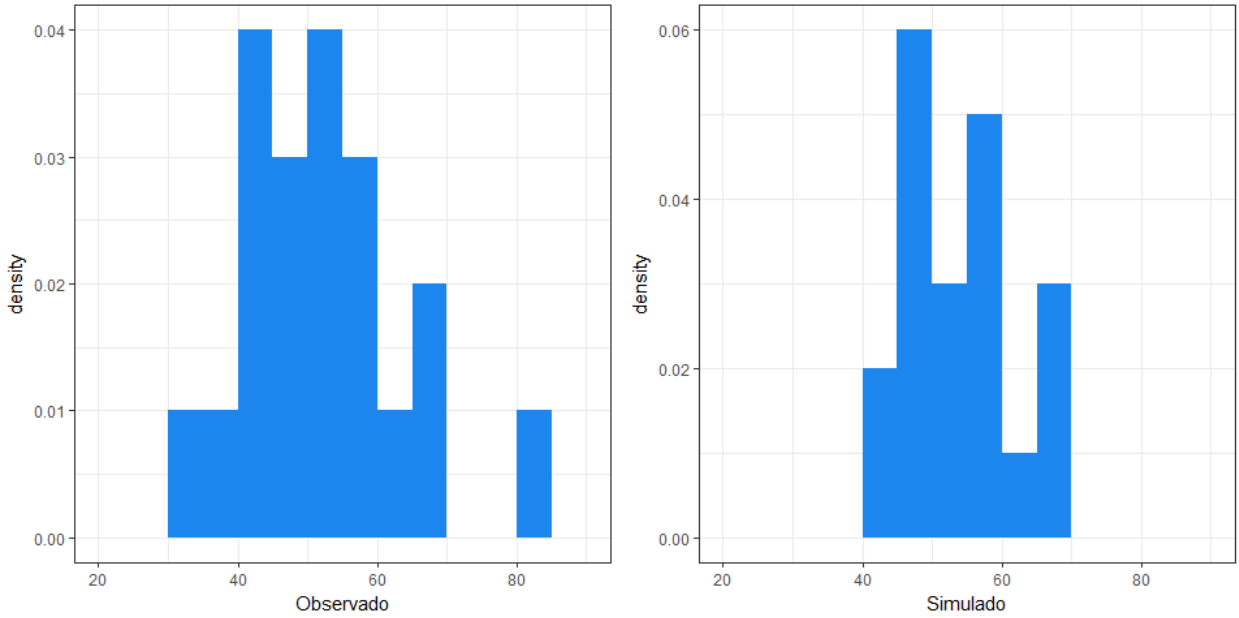

Figura 4.7. Comparação da distribuição de pontos observadas na Série A - Brasil 2015 e de um campeonato simulado em que todas as equipes possuem a mesma habilidade.

A segunda base de dados é utilizada para fazer o ajuste do modelo gráfico probabilístico que estima as habilidades das equipes em uma temporada/liga. Neste trabalho, o modelo é ajustado somente aos dados da liga de basquete americana *National Basketball Association*(NBA), mas sua teoria pode ser estendida para outras ligas de basquete e outros esportes. A NBA foi escolhida devido à grande quantidade de dados disponíveis e por ser uma liga em que a habilidade das equipes tem uma grande influência nos resultados finais de acordo com o coeficiente proposto ϕ. Todas essas informações das temporadas da NBA a partir de 2004 foram coletadas do site `www.basketballreference.com`.

A habilidade das equipes é estimada usando características das próprias equipes e os resultados dos jogos da temporada. O vetor de características $\mathbf{x_i}$ utilizado para estimar as habilidades das equipes possui algumas variáveis explicativas ex-

traídas de uma rede de conexões entre os jogadores e suas equipes, como propostas por Vaz de Melo et al. [2012]. Assim como feito em [Vaz de Melo et al., 2012], a rede de conexões irá utilizar as últimas 6 temporadas. Cada temporada possui uma rede de conexões associada que é representada por um grafo em que os jogadores e as equipes são vértices. Para um ano Y, dois jogadores da temporada são conectados no grafo por uma aresta se eles jogaram juntos em algum momento nas 6 temporadas anteriores à Y e ambos ainda jogam na NBA nesta equipe ou em outra; um jogador e uma equipe são conectados por uma aresta no grafo se o jogador atuou na equipe em algum momento nas 6 temporadas anteriores a Y e ainda joga na NBA, nesta ou em outra equipe. As variáveis explicativas $\mathbf{x_i}$ para um ano Y consideradas no ajuste do modelo são:

- CO: Representa a conferência ao qual as equipes pertencem. Assume o valor 0 se a equipe pertence a conferência Leste e 1 se a equipe jogar pela conferência Oeste.
- A5: Média dos 5 maiores salários pagos por cada equipe no ano Y.
- A6: Média dos salários entre o 6º e o 10º jogadores mais bem pagos por cada equipe no ano Y.
- SD: O desvio padrão dos salários pagos por cada equipe no ano Y.
- AP: Média do *Player Efficiency Rating*(PER) dos jogadores de cada equipe no ano $Y-1$. O PER foi criado por Hollinger [2005] e é uma medida de ranqueamento dos jogadores calculada de acordo com suas performances. Quando maior for o valor do PER de um jogador, melhor é sua posição no raqueamento. O valor do PER é padronizado de tal forma que sua média na temporada é sempre 15.
- VL: A volatibilidade de uma equipe mede o quanto ela troca seus jogadores. Seu cálculo é definido como $\Delta d_t^Y = d_t^Y - d_t^{Y-\epsilon}$, em que d_t^Y é o grau do vértice do time t no ano Y e ϵ é um parâmetro que representa uma janela de tempo. Neste trabalho adotou-se $\epsilon = 2$, como feito em [Vaz de Melo et al., 2012]. Valores altos de VL indicam que a equipe fez drásticas mudanças na composição de seu time.

- RV: A variável *Roster Aggregate Volatility* mede quanto os jogadores de uma equipe t mudaram de times nos 6 anos anteriores ao ano Y. Essa variável é definida como $\sum \Delta d_t^Y = \sum_{\forall v \in R_t^Y} \frac{d_v^t}{Y - Y0_v}$, em que R_t^Y representa o conjunto de jogadores da equipe t no ano Y, $y0_v$ representa o ano da primeira temporada do jogador $v \in R_t^Y$ no seu atual time t e d_v^t é o grau do vértice do jogador $v \in R_t^Y$ no ano Y. Valores altos de RV indicam que os jogadores de uma equipe qualquer t possuem uma grande tendência a trocarem de times.

- CC: Representa a inexperiência de uma equipe. Essa variável é calculada usando o coeficiente de clusterização de grafos. Esse coeficiente mede o grau com que os nós de um grafo tendem a agrupar-se. Um valor alto de CC indica que existem muitas conexões entre os jogares da equipe.

- RC: A variável *Roster Aggregate Coherence* mede força da relação entre os jogadores de uma equipe no ano Y. Um alto valor de $\overline{cc_t^y}$ indica que os jogadores da equipe t jogam juntos por um tempo substancial ou poucas mudanças ocorram na equipe nos anos anteriores a Y. Essa medida é definida como $\overline{cc_t^Y} = avg(cc_v^t \times (Y - Y0_v)), \forall v \in R_t^Y$, em que cc_v^t é o coeficiente de clusterização do jogador $v \in R_t^Y$.

- SI: Representa a quantidade de jogadores da equipe t no ano Y.

As Figuras 4.8, 4.9, 4.10, 4.11 e 4.12 mostram o comportamento de algumas variáveis explicativas. Cada Figura mostra o comportamento de uma variável explicativa em diferentes anos para cada uma das equipes presentes na NBA. Conclui-se a partir da Figura 4.9, por exemplo, que em 2016 a equipe do *Memphis*(MEM) pagou melhor os jogadores que recebem entre o 6º e 10º maiores salários da equipe do que pagava nos anos anteriores. Entretanto, essa mesma equipe não investiu mais dinheiro nos top 5 salários da sua equipe, como mostra a Figura 4.8. A Figura 4.12 mostra as oscilações na quantidade de jogadores que cada equipe contrata. O *Miami Heat*(MIA), por exemplo, quase não altera a quantidade de jogadores que contrata em cada ano, enquanto o *Philadelphia 76ers*(PHi) aumentou a quantidade de contratações nos últimos dois anos. Considerando o comportamento do desvio padrão mostrado na Figura 4.10 observa-se que o *Los Angeles Lakers*(LAL) possuía os maiores desvios padrões nos anos de 2013 e 2014, mas a partir de 2015

ele diminuiu drasticamente. Por fim, a Figura 4.11 mostra o PER médio de cada equipe nas últimas 5 temporadas. Observa-se que o *Portland*(POR) possui o PER quase constante ao longo dos anos, enquanto o *Miami Heat*(MIA) aumentou seu PER médio em todos os anos.

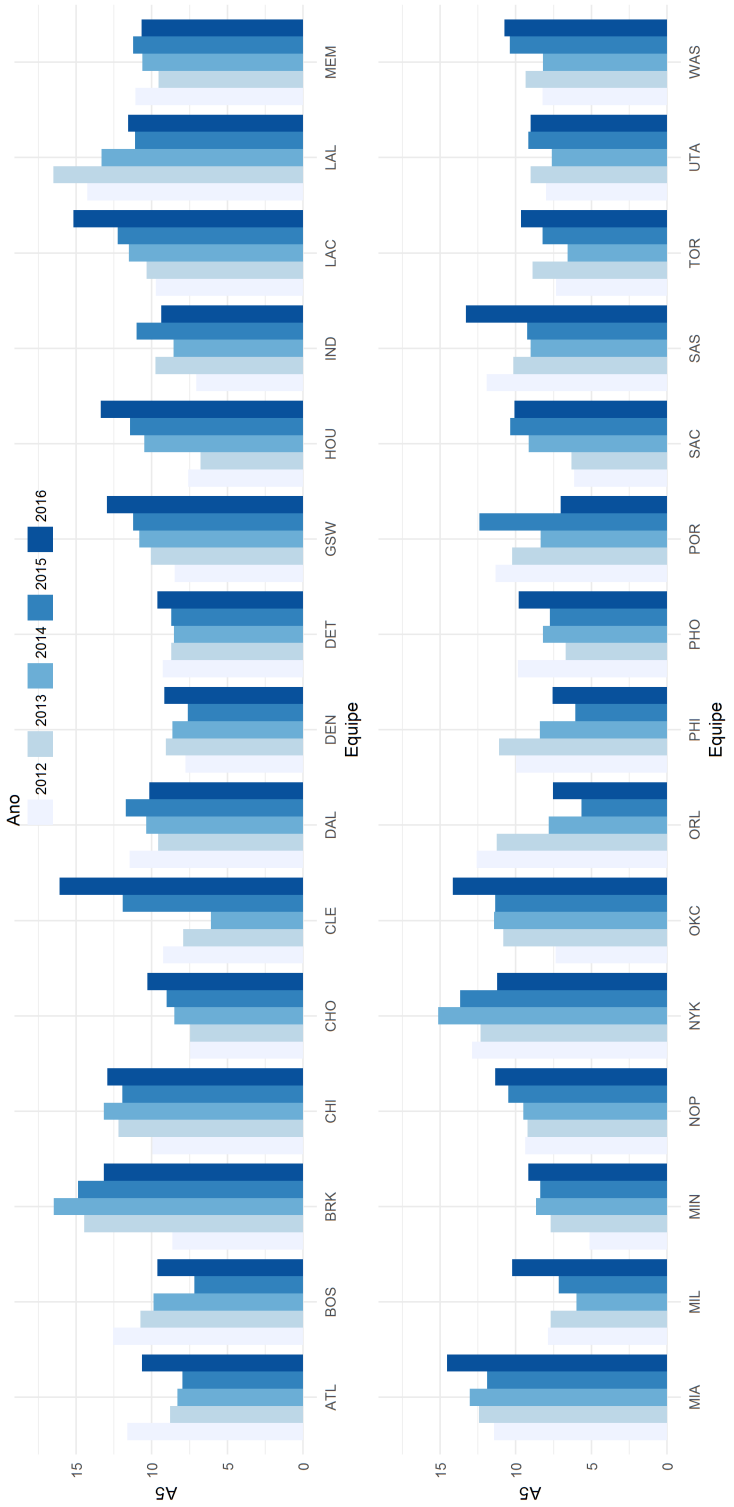

Figura 4.8. Variável A5 para cada equipe da NBA ao longo dos anos.

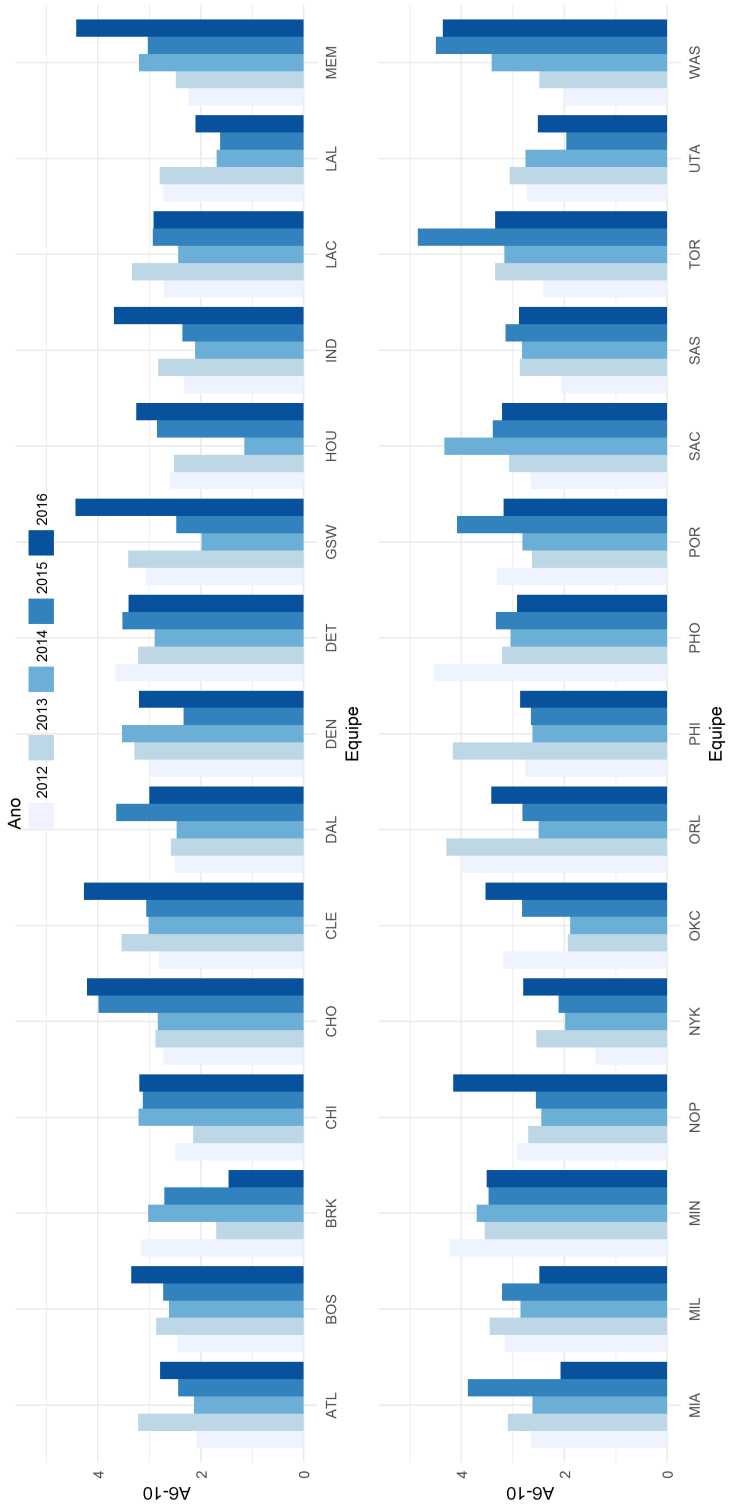

Figura 4.9. Variável A6-10 para cada equipe da NBA ao longo dos anos.

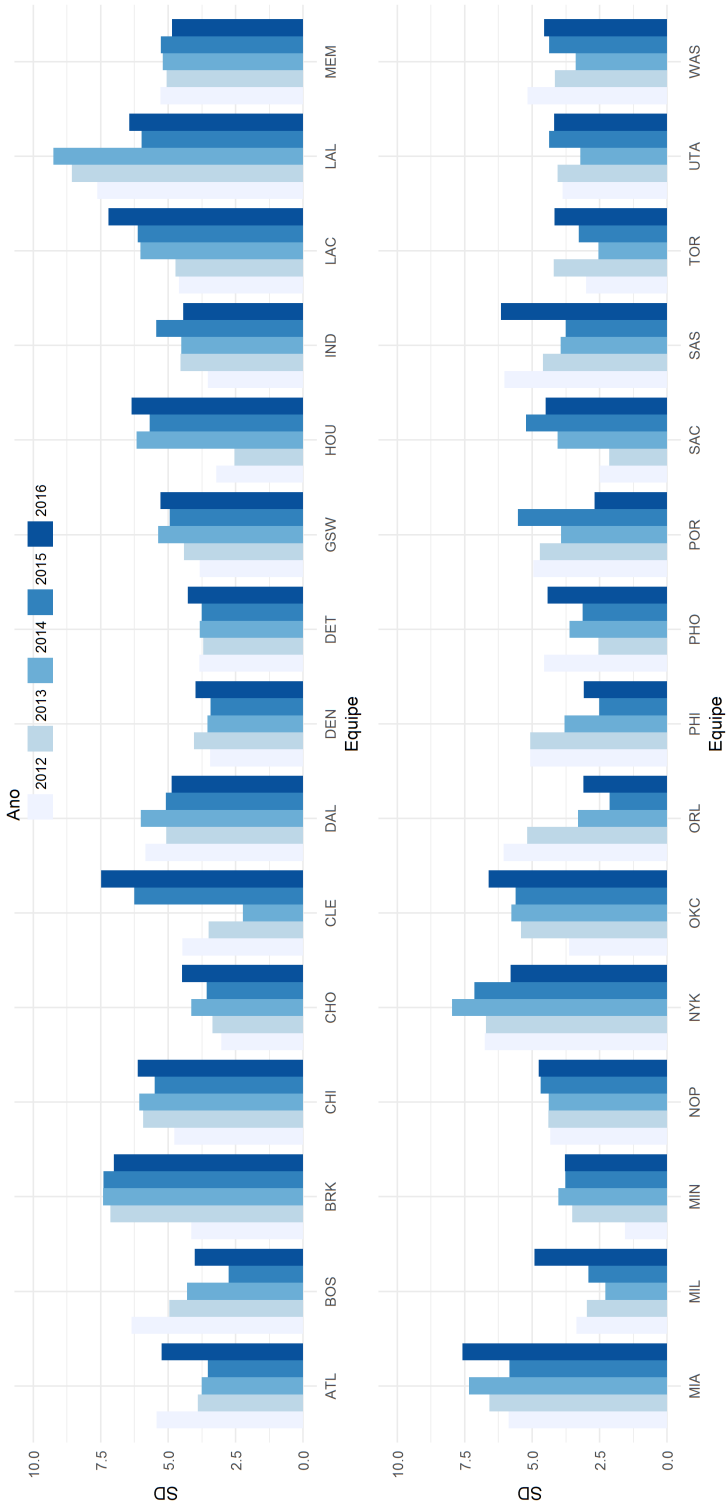

Figura 4.10. Variável SD para cada equipe da NBA ao longo dos anos.

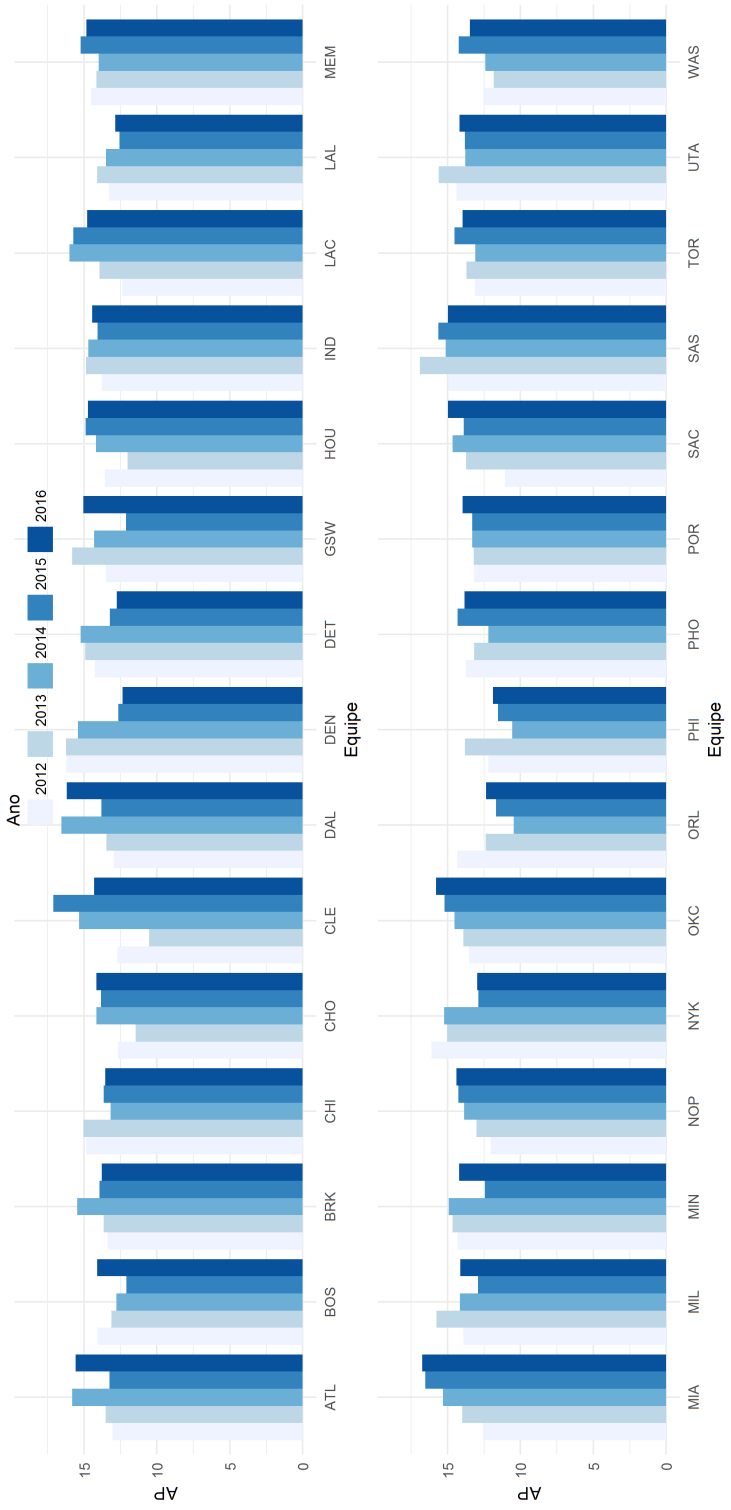

Figura 4.11. Variável AP para cada equipe da NBA ao longo dos anos.

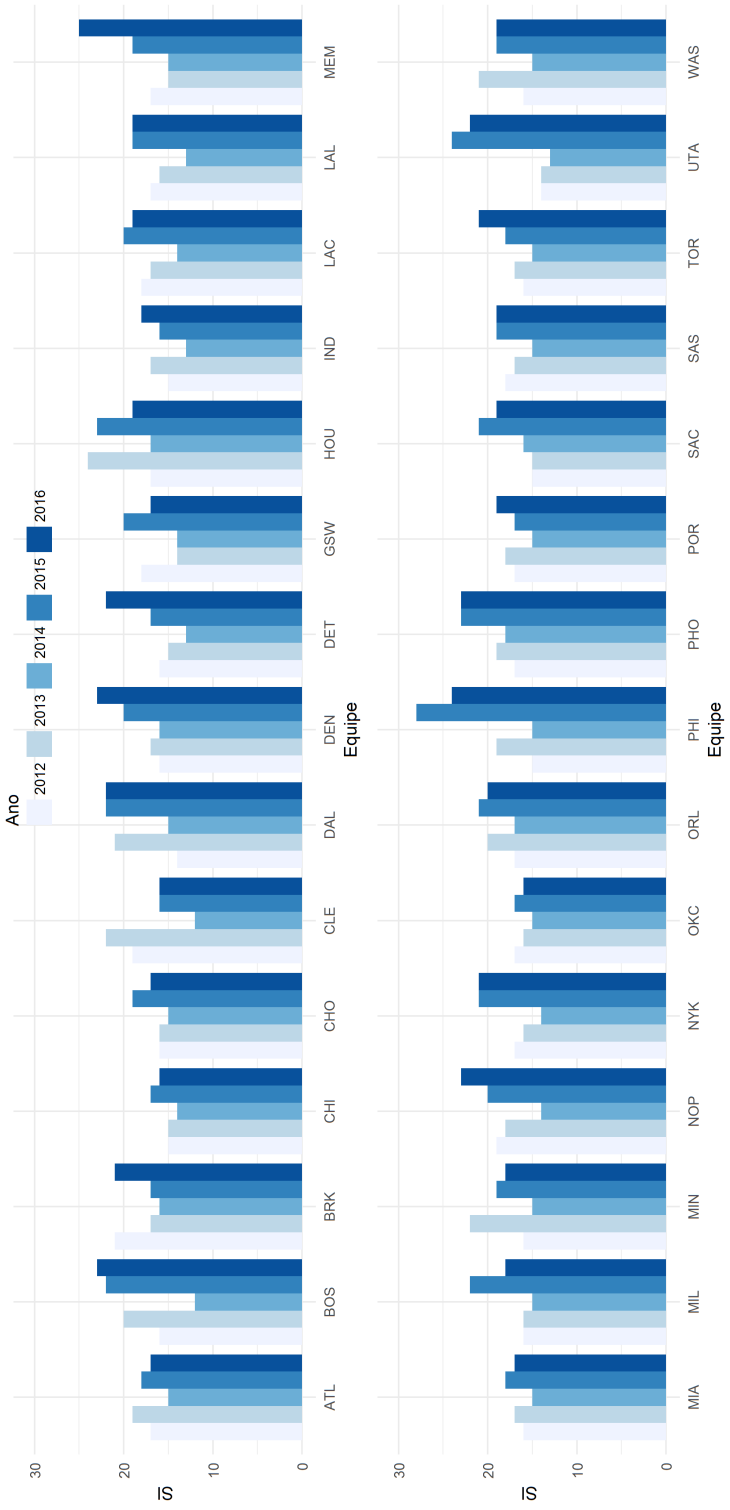

Figura 4.12. Variável SI para cada equipe da NBA ao longo dos anos.

4.2 Separando a sorte da habilidade

O coeficiente ϕ proposto na Seção 3.1 do Capítulo 3 foi estudado empiricamente na base de dados com 1503 temporadas esportivas de diferentes esportes, mostrada na Seção 4.1.

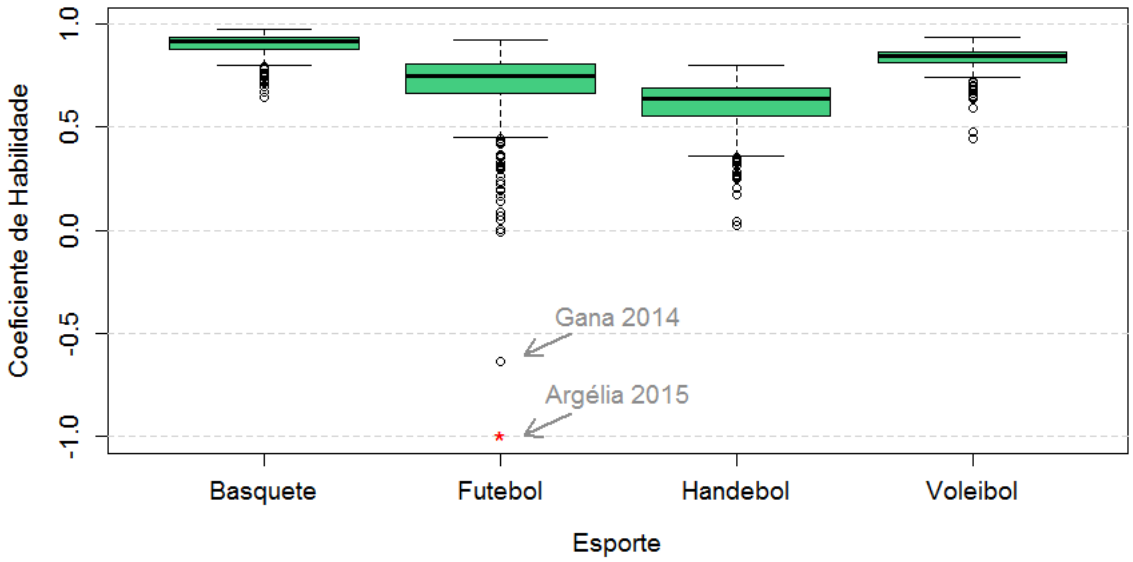

Figura 4.13. Boxplot dos valores do coeficiente ϕ das 1503 temporadas separadas por esporte.

A Figura 4.13 mostra o comportamento do coeficiente ϕ em cada esporte. Quanto mais comprimida em direção à 1 estiver a caixa do boxplot, mais distante do modelo aleatório é o comportamento médio das ligas desse esporte e mais influente é o componente habilidade para explicar a pontuação final. Dessa forma, conclui-se que dentre os esportes estudados o basquete é o mais competitivo, isto é, a habilidade possui uma maior influência nos resultados finais. O segundo esporte mais competitivo é voleibol, seguido por futebol e handebol. Um teste de Kruskal-Wallis foi feito para verificar se as distribuições dos coeficientes ϕ de cada esporte são significativamente diferentes entre si. Esse teste não-paramétrico foi escolhido devido ao fato das distribuições dos valores de ϕ não seguirem uma distribuição Normal. Ao nível de 5% de significância, a conclusão deste teste é que as distribuições do coeficiente ϕ não são iguais entre si.

Considerando a estrutura de cada esporte, é possível encontrar uma explicação para esses resultados. Normalmente, nas partidas de basquete e voleibol os atletas fazem muitos pontos ao longo da partida. Devido a essa longa sequência de eventos relevantes, é mais difícil para equipes menos habilidosas vencerem jogos por pura sorte. Entretanto, em jogos de futebol e handebol não existem tantos eventos relevantes ligados aos pontos. No futebol, por exemplo, o número médio de gols por partida é apenas 2.62. Isso faz com que seja mais fácil para um time menos habilidoso ganhar uma partida e consequentemente pontos no campeonato por pura sorte. Esse resultado mostra que o coeficiente proposto é capaz de identificar quais ligas são mais influenciadas pelo componente habilidade ou sorte.

Tabela 4.3. Separando o coeficiente ϕ por esporte e componentes habilidade ou sorte.

Esporte	Habilidade		Sorte		Total
	Freq.	%	Freq.	%	
Basquete	310	100%	0	0%	310
Futebol	586	92.87%	45	7.13%	631
Handebol	192	82.05%	42	17.95%	234
Voleibol	326	99.39%	2	0.61%	328
Total	1414	94.08%	89	5.92%	1503

A Tabela 4.3 mostra para cada esporte qual componente foi mais influente nos resultados das temporadas/ligas de acordo com o coeficiente proposto. Usando o intervalo de confiança definido na Seção 3.1, é possível avaliar se um coeficiente ϕ observado é significativamente diferente ou igual a 0. Se o coeficiente ϕ observado não for significativamente diferente de 0, essa temporada/liga é classificada como uma temporada cujo componente mais influente é a sorte. Da mesma forma, uma temporada/liga em que o coeficiente ϕ é significativamente diferente de 0 e positivo tem como componente principal a habilidade das equipes. Nota-se a partir da Tabela 4.3 que todas as temporadas de basquete e 99.39% das temporadas de voleibol são definidas principalmente pelo componente habilidade. Em contraste, no handebol e no futebol a pura sorte é o componente de maior influência em 17.95% e 7.13% das temporadas, respectivamente

A distribuição dos coeficientes ϕ do basquete e do voleibol mostrados na Figura 4.13 é muito próxima do seu valor máximo e isso pode levar a uma in-

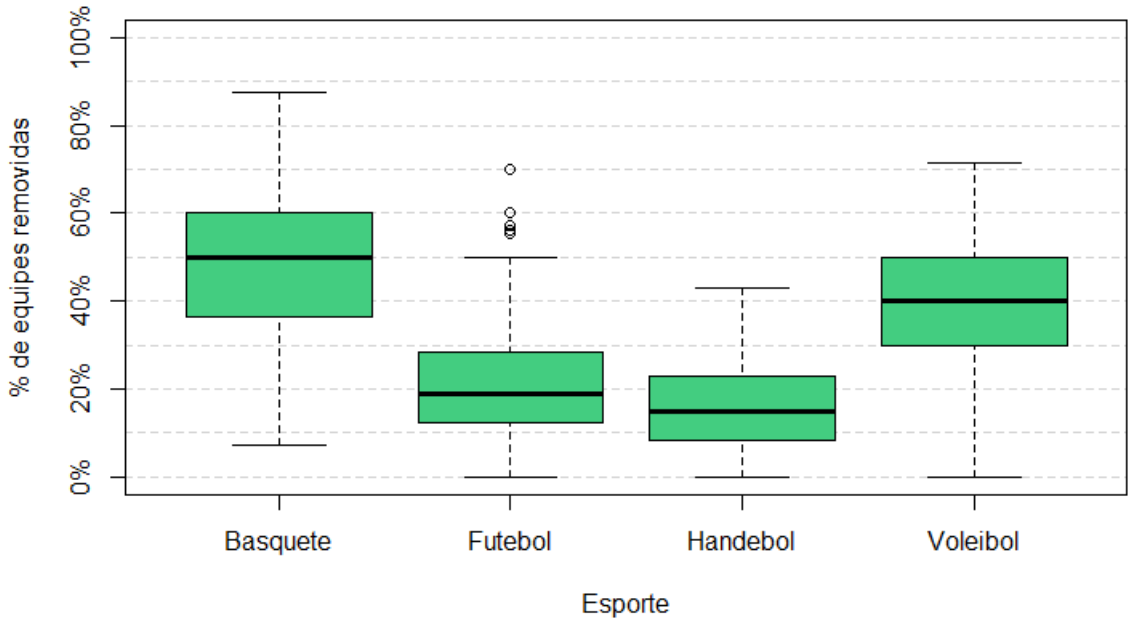

Figura 4.14. Percentual de equipes removidas

terpretação errônea de que a sorte não possui nenhuma influência nos resultados desses esportes. Mas ressalta-se que embora o componente habilidade possua uma forte influência nos resultados desses esportes, esse componente não determina 100% do ranking final. Um fato que pode ocorrer e ilustrar essa situação são as temporadas em que existem algumas equipes mais habilidosas enquanto as demais equipes possuem aproximadamente o mesmo nível de habilidade. As equipes mais habilidosas tendem a elevar o coeficiente ϕ e ter o resultado de suas partidas definidas através de suas habilidades, mas os jogos das demais equipes é muito provavelmente semelhante a uma loteria.

A Figura 4.14 mostra o percentual de equipes que deveriam ser removidas das temporadas/ligas para torna-las temporadas aleatórias. A forma com que é feita a remoção das equipes, a seleção de quais equipes serão removidas e os critérios de desempate foram explicados na Seção 3.1. O basquete e o voleibol são os esportes que exigem a remoção de mais equipes, sendo o percentual mediano de equipes a serem removidas 50% e 40% respectivamente. O handebol(14%) e o futebol(19%), que possuem tipicamente valores menores do coeficiente ϕ, exigem que um número menor de equipes sejam removidas para tornar suas ligas aleatórias. Para mostrar

o quão diferente o futebol e o handebol são, considere duas importantes ligas de futebol: a *Primera División* da Espanha e a *Premier League* da Inglaterra. Em cada temporada ocorrida entre 2007 e 2016, a quantidade média de equipes que tiveram que ser removidas foi 3.2 e 4.9 respectivamente. Considerando que essas ligas possuem 20 equipes disputando a competição, isso significa remover 16% e 25% das equipes. Mais importante que isso, todos os anos são removidas praticamente as mesmas equipes. Por exemplo, na liga espanhola, as 3.2 equipes que devem ser removidas incluem o Real Madrid 10 vezes e o Barcelona 9 das 10 temporadas estudadas. Isso significa que, removendo essas duas equipes das 20 que disputam o campeonato, tem-se uma competição que produz um *rank* praticamente aleatório entre as demais equipes. No campeonato inglês *Premier League*, o *Manchester United* deve ser removido em 9 das 10 temporadas estudadas e outras equipes aparecem frequentemente na lista dos times que, se removidos, deixam o campeonato entre as equipes restantes aleatório. Esse resultado mostra como a influência do componente habilidade é menor no futebol e no handebol. Para fazer um paralelo com outros esportes, considere a competição de basquete americana NBA. Nessa liga é necessário remover entre 17 e 25 equipes das 30 que disputam o campeonato para ter uma competição aleatória, refletindo a influência do componente habilidade nessa competição.

O valor do coeficiente ϕ mais extremo mostrado na Figura 4.13 foi observado na liga de futebol *Division 1* da Argélia. Seu valor foi igual a -1.93 na temporada 2014-2015. O valor do coeficiente ϕ da Argélia nessa temporada é mostrado somente figurativamente no gráfico, ou seja, sua altura vertical é diferente do seu verdadeiro valor. Fez-se essa alteração para que a visualização dos resultados das demais temporadas não fosse prejudicada. Valores negativos não são muito comuns e esse foi um valor extremo, considerando que o segundo menor valor observado foi $\phi \approx -0.6$. Na temporada 2014-2015 da *Division 1* da Argélia, a pontuação das equipes era muito similar e perto do final do campeonato todos os times ainda tinham chance de serem campeões, como destaca a reportagem Shergold [2015](veja Figura 4.15). Essa temporada foi disputada por 16 equipes e durante a terceira semana do campeonato, Albert Bodjongo, um jogador da equipe JSK, foi dramaticamente morto durante um jogo. Nessa partida, o JSK jogava como mandante e no final de um jogo contra o USM Alger, quando perdia por 2 a 1, torcedores

Figura 4.15. Título de uma reportagem mostrando o quão próximo era a pontuação das 16 equipes na temporada 2014-2015 do campeonato argelino.

do JSK infelizes com o resultado lançaram um objeto no campo. Esse objeto feriu Bodjongo, que faleceu horas depois do ocorrido. Como consequência, a Federação de Futebol da Argélia suspendeu todos os jogos de futebol indefinidamente e ordenou o fechamento do estádio em que a fatídica partida ocorreu. A temporada voltou a ser disputada mais tarde, entretanto, não é claro o quanto que esses infelizes eventos influenciaram o valor extremante negativo de ϕ obtido para essa temporada/liga.

Por fim, será avaliada a diferença entre o coeficiente ϕ de ligas disputadas por atletas do sexo feminino e do sexo masculino. A Figura 4.6 da Seção 4.1 mostra a quantidade de temporadas disputadas por homens e mulheres em cada esporte. A Figura 4.16 mostra a distribuição do coeficiente ϕ nos campeonatos disputados por cada um dos sexos. Nota-se que o comportamento do coeficiente ϕ dos campeonatos masculinos é similar ao comportamento do coeficiente ϕ nos

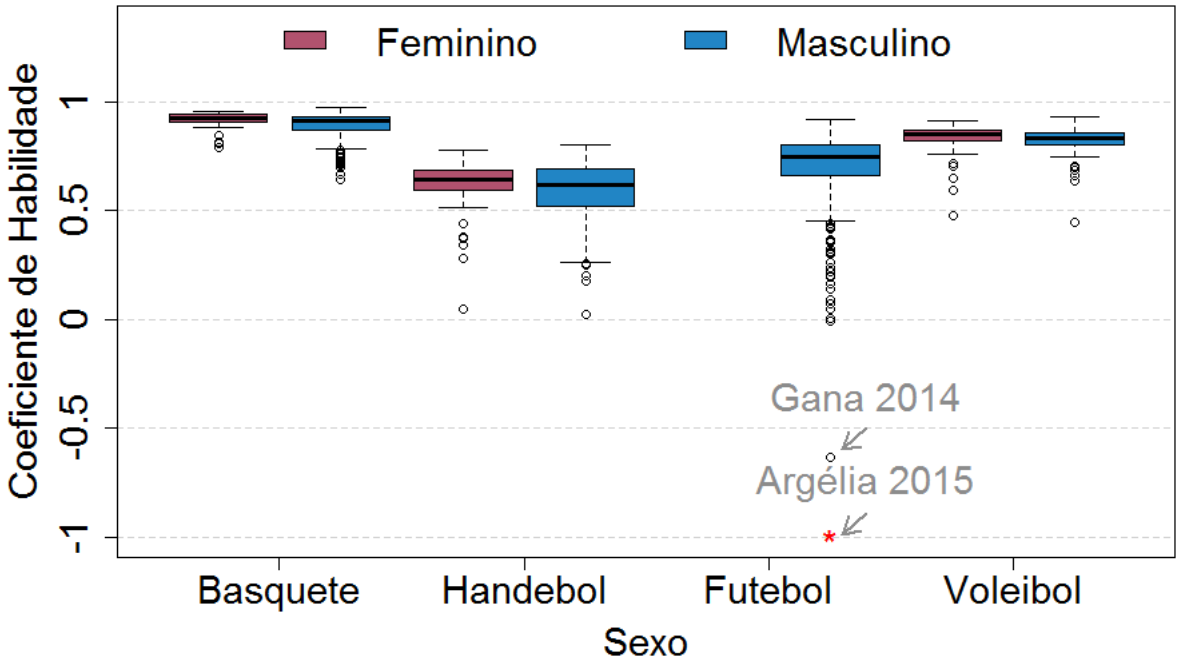

Figura 4.16. Comparação do coeficiente ϕ pelo gênero dos atletas

campeonatos femininos em todos os esportes.

O nome de cada liga, as temporadas, o campeão, a quantidade de jogos da temporada, quantidade de equipes, coeficiente ϕ e outras características das ligas avaliadas neste trabalho podem ser vistas em `https://docs.google.com/spreadsheets/d/10Gg7vC3wuatpqdstqH0AFo0CgF6nDX3WrGX60b27PuY/edit?usp=sharing`

4.3 Estimação das habilidades

O modelo Bayesiano proposto será ajustado aos dados da *National Basketball Associaion*(NBA). Essa liga foi escolhida devido a grande disponibilidade de dados, prestígio e alto valor do coeficiente ϕ em suas temporadas. A liga da NBA é dividida em duas partes: a temporada regular e os *playoffs*. Na temporada regular todas as equipes jogam uma contra a outra pelo menos duas vezes: uma como equipe mandante e outra como visitante. Nessa primeira etapa do campeonato as

equipes apenas acumulam os pontos de suas vitórias. No final da fase regular, essa pontuação define quais equipes irão disputar a fase de *playoffs*, que são jogos de caráter eliminatório. No basquete não existem empates e as temporadas possuem 30 equipes divididas igualmente entre as conferências Leste e Oeste. Apenas 8 equipes de cada conferência passam para a fase de *playoffs*, que define as posições finais das equipes no campeonato. No ajuste do modelo gráfico probabilístico são considerados somente os jogos da temporada regular, devido aos poucos jogos que ocorrem na fase de *playoffs* e sua dinâmica diferente.

A Seção 4.1 apresentou a base de dados que será utilizada no ajuste do modelo. Com exceção da variável explicativa CO, todas as demais variáveis foram padronizadas. O modelo bayesiano foi ajustado utilizando 10000 iterações do algoritmo Metropolis-Hastings e adotou-se um *burn-in* igual a 2000. A taxa de aceitação média dos novos parâmetros propostos pelo Metropolis-Hastings foi 0.395, que é um valor razoável, e o desvio padrão foi de 0.02.

Tabela 4.4. DIC dos melhores modelos ajustados.

	Variáveis do Modelo	DIC				
		2012	2013	2014	2015	2016
1	CO+A5+AP+VL+RC+SI	**-683589.7**	**-871853.9**	**-901784.2**	**-890471.0**	**-922389.1**
2	CO+A5+A6+AP+SD+VL+RC+SI	-678306.4	-865637.7	-895432.8	-882730.4	-915692.8
3	CO+A5+AP+SD+VL+RC+SI	-681779.5	-869800.7	-899301.4	-888251.6	-920525.5
4	CO+A5+A6+AP+VL+RV+CC+RC+SI	-675140.2	-860787.3	-890335.5	-87362.2	-908063.3

A Tabela 4.4 apresenta o valor do DIC para os melhores modelos ajustados. Conclui-se a partir desses resultados que o modelo 1, que inclui as variáveis $CO + A5 + AP + VL + RC + SI$, é o melhor modelo em todas as temporadas analisadas. Assim como ocorreu em Vaz de Melo et al. [2012], as variáveis VL e RC se mostraram relevantes no ajuste do modelo enquanto as demais variáveis da rede de conexões não foram relevantes. As variáveis CO, $A5$ e AP não estavam presentes em Vaz de Melo et al. [2012], mas essas variáveis influenciam nas habilidades das equipes, de acordo com o modelo gráfico probabilístico ajustado.

A Figura 4.17 apresenta os coeficientes estimados pelo modelo 1 em todas as temporadas da NBA presentes no estudo. O salário médio dos 5 maiores salários da equipe ($A5$) e o PER médio (AP) possuem um efeito positivo na habilidade, como era o esperado. Isto significa que, fixando as demais variáveis, quanto mais

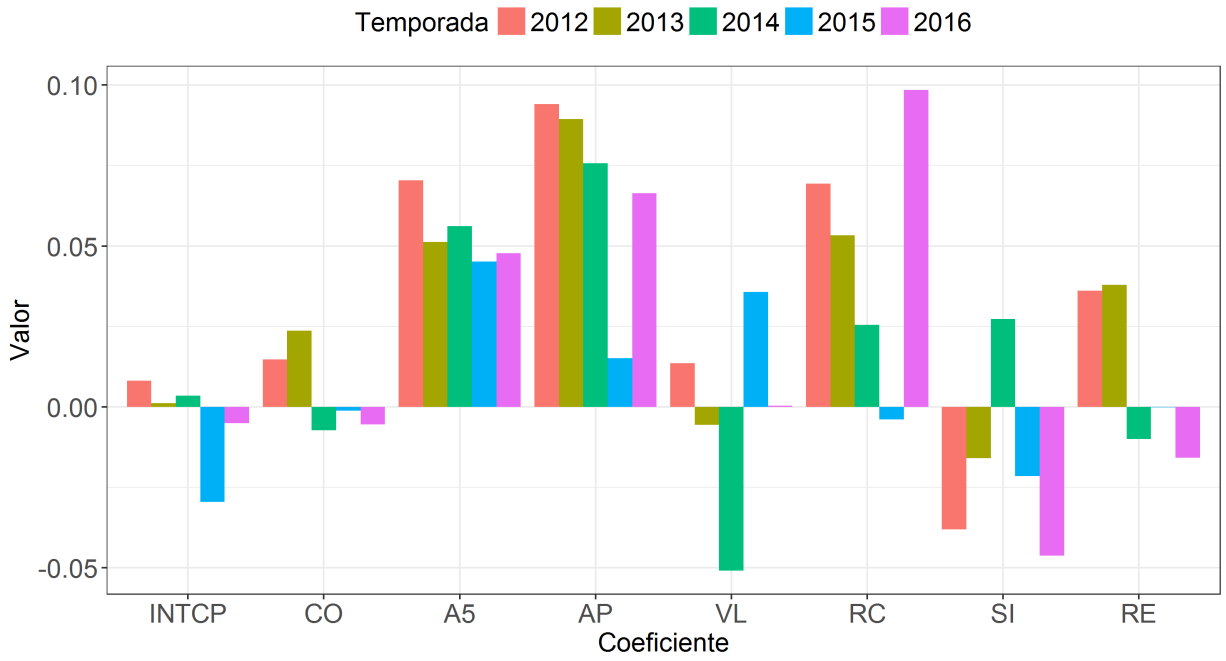

Figura 4.17. Coeficientes estimados a partir do Metropolis-Hastings - Modelo 1

bem pagos são os 5 jogadores com maiores salários da equipe e maior o PER médio da equipe, maior tende a ser a habilidade esperada. Com exceção da temporada 2014-15, a variável *Roster Aggregate Coherence* (RC) também possui um efeito positivo na habilidade das equipes. Por outro lado, com exceção da temporada 2013-14, o tamanho das equipes (SI) possui um efeito negativo na habilidade das equipes.

Os gráficos mostrados na Figura 4.18 mostram as correlações entre as habilidades estimadas pelo modelo 1 versus a quantidade de jogos vencida pelas equipes durante a fase regular das temporadas avaliadas. Cada círculo representa uma equipe e o símbolo dentro do círculo representa a posição da equipe após a disputa da fase de *playoffs*: o número representa a posição da equipe no ranking, as letras P e N representam as equipes que foram eliminadas na primeira disputa dos *playoffs* e as equipes que não passaram para a fase de *playoff*, respectivamente.

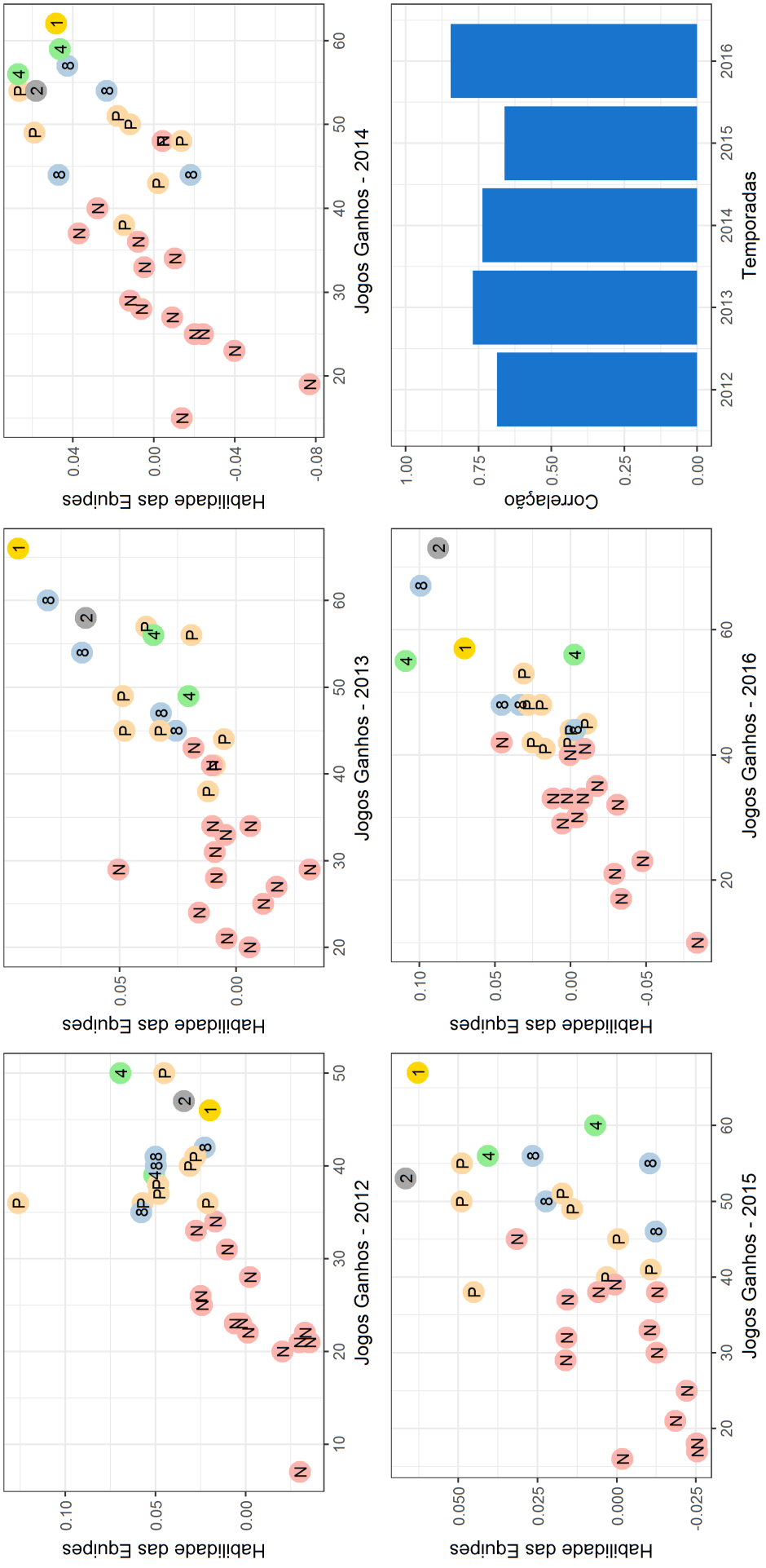

Figura 4.18. Correlação entre a habilidade estimada pelo modelo 1 e a quantidade de jogos ganhos em cada temporada da NBA analisada

Os gráficos de dispersão da Figura 4.18 mostram que as habilidades estimadas $\hat{\alpha}$ são muito correlacionadas com o número de vitórias da temporada regular e com a posição final das equipes após a fase de *playoffs*. Entretanto, disputas esportivas são sempre suscetíveis a ocorrência de eventos surpreendentes e influências adicionais durante os jogos ocorridos na fase de *playoff*. Consequentemente, o ranking final também não possui uma correlação linear perfeita com o número de vitórias. Por exemplo, a equipe com a maior habilidade estimada pelo modelo 1 na temporada de 2016 terminou na 4ª posição no rank final; nesse mesmo ano a equipe com a segunda maior habilidade estimada foi o time com o segundo maior número de vitórias, mas terminou o campeonato somente na 8ª posição. Ainda analisando a temporada de 2016, o modelo estimou a 4ª maior habilidade para o campeão e ele também teve o 4º maior número de vitórias durante a temporada regular.

Em 2012, a maior habilidade estimada pelo modelo 1 foi associado a uma equipe eliminada na primeira fase dos *playoffs*. Essa equipe em questão era o *New York Knicks*, que perdeu na primeira da fase de *playoffs* para o *Miami Heat*, o campeão dessa temporada. O *Miami Heat* não se destacou muito pelo número de vitórias na temporada regular e nem por sua habilidade estimada pelo modelo 1, por isso sua colocação final é um exemplo de que eventos de sorte ou externos possuem certa influência no ranking final, mesmo em ligas extremamente competitivas. Nesse ano em questão a equipe *New York Knicks* perdeu dois jogadores importantes devido a contusões antes da partida contra *Miami Heat*, o que certamente contribuiu com sua derrota [Wikipedia, 2017]. Por outro lado, no ano de 2013 a equipe com o maior número de vitórias e a maior habilidade estimada pelo modelo 1 foi também a equipe campeã, mostrando que embora a sorte aconteça e tenha certa influência nos resultados, ela claramente não é a regra.

O último gráfico da Figura 4.18 mostra o valor exato da correlação entre as habilidades estimadas $\hat{\alpha}$ e o número de vitórias na temporada regular de cada uma das temporadas avaliadas. A temporada com a maior correlação foi 2016, seguido pelo ano de 2013. Já a temporada com a menor correlação observada foi 2015. De modo geral, a correlação média foi 0.7399.

A Tabela 4.5 mostra um resultado muito importante. Com base nas habilidades estimadas $\hat{\alpha}_i$ para cada equipe, foi possível identificar em cada partida qual

equipe seria classificada como a mais habilidosa. Desta forma, é possível estimar a partir das habilidades relativas $\alpha_{h(k)}$ e $\alpha_{a(k)}$ a probabilidade do pior time da partida, o *underdog*, vencer a partida. Essa probabilidade é representada pelo termo $P(U)$. Observa-se a partir da tabela que esse valor é estável ao longo das temporadas e vale aproximadamente 0.36. O modelo gráfico probabilístico desenvolvido também permite estimar a probabilidade do pior time vencer dado que ele joga fora de casa ($P(U|A)$) e em casa ($P(U|H)$). A conclusão neste caso é que jogar como mandante aumenta a probabilidade do time da casa *underdog* vencer com relação ao time visitante *underdog* em aproximadamente $0.45 - 0.27 = 0.18$, uma vantagem substancial.

Tabela 4.5. Probabilidades condicionais e não-condicionais dado o modelo da pior equipe da partida (*underdog*) vencer.

Temporada	$P(U)$	$P(U\|A)$	$P(U\|H)$	$P(U\|A, R+)$	$P(U\|H, R+)$
2012	0.35	0.27	0.44	0.25	0.24
2013	0.36	0.25	0.47	0.16	0.13
2014	0.37	0.29	0.45	0.15	0.14
2015	0.37	0.30	0.44	0.22	0.20
2016	0.34	0.26	0.43	0.19	0.13
Média	0.36	0.27	0.45	0.19	0.17

Por fim, calculou-se a probabilidade da equipe *underdog* vencer quando a equipe adversária é uma das equipes que deveriam ser removidas para a liga se tornar aleatória de acordo com o coeficiente ϕ estimado na Seção 4.2. Nessa etapa de remoção, de acordo com a metodologia adotada, são removidas iterativamente as melhores ou piores equipes da temporada até que a liga se torne aleatória. Nos resultados mostrados na Tabela 4.5 são consideradas somente as equipes removidas e que estão as melhores. Essas probabilidades foram denotadas como $P(U|A, R+)$ e $P(U|H, R+)$. As probabilidades deveriam obviamente decrescer devido ao fato das equipes $R+$ serem as mais habilidosas do campeonato na fase regular e isto é mostrado na Tabela. A vantagem da equipe mandante praticamente desaparece neste caso. Nota-se inclusive que nessa situação as equipes *underdog*'s visitantes possuem uma ligeira vantagem sob as equipes *underdog*'s mandantes, mas frisa-se que essas proporções possuem uma maior variabilidade devido ao fato delas serem calculadas com uma pequena quantidade de jogos.

Capítulo 5

Conclusão

A proposta desse trabalho era estudar o papel da sorte e da habilidade em competições esportivas. Para cumprir esse objetivo, selecionou-se algumas temporadas/ligas de basquete, handebol, futebol e voleibol para desenvolver dois estudos: identificar qual componente tem maior influência nos resultados de uma liga e estimar as habilidades das equipes de uma temporada/liga.

O primeiro estudo propõe o coeficiente ϕ, que é uma forma de identificar qual componente influência mais os resultados da temporada. Esse coeficiente possibilitou identificar o basquete como o esporte mais competitivo quando comparado com voleibol, futebol e handebol. Além disso, apesar da pontuação final do voleibol ser medida em sets, que são limitados a 5 por partida e no máximo 3 por equipe, a longa sequência de pontos necessários durante a partida para obter esses sets fazem com que esse esporte seja mais competitivo que o futebol e o handebol. Uma outra contribuição foi a identificação de quais equipes deveriam ser removidas de uma temporada/liga para torna-lá aleatória. No basquete e no voleibol, por exemplo, precisa-se remover aproximadamente 50% e 40% das equipes da liga, enquanto no futebol basta remover 20% das equipes para tornar a temporada/liga aleatória.

O segundo estudo desenvolvido foi um modelo bayesiano para estimar as habilidades das equipes de uma liga. Essas ligas são aquelas em que o coeficiente ϕ foi positivo e ficou de fora do intervalo de confiança. Esse modelo foi ajustado em temporadas liga americana de basquete, a NBA. A NBA foi selecionada devido

a grande quantidade de dados disponíveis e seu coeficiente ϕ próximo de 1. O modelo 1 mostrado na Tabela 4.4 alcança uma correlação média de 0.7399 entre as habilidades das equipes estimadas e o número de vitórias durante a temporada regular.

Analisando as variáveis explicativas relevantes no modelo 1, o resultado encontrado sugere que uma estratégia para gerenciar equipes de basquete é investir na formação de equipes menores (SI), que permitem pagar altos salários para bons jogadores ($A5$). Salários mais altos também atraem jogadores de bom desempenho, o que eleva o PER médio (AP) e evita que jogadores da equipe procurem por oportunidades em equipes adversárias (RC). Equipes pequenas, mas coesas possuem menos conflitos e são mais fáceis de administrar.

A partir desses dois estudos foi encontrado que, nas temporadas da NBA, 35% das partidas serão vencidas pelas equipes menos habilidosas da partida, os chamados *underdogs*. Além disso, foi estimada que a vantagem por ser o mandante da partida adiciona uma vantagem para os *underdogs* de 0.18. A probabilidade média dos *underdogs* vencerem quando jogam como visitantes é de 0.27.

A mensagem final deste trabalho é que embora a sorte dificulte a modelagem e previsão de resultados esportivos, é ela que traz a alegria para a audiência e torna os esportes interessantes. Além disso, esse estudo mostra quais esportes são mais suscetíveis a eventos aleatórios e propõe uma forma de estimar as habilidades das equipes de uma temporada quando o componente habilidade possui uma maior influência nos resultados.

5.1 Trabalhos Futuros

O estudo de dados esportivos é uma área desafiadora e em grande expansão. Uma possível expansão desse trabalho é adaptar o coeficiente ϕ para competições que não seguem o sistema de pontos corridos. Dessa forma, muitos campeonatos tradicionais, como os campeonatos de futebol de países da América Latina, poderão ser adicionados a base de dados.

Outra possibilidade é tentar novas abordagens no modelo Bayesiano. Essa classe de modelos é muito flexível e permite inúmeras modelagens, e alguma delas

pode trazer um melhoramento nas estimativas do modelo e um melhor entendimento das habilidades das equipes.

Uma terceira expansão desse trabalho é acrescentar mais esportes nas análises do coeficiente ϕ e do modelo, para que seja possível comparar cada vez mais um universo maior de esportes. Além disso, é de interesse procurar por novas variáveis explicativas para adicionar no modelo Bayesiano e tentar melhorar as estimativas das habilidades das equipes.

Por fim, como último trabalho futuro, deseja-se investigar a influência e o comportamento de um efeito que ocorre principalmente no basquete e no voleibol conhecido como *Hot Hand*. Esse efeito faz com que um jogador ou uma equipe que está marcando pontos continue a marcar continuamente e influência diretamente na previsão dos resultados de partidas esportivas [Ayton & Fischer, 2004; Xu & Harvey, 2014; Wardrop, 1999]

Referências Bibliográficas

Aoki, R.; Assunção, R. & de Melo, P. V. (2016). Medindo o tamanho da caixinha de surpresas em ligas de futebol. *Simpósio Brasileiro de Banco de Dados*.

Ayton, P. & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & cognition*, 32(8):1369--1378.

Ben-Naim, E.; Hengartner, N.; Redner, S. & Vazquez, F. (2013). Randomness in competitions. *Journal of Statistical Physics*, 151(3-4):458--474.

Ben-Naim, E.; NW; Vazquez, F. & Redner, S. (2007). What is the most competitive sport? *Journal of the Korean Physics Society*, 50:124--126.

Ben-Naim, E.; Vazquez, F. & Redner, S. (2006). Parity and predictability of competitions. *Journal of Quantitative Analysis in Sports*, 2(4).

Brooks, J.; Kerr, M. & Guttag, J. (2016). Developing a data-driven player ranking in soccer using predictive model weights. Em *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 49--55.

Chan, W.; Courty, P. & Hao, L. (2009). Suspense: Dynamic incentives in sports contests. *The Economic Journal*, 119(534):24--46.

Chen, S. & Joachims, T. (2016). Predicting matchups and preferences in context. *KDD*.

Chen, W.-S. & Du, Y.-K. (2009). Using neural networks and data mining techniques for the financial distress prediction model. *Expert Systems with Applications*, 36(2):4075--4086.

Chetrite, R.; Diel, R. & Lerasle, M. (2015). The number of potential winners in bradley-terry model in random environment. *arXiv preprint arXiv:1509.07265*.

Fort, R. & Maxcy, J. (2003). "competitive balance in sports leagues: An introduction". *Journal of Sports Economics*, 4(2):154–160.

Fort, R. & Quirk, J. (2011). Optimal competitive balance in a season ticket league. *Economic inquiry*, 49(2):464--473.

Gabel, A. & Redner, S. (2012). Random walk picture of basketball scoring. *Journal of Quantitative Analysis in Sports*, 8(1):1--18.

Haughton, D.; McLaughlin, M.-D.; Mentzer, K. & Zhang, C. (2015). Oscar prediction and prediction markets. Em *Movie Analytics*, pp. 37--39. Springer.

Hollinger, J. (2005). *Pro Basketball Forecast, 2005-06*. Potomac Books.

Khanin, I. (2000). *Emotions in Sport*. Human Kinetics.

Martin, T.; Hofman, J. M.; Sharma, A.; Anderson, A. & Watts, D. J. (2016). Exploring limits to prediction in complex social systems. Em *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pp. 683--694.

Merritt, S. & Clauset, A. (2014). Scoring dynamics across professional team sports: tempo, balance and predictability. *EPJ Data Science*, 3(1):4.

Miljković, D.; Gajić, L.; Kovačević, A. & Konjović, Z. (2010). The use of data mining for basketball matches outcomes prediction. Em *Intelligent Systems and Informatics (SISY), 2010 8th International Symposium on*, pp. 309--312. IEEE.

Owen, P. D. (2013). Measurement of competitive balance and uncertainty of outcome. *Handbook on the economics of professional football*, pp. 41--59.

Peel, L. & Clauset, A. (2015). Predicting sports scoring dynamics with restoration and anti-persistence. Em *Data Mining (ICDM), 2015 IEEE International Conference on*, pp. 339--348. IEEE.

Pelechrinis, K.; Papalexakis, E. & Faloutsos, C. (2016). Sportsnetrank: Network-based sports team ranking. *ACM SIGKDD Workshop on Large Scale Sports Analytics*.

Ribeiro, H. V.; Mukherjee, S. & Zeng, X. H. T. (2016). The advantage of playing home in nba: Microscopic, team-specific and evolving features. *PloS one*, 11(3):e0152440.

Shergold, A. (2015). Algerian league is so tight all 16 teams can mathematically still win the title with four rounds of matches to go.

Spiegelhalter, D. (2007). Football leagues. [http://understandinguncertainty.org/node/314; accessed 26-June-2016].

Spiegelhalter, D. J.; Best, N. G.; Carlin, B. P. & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583--639.

Tarlow, D.; Graepel, T. & Minka, T. (2014). Knowing what we don't know in ncaa football ratings: Understanding and using structured uncertainty. Em *Proceedings of the 2014 MIT Sloan Sports Analytics Conference (SSAC 2014)*, pp. 1--8. Citeseer.

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G. & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10(1):178--185.

Van Haaren, J.; Ben Shitrit, H.; Davis, J. & Fua, P. (2016). Analyzing volleyball match data from the 2014 world championships using machine learning techniques. Em *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Vaz de Melo, P. O.; Almeida, V. A.; Loureiro, A. A. & Faloutsos, C. (2012). Forecasting in the nba and other team sports: Network effects in action. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(3):13.

Vračar, P.; Štrumbelj, E. & Kononenko, I. (2016). Modeling basketball play-by-play data. *Expert Systems with Applications*, 44:58 – 66.

Wang, Q.; Zhu, H.; Hu, W.; Shen, Z. & Yao, Y. (2015). Discerning tactical patterns for professional soccer teams: an enhanced topic model with applications. Em *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2197--2206.

Wardrop, R. L. (1999). Statistical tests for the hot-hand in basketball in a controlled setting. *American Statistician*, 1:1--20.

Wikipedia (2017). 2011–12 nba season — wikipedia, the free encyclopedia. [Online; accessed 2-March-2017].

Xu, J. & Harvey, N. (2014). Carry on winning: The gamblers' fallacy creates hot hand effects in online gambling. *Cognition*, 131(2):173--180.

Zimbalist, A. S. (2002). Competitive balance in sports leagues: An introduction. *Journal of Sports Economics*, 3(2):111--121.

Capítulo 6

Anexo I

Habilidade entre as equipes

As Figuras de 6.1 à 6.5 mostram as habilidades das equipes da NBA nas ligas estudadas nesta dissertação. Os gráficos representam a habilidade da equipe mandante vencer a partida contra cada uma das demais equipes. Quanto mais clara for a cor, maior a habilidade daquela equipe com relação aos times adversários.

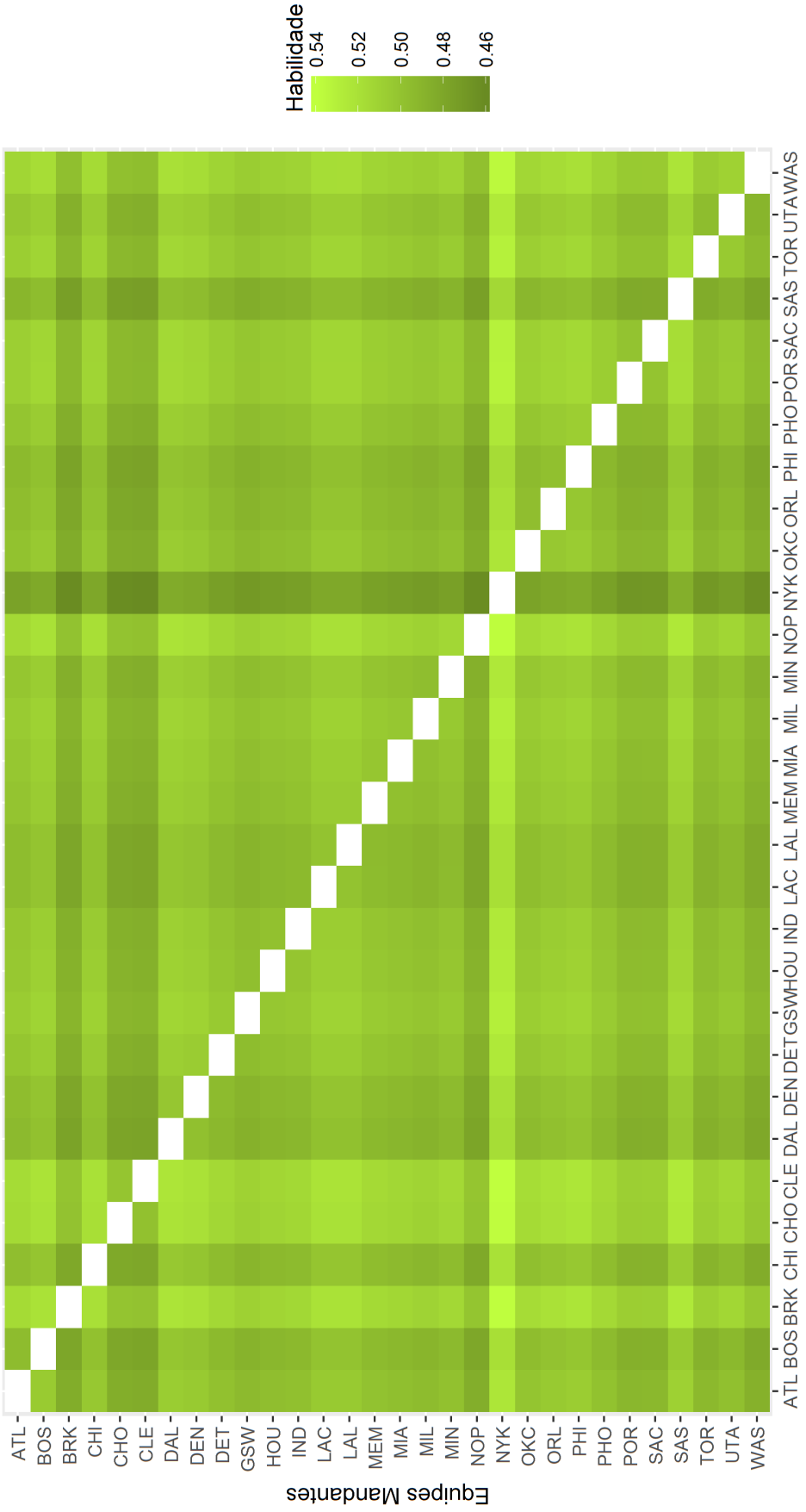

Figura 6.1. Habilidade relativa entre as equipes da temporada NBA 2012

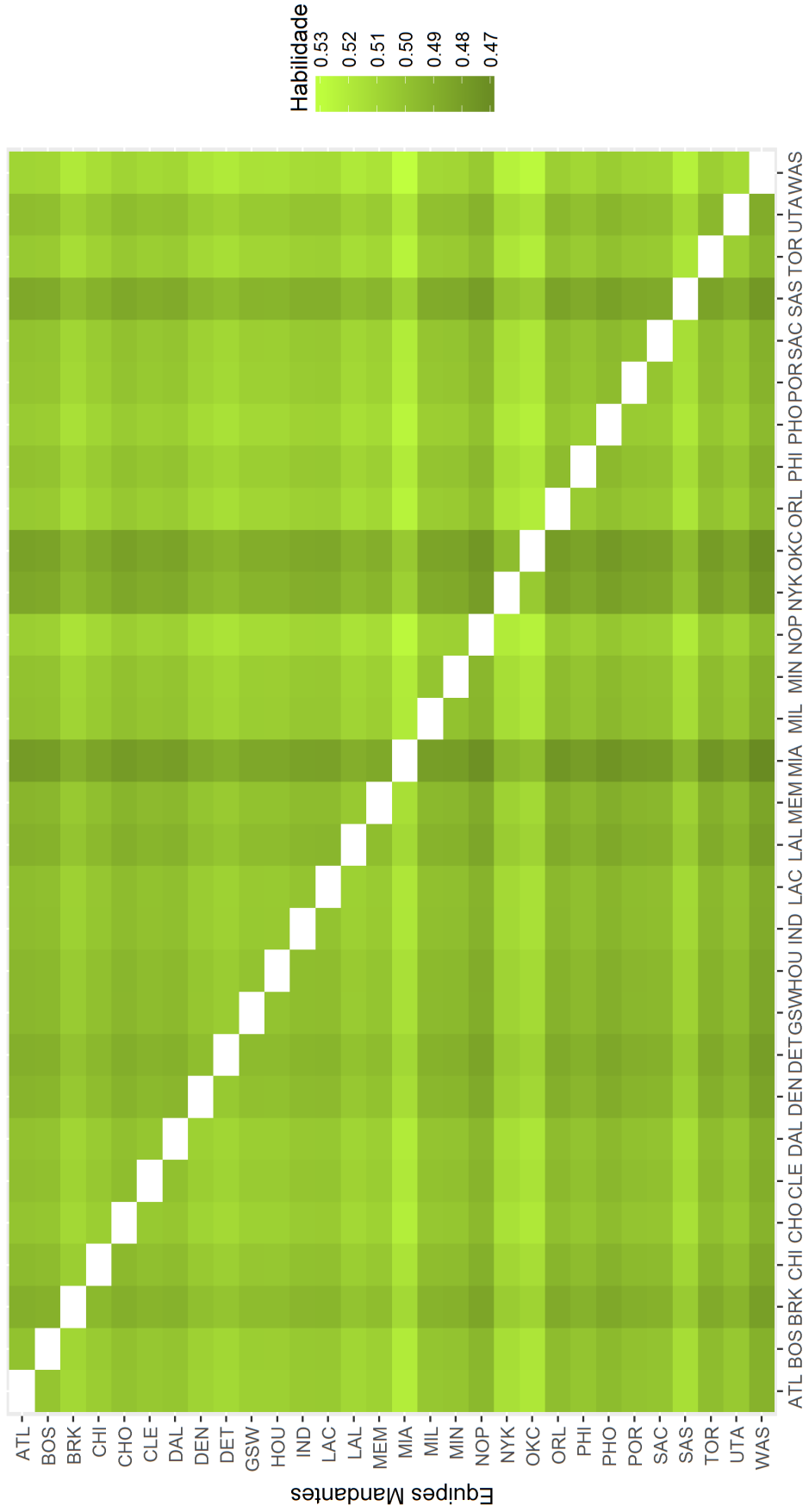

Figura 6.2. Habilidade relativa entre as equipes da temporada NBA 2013

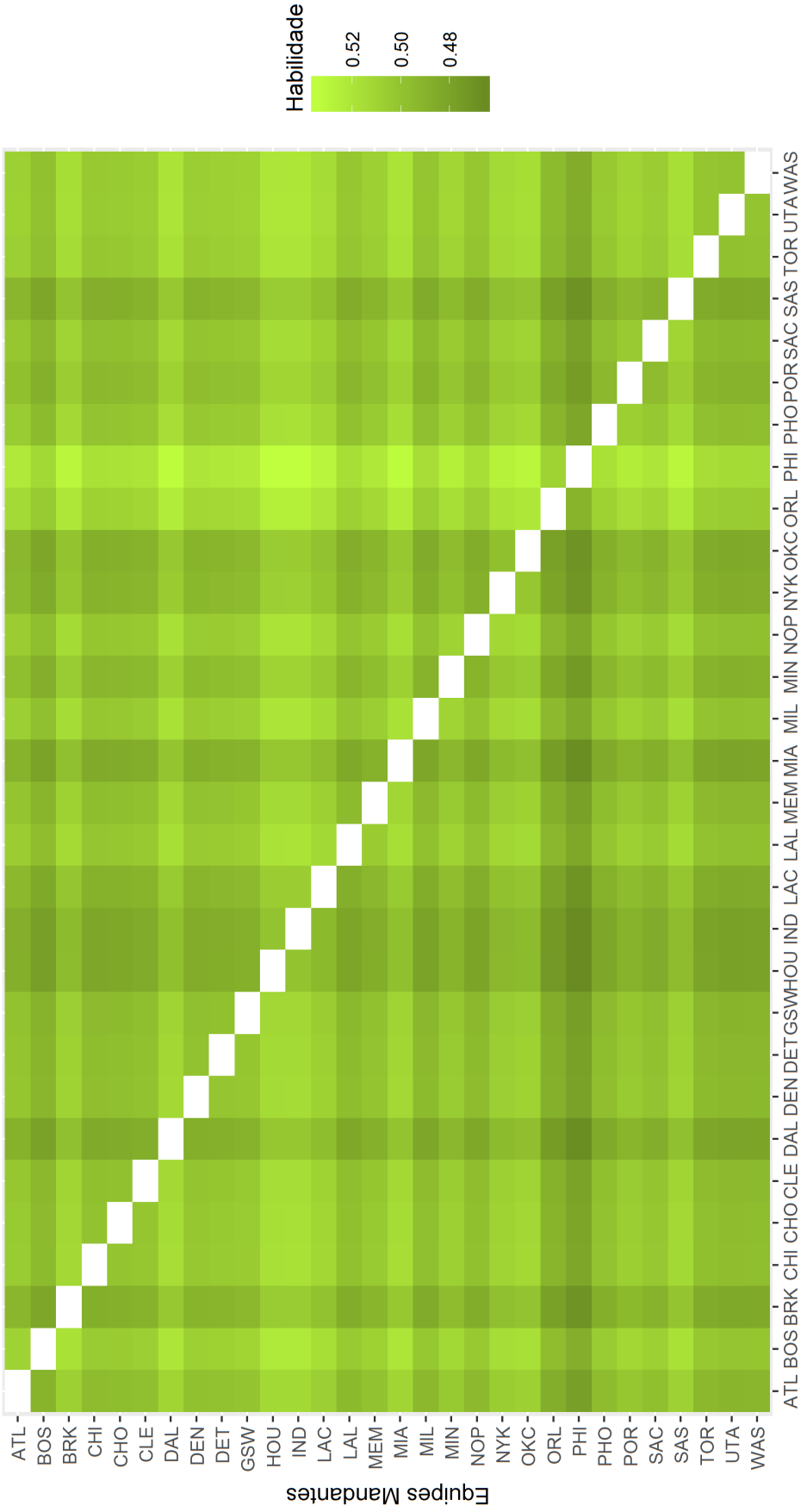

Figura 6.3. Habilidade relativa entre as equipes da temporada NBA 2014

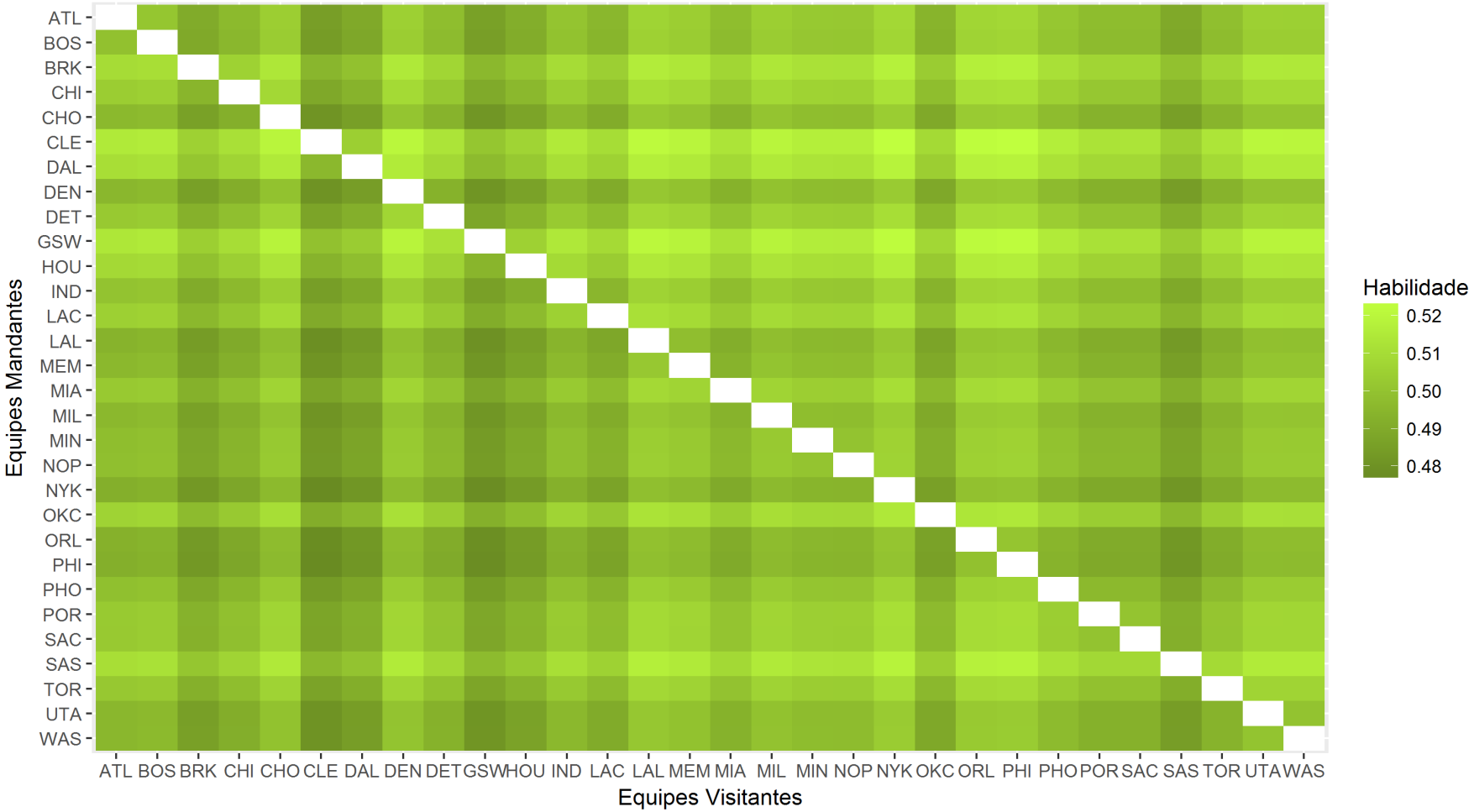

Figura 6.4. Habilidade relativa entre as equipes da temporada NBA 2015

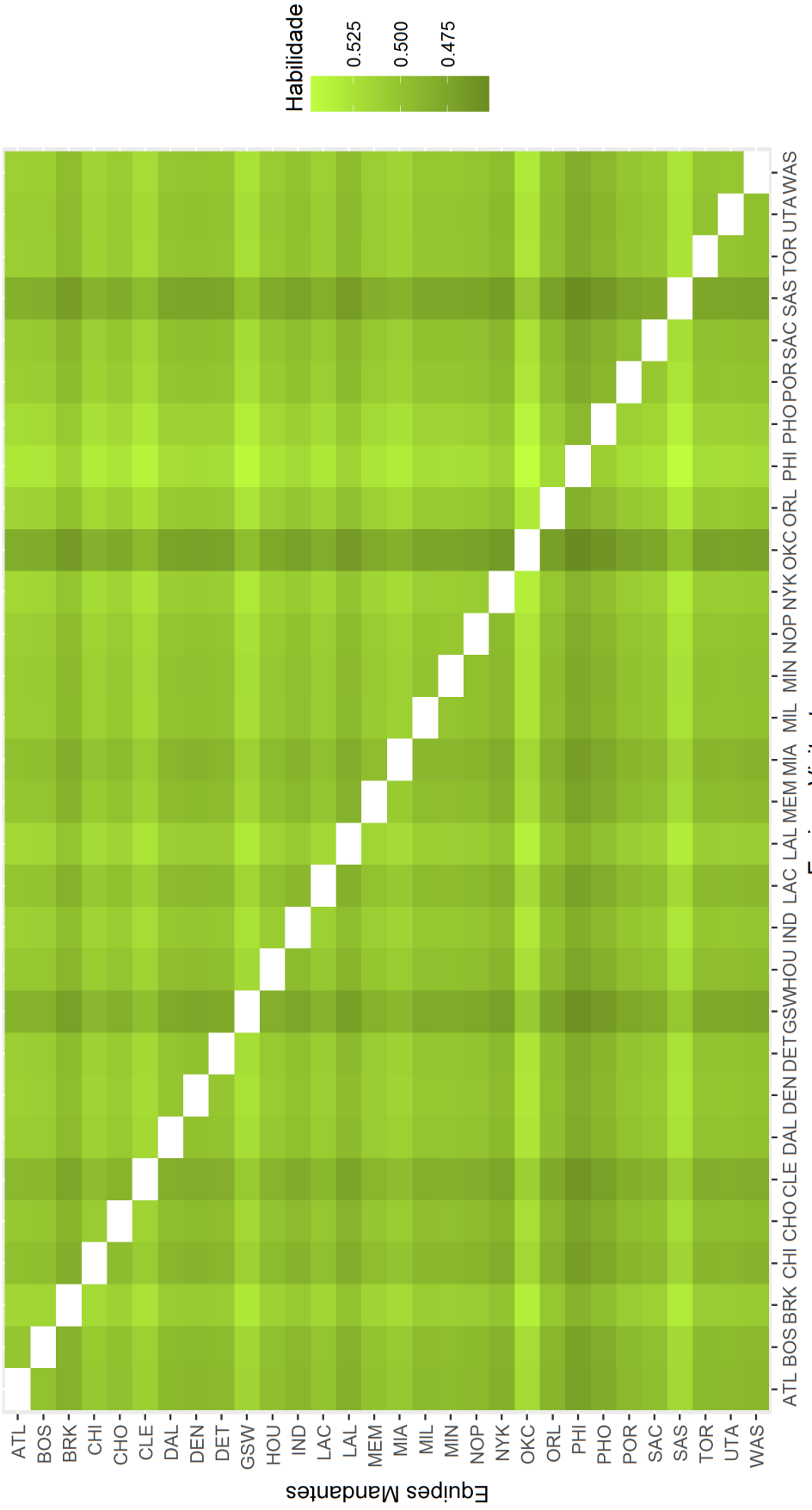

Figura 6.5. Habilidade relativa entre as equipes da temporada NBA 2016