

**Fundação Getúlio Vargas
Escola de Matemática Aplicada**

Flavio Cordeiro Fontanella

**Métodos de Avaliação de Modelos de
Previsão de Resultados de Futebol**

Rio de Janeiro
2021

Flavio Cordeiro Fontanella

**Métodos de Avaliação de Modelos de
Previsão de Resultados de Futebol**

Dissertação submetida à Escola de Matemática Aplicada como requisito parcial para a obtenção do grau de Mestre em Modelagem Matemática.

Área de Concentração: Ciência de Dados

Orientador: Moacyr Alvim Horta Barbosa da Silva

Rio de Janeiro
2021

Fontanella, Flavio Cordeiro

Métodos de avaliação de modelos de previsão de resultados de futebol /
Flavio Cordeiro Fontanella. – 2021.
98 f.

Dissertação (mestrado) -Fundação Getulio Vargas, Escola de Matemática
Aplicada.

Orientador Moacyr Alvim Horta Barbosa da Silva.
Inclui bibliografia.

1. Futebol - Previsão. 2. Futebol - Modelos matemáticos. .3. Gerenciamento de
resultados. I. Silva, Moacyr Alvim Horta Barbosa da, II. Fundação Getulio Vargas.
Escola de Matemática Aplicada. III. Título.

CDD – 796.334

FLÁVIO CORDEIRO FONTANELLA

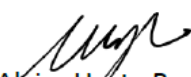
“MÉTODOS DE AVALIAÇÃO DE MODELOS DE PREVISÃO DE RESULTADOS DE FUTEBOL”.

Dissertação apresentado(a) ao Curso de Mestrado em Modelagem Matemática do(a) Escola de Matemática Aplicada para obtenção do grau de Mestre em Modelagem Matemática.

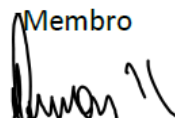
Data da defesa: 02/02/2021

ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA

Presidente da Comissão Examinadora: Prof^o Moacyr Alvim Horta Barbosa da Silva


Moacyr Alvim Horta Barbosa da Silva
Orientador



Paulo Cezar Pinto Carvalho
Membro


Gilcione Nonato Costa
Membro


Bernardo Nunes Borges de Lima
Membro

Nos termos da Lei nº 13.979 de 06/02/20 - DOU nº 27 de 07/02/20 e Portaria MEC nº 544 de 16/06/20 - DOU nº 114 de 17/06/20 que dispõem sobre a suspensão temporária das atividades acadêmicas presenciais e a utilização de recursos tecnológicos face ao COVID-19, as apresentações das defesas de Tese e Dissertação, de forma excepcional, serão realizadas de forma remota e síncrona, incluindo-se nessa modalidade membros da banca e discente.


César Leopoldo Camacho Manco
Diretor


Antonio de Araujo Freitas Junior
Pró-Reitor de Ensino, Pesquisa e Pós-Graduação FGV
Antonio Freitas, PhD
Pró-Reitor de Ensino, Pesquisa e Pós-Graduação
Fundação Getúlio Vargas

Instrução Normativa nº 01/19, de 09/07/19 - Pró-Reitoria FGV

Em caso de participação de Membro(s) da Banca Examinadora de forma não-presencial*, o Presidente da Comissão Examinadora assinará o documento como representante legal, delegado por esta I.N.

*Skype, Videoconferência, Apps de vídeo etc

Agradecimentos

Agradeço aos meus colegas, pelo companheirismo nas horas de estudo e de lazer e por toda a ajuda ao longo do curso.

Agradeço aos professores, por todo o conhecimento transmitido e pela paciência quando necessário.

Agradeço aos membros da banca e, em especial, ao meu orientador, Moacyr Alvim Horta Barbosa da Silva, pela disponibilidade, pelas contribuições e pelos bons conselhos, que possibilitaram a realização deste trabalho.

Agradeço à FGV e a toda a equipe da EMAP, que tanto ajudaram a viabilizar esse curso e tornar o ambiente tão agradável.

Agradeço ao grupo Esporte em Números, com todos os seus participantes, pela inspiração e pela motivação com que lidamos com todas as pequenas questões, a mim tão interessantes.

Agradeço a toda a minha família e amigos, por todo o suporte e todo o amor, que me trouxeram até aqui e me deram forças para alcançar meus objetivos.

Agradeço, sobretudo, a minha mulher, Natália, e a meus filhos, Nathalia e Jonathan, pelo amor e apoio incondicionais e por toda a compreensão nas longas horas de estudo, em que tive que abrir mão de atividades mais coletivas.

Obrigado!

Resumo

Este trabalho tem por objetivos revisar a teoria de avaliação de modelos de previsão, apresentando conceitos como a utilização de medidas de avaliação próprias e suas decomposições em somas de atributos distintos; adequar essa teoria ao contexto de previsões de placares e resultados de partidas de futebol, propondo medidas apropriadas ao tema; e aplicar as medidas a partidas do campeonato brasileiro de futebol masculino, entre as temporadas de 2014 e 2019, buscando avaliar diferentes modelos de previsão e inferir atributos como a incerteza dos resultados, o refinamento das previsões e a discriminação das previsões ante os resultados.

Abstract

This work main goals are to review the theory of forecasting evaluation, presenting notions like the use of proper scoring rules and their decompositions as sums of attributes; to address this theory to the context of football (soccer) match results/scores forecasting, proposing scoring rules that are suited to this environment; and to apply the scoring rules to the matches of Brazilian championship of men's football, from 2014 to 2019, evaluating different forecasting models and estimating attributes such as the uncertainty of the results, the refinement of the forecasts and the discrimination of the forecasts given the results.

Sumário

1	Introdução	7
1.1	Contexto e Problema	7
1.2	Objetivos	8
1.3	Estrutura da Dissertação	8
2	Referencial Teórico	9
2.1	Previsões de Resultados e Placares de Futebol	9
2.2	Medidas de Avaliação de Previsões	10
2.2.1	Espaço de Eventos Possíveis, Espaço de Previsões, Definição e Orientação de Medida de Avaliação de Previsões	10
2.2.2	Medida 0-1	12
2.2.3	Medida Probabilidade Média do Evento Ocorrido	13
2.2.4	Medidas que Correspondem a Distâncias no Espaço de Previsão	14
2.3	Propriedades das Medidas de Avaliação	16
2.3.1	Previsões Consistentes	17
2.3.2	Medidas Próprias, Medidas Estritamente Próprias e a Medida de Brier	18
2.3.3	Medidas Locais e a Medida de Ignorância	19
2.3.4	Medidas Sensíveis à Distância e a Medida RPS	20
2.3.5	Localidade x Sensibilidade à Distância	21
2.4	Atributos das Previsões e Decomposições de Medidas Próprias	23
2.4.1	Distribuição Conjunta de Previsões e Observações e suas Decomposições em Distribuições Condicionais e Marginais	24
2.4.2	Atributos de um Conjunto de Previsões	25
2.4.3	Um exemplo ilustrativo	26
2.4.4	Primeira Decomposição (Primária) de Medidas Próprias: Decomposição URR	32
2.4.5	Segunda Decomposição (Dual) de Medidas Próprias: Decomposição RDC	35
2.4.6	O problema da esparsidade na distribuição conjunta de previsões e observações	37
3	Metodologia	40

3.1	Primeira Abordagem: Previsões da EMaP para o Brasileiro 2019	40
3.2	Outros Modelos de Previsão	44
3.2.1	Modelos Poisson Independentes Baseados em Retrospecto de Partidas e/ou Dados Econômicos	44
3.2.2	Modelos Idealizados para Benchmarking: PIP, PIS e FREQ	46
3.3	Base de Dados Utilizada Para Avaliação dos Modelos	48
3.4	Medidas de Avaliação Mais Adequadas a Previsões de Placares e Resultados de Futebol	50
3.5	Decomposições das Medidas de Avaliação Seleccionadas	52
3.5.1	Convenções e métodos adotados para o cálculo das esperanças	53
3.5.2	Expressões das decomposições da medida 2RPS	54
4	Resultados e Análise	61
4.1	Resultados dos Diferentes Modelos	61
4.1.1	Modelos Poisson Independentes com 4 Parâmetros por Time e 1 Fator de Esquecimento	61
4.1.2	Modelos Poisson Independentes com 2 Parâmetros por Time, 1 Fator de Esquecimento e 1 Parâmetro Geral para o Fator Casa	64
4.1.3	Modelos Poisson Independentes com 2 Parâmetros por Time, 1 Fator de Esquecimento, 1 Parâmetro Geral para o Fator Casa e Utilização de Dados Econômicos	67
4.1.4	Modelo PIP e as limitações dos Modelos Poisson Independentes	71
4.2	Comparativo entre Modelos	74
4.3	Decomposições das Avaliações	87
4.3.1	Modelo FREQ e a Incerteza dos Eventos Ocorridos	87
4.3.2	Decomposições da medida 2RPS	88
5	Conclusão	93
6	Referências	96

1 Introdução

1.1 Contexto e Problema

Futebol é o esporte mais popular do mundo, com mais de 270 milhões de praticantes e 4 bilhões de espectadores ao redor do globo. [Editors, 2018] Como tal, o resultado e o placar de uma partida de futebol atraem grande interesse do público, composto, entre outros, pelos jogadores e profissionais envolvidos na partida, pelos torcedores de ambos os times, por jornalistas que cobrem o evento, por expectadores que acompanham a distância ou por aqueles que apenas consomem o noticiário relacionado. Dependendo do grau de interesse em uma partida específica, o alcance pode chegar a dezenas ou centenas de milhões de interessados.

Tentar prever o placar ou o resultado de uma partida de futebol é uma atividade há muito tempo em uso com diferentes objetivos: desde a mera recreação numa conversa entre amigos, passando por fins comerciais e profissionais para um jornalista num programa esportivo, e até no planejamento dos times, como numa decisão de poupar jogadores importantes da equipe numa partida considerada “fácil” ou “difícil demais”. Com tantos usos, é necessário destacar um como, provavelmente, aquele que mais se beneficia das previsões: o mercado de apostas esportivas. Sem as previsões, esse mercado dificilmente existiria, ou, ao menos, teria imensa dificuldade para precificar as apostas.

Dado o grande interesse e a variedade de usos para as previsões de futebol (e esportivas em geral), fica no ar a pergunta: “Como avaliar a qualidade de uma previsão?”.

É claro que essa pergunta é muito ampla e aplicável a diversos tipos de previsão, não ficando restrita a previsões esportivas e, menos ainda, especificamente a resultados e placares de partidas de futebol. Por outro lado, é de se esperar que uma resposta a essa pergunta passe não somente por uma análise geral, baseada em conceitos de matemática e estatística, como também por questões específicas relacionadas ao contexto de partidas de futebol.

Vale observar também que, para uma previsão determinística (assertiva) do tipo: “o time A vencerá por 2 a 0”, é relativamente simples avaliar a qualidade no sentido de verificar, após a realização da partida, se a asserção foi verdadeira ou não, enquanto numa previsão probabilística do tipo: “a partida tem 60% de chances de terminar empatada”, não parece haver uma maneira simples e óbvia de avaliar a previsão após o resultado ocorrido.

Afinal, essa probabilidade não representa, essencialmente, um prognóstico daquilo que vai ocorrer, mas um diagnóstico das chances estimadas para um resultado possível.

Esse é o tipo de problema que vamos abordar no presente trabalho. Mais precisamente, dado um conjunto de previsões probabilísticas de resultados e/ou de placares de partidas de futebol e dados seus respectivos resultados e/ou placares ocorridos, queremos avaliar a qualidade dessas previsões.

1.2 Objetivos

Nosso principal objetivo é definir/selecionar métodos de avaliação que possam aferir a qualidade das previsões e que permitam comparar diferentes modelos num determinado escopo e concluir qual modelo se saiu melhor. De maneira semelhante, os métodos devem ser capazes de avaliar um mesmo modelo em diferentes escopos e concluir em qual escopo ele se saiu melhor.

Uma vez alcançado esse objetivo, vamos colocar em prática os métodos adotados, passando a avaliar diferentes modelos num conjunto de partidas selecionado.

1.3 Estrutura da Dissertação

Essa dissertação está estruturada da seguinte maneira: na seção 2 fazemos uma revisão da literatura e apresentamos os conceitos com que vamos trabalhar, através de definições e teoremas; na seção 3 apresentamos nossa metodologia, com as medidas de avaliação selecionadas, os dados utilizados e o tipo de resultados e tratamentos que esperamos obter; na seção 4 mostramos os resultados, através de tabelas e gráficos formulados a partir da aplicação da metodologia ao conjunto de dados; na seção 5 temos a conclusão e as considerações finais do trabalho. A última parte é composta pelas referências aos trabalhos anteriores.

2 Referencial Teórico

2.1 Previsões de Resultados e Placares de Futebol

“Previsões probabilísticas de eventos esportivos como partidas de futebol se tornaram uma área de considerável interesse nos anos recentes. Uma razão para isso é que as previsões podem ajudar no mercado apostas, tendo o potencial de suportar a identificação de estratégias de aposta lucrativas. Previsões probabilísticas também cresceram em popularidade na imprensa esportiva. Em algumas mídias, por exemplo, a apresentação de probabilidades se tornou rotina nas discussões pré-jogo e mesmo durante as partidas.”

[Wheatcroft, 2019]

Constantinou and Fenton [2012] definiram previsão probabilística como uma sequência de probabilidades correspondentes aos eventos possíveis.

Formalmente, vamos acrescentar à definição que os eventos possíveis sejam exaustivos e mutuamente exclusivos e que, portanto, $\sum_{i=1}^d \text{prob}(x_i) = 1$, para um conjunto com d eventos possíveis.

Assim, uma previsão de resultado de uma partida de futebol é um trio ordenado (p_H, p_D, p_A) , com $0 \leq p_i \leq 1$, $p_H + p_D + p_A = 1$, onde H , D e A correspondem, respectivamente, a vitória do mandante, empate e vitória do visitante. Uma previsão de placar de uma partida de futebol é uma sequência (p_i) , com $\sum p_i = 1$, onde cada i corresponde a um placar possível.

Maher [1982] apresentou um modelo de previsão de placares de partidas de futebol, em que o número de gols marcados por cada time foi modelado como uma variável aleatória com distribuição de Poisson, independentes uma em relação a outra. O parâmetro λ de cada variável Poisson foi definido como uma função de parâmetros específicos de cada time, que representavam sua capacidade ofensiva e defensiva, e eram estimados a partir de resultados (placares) anteriores no campeonato.

Lee [1997] adotou um modelo similar ao de Maher para calcular os parâmetros ofensivos e defensivos de cada time na temporada 95/96 da *English Premier League* e o utilizou para simular a temporada de acordo com as probabilidades atribuídas pelo modelo para os resultados para cada partida.

Dixon and Coles [1997] utilizaram um modelo semelhante, com o intuito de desenvolver uma estratégia lucrativa para apostas em partidas de futebol. Eles incorporaram ao modelo uma função de esquecimento, que dava peso maior a resultados recentes na estimação dos parâmetros.

Outros modelos de previsão de partidas de futebol foram desenvolvidos ao longo do tempo, baseados em diferentes técnicas como modelos binomiais probit, sistemas de rating e aprendizado de máquinas, entre outras. [Constantinou, 2019]

2.2 Medidas de Avaliação de Previsões

“Dada a criticidade em se validar modelos de previsão, fica claro que a escolha da medida de avaliação pode ser tão importante quanto o próprio desenvolvimento do modelo.”

[Constantinou and Fenton, 2011]

2.2.1 Espaço de Eventos Possíveis, Espaço de Previsões, Definição e Orientação de Medida de Avaliação de Previsões

De modo bastante genérico, dados uma previsão p e um evento ocorrido x , uma avaliação da previsão ante o evento ocorrido é um número real $S(p, x)$, onde S é uma função real definida no espaço produto cartesiano dos espaços de previsão e de eventos possíveis, $S : \mathcal{P} \times \mathcal{E} \rightarrow \mathbb{R}$.

No caso de previsões de placares de futebol, o espaço de eventos possíveis contém qualquer placar final possível para uma partida. Teoricamente, seria $\mathbb{N} \times \mathbb{N}$, onde \mathbb{N} é o conjunto dos números naturais (incluindo o zero). Na prática, não existe tempo hábil para placares muito altos, por exemplo, acima de 100 gols numa partida. Por isso, e para possibilitar o processamento computacional, consideramos geralmente um truncamento desse espaço, para números naturais menores que um determinado limite. O espaço de previsões consiste de todas as possíveis distribuições de probabilidade para os placares possíveis, ou seja, $\mathcal{P} = \{p \in [0, 1]^d ; p_1 + \dots + p_d = 1\}$, onde d é a dimensão (cardinalidade) do espaço de placares \mathcal{E} .

Para previsões de resultados de futebol, o espaço de eventos possíveis consiste de 3 pontos, representando a vitória do mandante, o empate e a vitória do visitante. Já o espaço de previsões consiste de todas as distribuições de probabilidade para os resultados possíveis, isto é, $\mathcal{P} = \{p \in [0, 1]^3 ; p_1 + p_2 + p_3 = 1\}$ (ou $p_H + p_D + p_A = 1$, como usado anteriormente). Geometricamente, \mathcal{P} pode ser identificado como o triângulo em \mathbb{R}^3 , com vértices em $(1, 0, 0)$, $(0, 1, 0)$ e $(0, 0, 1)$, enquanto \mathcal{E} seria o conjunto desses três vértices. Essa representação é conhecida como 2-simplexo ou simplexo padrão.

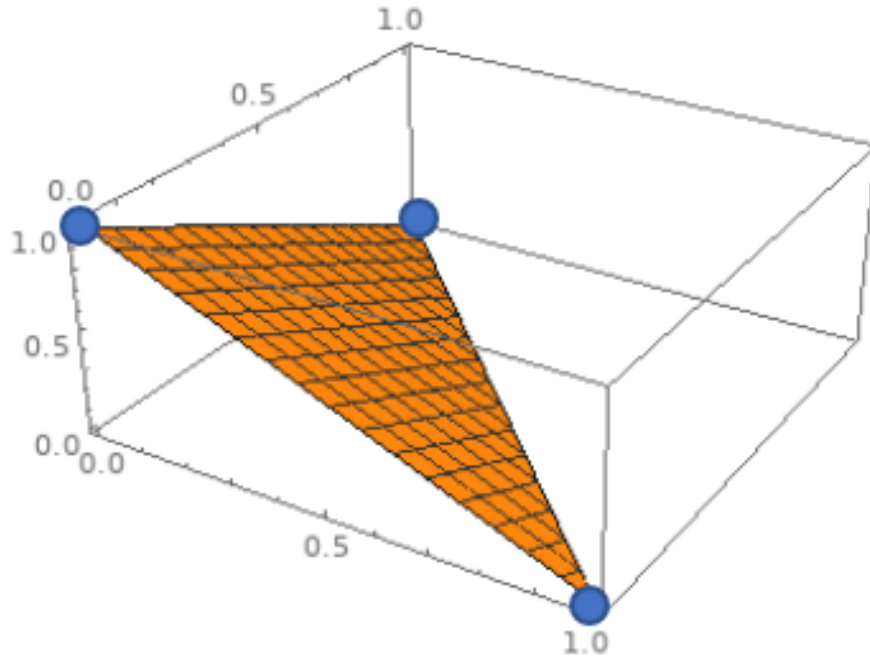


Figura 1: 2-simplexo com destaque (em azul) para os pontos correspondentes aos resultados possíveis

Chamaremos a função $S : \mathcal{P} \times \mathcal{E} \rightarrow \mathbb{R}$. de uma **medida de avaliação de previsões**, muito embora, em geral, essa função não corresponda ao objeto matemático definido na teoria da medida. A função também não é, necessariamente, uma métrica ou uma pseudométrica, embora o seja em alguns casos.

Dizemos que uma medida de avaliação de previsões tem **orientação positiva**, quando a um maior valor da medida corresponder uma melhor avaliação da previsão. A medida terá **orientação negativa** quando melhores avaliações corresponderem a menores valores da medida. [Winkler and Murphy, 1968]

É importante esclarecer também que a cada função $S : \mathcal{P} \times \mathcal{E} \rightarrow \mathbb{R}$, faremos corresponder funções $\tilde{S} : (\mathcal{P} \times \mathcal{E})^N \rightarrow \mathbb{R}$, que representarão a avaliação de um conjunto de N previsões ante os respectivos N eventos ocorridos. Em geral, a função \tilde{S} é a média aritmética das N avaliações individuais S , sendo, para alguns autores, a própria definição de \tilde{S} . Neste trabalho, vamos considerar \tilde{S} de forma mais ampla, podendo ser a soma, a média geométrica ou alguma outra função das avaliações individuais S , ou seja, dadas N partidas de futebol, $\tilde{S}((p_1, x_1), \dots, (p_N, x_N)) = G(S(p_1, x_1), \dots, S(p_N, x_N))$, para alguma função real G . Sem causar confusão (esperamos), essas funções \tilde{S} serão também chamadas de medidas de avaliação de previsões e ficará claro, no restante do trabalho, quando uma medida for definida ou utilizada, qual a função S e quais as funções \tilde{S} a que se refere.

2.2.2 Medida 0-1

A medida mais simples que vamos utilizar é a medida 0-1, que avalia com 1 uma previsão quando o placar (resultado) ocorrido corresponder à maior probabilidade da previsão, e com 0 caso contrário, i.e., $S(p, x) = 1$, se $x \equiv e_k$ e $p_k > p_i \forall i \neq k$. (e_k é o k -ésimo vetor da base canônica do espaço de previsões, e corresponde à previsão que aponta probabilidade 1 para o k -ésimo evento e 0 para todos os demais, podendo ser identificada ao próprio evento, como no caso do 2-simplexo.)

Essa medida tem orientação positiva e é popularmente conhecida como “acertar” ou “errar” a previsão, no sentido de verificar se o evento previsto como mais provável ocorreu ou não.

Para um conjunto de N previsões e resultados, a função \tilde{S} dessa medida seria a soma (ou média aritmética) dos N pares de previsão e resultado e

corresponderia à quantidade (ou proporção) de “acertos” para as N partidas.

É fácil perceber que a medida 0-1 avalia de maneira igual previsões “bem diferentes”. Por exemplo, para uma partida que termine empatada $(0, 1, 0)$, ambas as previsões $(0.1, 0.8, 0.1)$ e $(0.3, 0.4, 0.3)$ seriam avaliadas com 1 (e interpretadas como “acertos”).

Analogamente, duas previsões “semelhantes” podem ser avaliadas de maneira diferente. É o caso das previsões $(0.1, 0.46, 0.44)$ e $(0.1, 0.44, 0.46)$, em que a primeira receberia 1 (“acerto”) e a segunda 0 (“erro”).

Os termos entre aspas “bem diferentes” e “semelhantes” são relativos às probabilidades atribuídas aos resultados possíveis. É intuitivo imaginar que previsões próximas (no sentido matemático de métrica/distância em \mathbb{R}^3) deveriam ter uma avaliação próxima (no sentido de distância em \mathbb{R}). Assim, a previsão $(0.3, 0.4, 0.3)$ deveria ter uma avaliação mais próxima à avaliação de $(0.4, 0.3, 0.3)$ do que à de $(0.1, 0.8, 0.1)$.

Numa outra abordagem, a probabilidade prevista para um resultado possível pode ser interpretada como “o grau de confiança” da previsão naquele resultado. Nesse sentido, uma previsão de 0.8 para o resultado ocorrido representaria uma confiança muito maior do que uma previsão de 0.4, sendo esperado, portanto, que a avaliação da previsão $(0.1, 0.8, 0.1)$ seja melhor que a de $(0.3, 0.4, 0.3)$. Nada disso ocorre com a medida 0-1.

2.2.3 Medida Probabilidade Média do Evento Ocorrido

Uma outra medida comum, que leva em conta parte das questões levantadas anteriormente é a que avalia uma previsão com a exata probabilidade atribuída ao evento ocorrido, i.e., $S(p, x) = p_k$ se $x \equiv e_k$.

Para N partidas, a função \tilde{S} dessa medida é usualmente a média aritmética das N avaliações e corresponde à probabilidade média atribuída aos eventos ocorridos. Em alguns casos, é utilizada a média geométrica em vez da média aritmética, por motivos que serão discutidos mais adiante neste trabalho.

A medida probabilidade média do evento ocorrido tem orientação positiva.

Embora essa medida respeite as ponderações anteriores sobre distância e confiança no que diz respeito ao evento ocorrido, o mesmo não ocorre, de

modo geral, quando considerados os demais resultados possíveis.

Por exemplo, numa vitória do mandante, as previsões $(0.4, 0.1, 0.5)$ e $(0.4, 0.3, 0.3)$ recebem a mesma avaliação, que é diferente da avaliação da previsão $(0.39, 0.11, 0.5)$, muito mais próxima em \mathbb{R}^3 à primeira delas. Pode-se dizer também que a segunda previsão é de um jogo equilibrado com confiança ligeiramente superior na vitória do mandante, enquanto as duas outras previsões apresentam confiança ligeiramente superior na vitória do visitante.

Ao levar em conta somente a probabilidade prevista para o evento ocorrido, a medida deixa de avaliar como a probabilidade restante foi distribuída entre os demais resultados possíveis, o que pode ser de interesse do avaliador.

2.2.4 Medidas que Correspondem a Distâncias no Espaço de Previsão

Para definir uma medida que leve em consideração toda a distribuição de probabilidade da previsão, podemos utilizar uma distância no espaço de previsões \mathcal{P} e tomá-la entre o ponto p que representa a previsão e o ponto e_k que corresponde ao evento ocorrido.

No caso de previsão de resultados, podemos tomar a distância euclidiana em \mathbb{R}^3 , restrita ao 2-simplexo.

Por exemplo, para uma vitória do visitante, uma previsão $(0.1, 0.1, 0.8)$ teria avaliação $\sqrt{(0 - 0.1)^2 + (0 - 0.1)^2 + (1 - 0.8)^2} = \sqrt{0.06} \approx 0.245$, enquanto uma previsão $(0.4, 0.3, 0.3)$ teria avaliação $\sqrt{(0 - 0.4)^2 + (0 - 0.3)^2 + (1 - 0.3)^2} = \sqrt{0.74} \approx 0.86$.

É claro que a melhor previsão possível $(0, 0, 1)$ teria avaliação 0 e que, nessa medida, quanto menor a avaliação, melhor seria considerada a previsão, ou seja, a medida tem orientação negativa.

Para previsões de placares, poderíamos definir como medida, analogamente, uma distância em \mathbb{R}^d (onde d é a cardinalidade do espaço de placares truncado) entre a previsão e o ponto correspondente ao placar ocorrido (a previsão com probabilidade 1 para o placar ocorrido).

Alternativamente, podemos utilizar uma distância em \mathbb{R}^2 , entre o ponto correspondente ao placar (ponto $(3, 2)$ para um placar de 3 x 2 para o mandante) e um ponto representativo da previsão, como o placar médio previsto

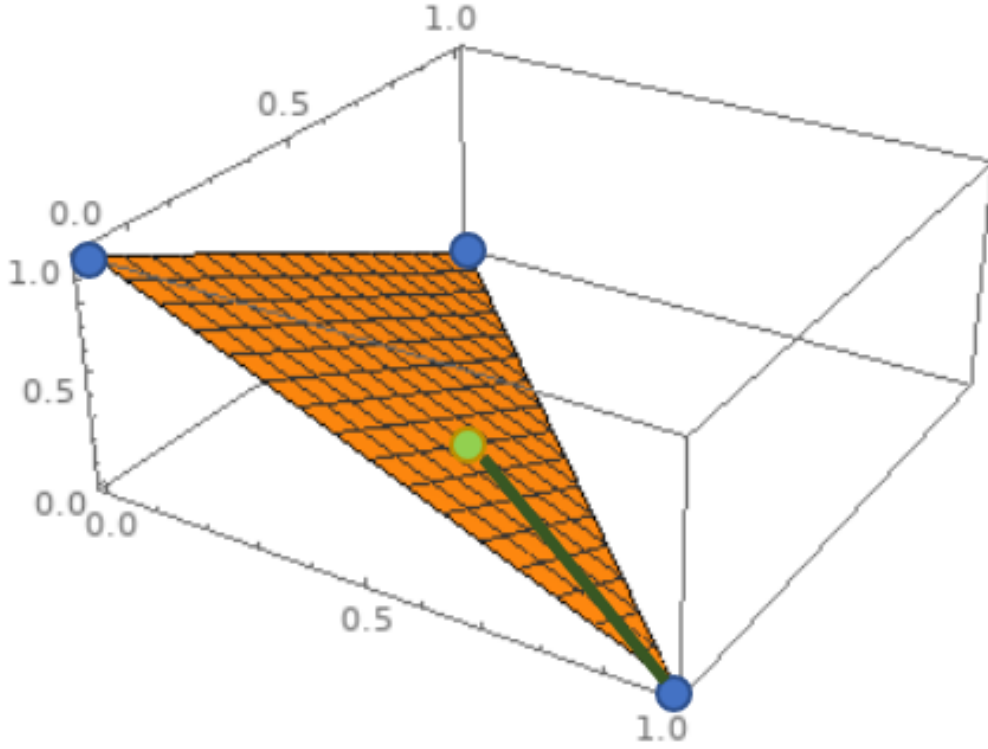


Figura 2: 2-simplexo. O ponto verde representa a previsão e o segmento verde-escuro a distância euclidiana para o resultado ocorrido.

ou o placar mais provável previsto.

Neste caso, se, por exemplo, escolhermos a distância euclidiana e a moda, para um placar ocorrido de 3 x 2, uma previsão com moda 2 x 1 teria avaliação $\sqrt{(3-2)^2 + (2-1)^2} = \sqrt{2} \approx 1.41$, enquanto uma previsão com moda 3 x 2 teria avaliação 0, a melhor possível.

Essa última medida pode ser interpretada como o “erro euclidiano” em \mathbb{R}^2 da (moda da) previsão em relação ao placar.

Em todos os casos, o erro ou distância para o placar ocorrido é medida com orientação negativa.

2.3 Propriedades das Medidas de Avaliação

“Verificação de previsões climáticas tem sido um assunto controverso por mais de meio século. Há uma diversidade de razões pelas quais esse problema tem sido tão perturbador para meteorologistas e outros, mas uma das dificuldades mais importantes parece ser chegar a um acordo de uma escala de ‘ser boa’ para previsões climáticas.”

[Brier, 1950]

“Existem diferentes opiniões entre os preditores – e entre preditores e usuários de previsões – a respeito do significado da frase ‘previsão boa (ruim)’. Essas diferenças de opinião são fruto de uma falta de clareza e/ou entendimento no que concerne a natureza de ser bom para previsões. Essa falta de clareza e entendimento atrapalha os processos de formular e avaliar previsões e prejudica sua utilidade final.”

[Murphy, 1993]

Segundo Murphy [1993], no que tange a uma previsão, a natureza de ‘ser boa’ deve ser avaliada segundo três óticas diferentes. Uma boa previsão deve ter consistência (tipo 1), característica relacionada à correspondência entre a crença do preditor e sua previsão. Uma boa previsão também deve ter qualidade (tipo 2), relacionada à correspondência entre a previsão e o resultado observado. Uma boa previsão deve ter valor (ou utilidade) (tipo 3), que é o incremento econômico e/ou de outros benefícios auferidos por tomadores de decisão a partir do uso da previsão. Neste trabalho, vamos abordar os dois primeiros tipos e procurar medidas de avaliação que sejam adequadas nesse contexto.

2.3.1 Previsões Consistentes

“Diversas medidas de avaliação foram propostas, mas um dos maiores argumentos contrários à verificação de previsões é que previsões consideradas ‘melhores’ de acordo com determinada medida podem não ser as previsões mais úteis. Tentando superar essa dificuldade, o preditor pode se ver na posição de escolher entre ignorar a medida ou deixá-la orientá-lo, ‘especulando’ ou ‘fazendo a previsão de acordo com o sistema’. Isso pode levar a previsão a ser diferente daquilo que o preditor realmente acredita que irá ocorrer (...), um critério essencial para que a verificação seja satisfatória é que a medida de avaliação não influencie o preditor de maneira indevida.”

[Brier, 1950]

Murphy [1993] define uma previsão como **consistente** quando ela reflete “a melhor crença” do preditor, ou seja, quando a probabilidade atribuída a cada resultado possível coincide com a probabilidade que o preditor acredita ter aquela possibilidade.

Por exemplo, se o evento a ser previsto é o lançamento de uma moeda de duas caras, a previsão $(1, 0)$, i.e., probabilidade 1 para “cara” e probabilidade 0 para “coroa”, é consistente para um preditor que conheça a moeda, enquanto a previsão $(0.5, 0.5)$ é consistente para um preditor que pense que a moeda é honesta.

Algumas medidas de avaliação de previsão podem estimular o preditor a não ser consistente em suas previsões. Por exemplo, suponha que, ao observar que a frequência relativa do resultado de vitória do mandante é muito maior que a dos outros resultados, alguém proponha uma medida que conceda 1 ponto caso o resultado do jogo seja vitória do mandante e a previsão tenha atribuído maior probabilidade para esse resultado que para outros, e que conceda 2 pontos caso o resultado do jogo seja empate ou vitória do visitante e a previsão tenha atribuído maior probabilidade para o resultado que de fato ocorreu. Suponha ainda que, para determinada partida, o preditor acredite que as probabilidades dos resultados sejam $(0.5, 0.4, 0.1)$. Caso o preditor relate a previsão correspondente a sua melhor crença, o valor que ele espera obter da medida da previsão será $0.5 \times 1 = 0.5$. Por outro lado, se ele

relatar uma previsão de $(0.4, 0.5, 0.1)$, diferente de sua crença, o valor que ele esperará obter da medida será $0.4 \times 2 = 0.8$. Dessa forma, a medida proposta estimula o preditor a apresentar uma previsão não consistente.

2.3.2 Medidas Próprias, Medidas Estritamente Próprias e a Medida de Brier

[Brier \[1950\]](#) propôs uma medida que, em suas palavras, “*não pode influenciar o preditor de maneira indesejada*”.

Dados uma previsão p e um evento ocorrido $x \equiv e_k$, a medida de Brier (Brier score) é definida por:

$$S(p, x) = \sum_{i=1}^d (p_i - e_{k_i})^2,$$

onde d é a dimensão (cardinalidade) do espaço de eventos possíveis.

Para N pares de previsão/observação, a medida de Brier corresponde à média aritmética das avaliações de cada par, ou seja:

$$\tilde{S}((p_1, x_1), \dots, (p_N, x_N)) = \frac{1}{N} \sum_{n=1}^N \sum_{i=1}^d (p_{n_i} - e_{k_{n_i}})^2.$$

As medidas que estimulam os preditores a serem sempre consistentes em suas previsões, como é o caso da medida de Brier, são chamadas próprias. [\[Murphy and Epstein, 1967\]](#)

Formalmente, uma medida é dita **própria** quando a melhor avaliação (esperada) é obtida pela previsão que corresponde à distribuição dos eventos.

Uma medida é dita **estritamente própria** quando a previsão correspondente à distribuição dos eventos é a única que maximiza (ou minimiza, em caso de orientação negativa) seu valor esperado. [\[Gneiting and Raftery, 2007\]](#)

Por exemplo, para o lançamento de uma moeda honesta, o melhor valor esperado da avaliação por uma medida própria será obtido pela previsão $(0.5, 0.5)$. Além disso, se a medida for estritamente própria, seu valor esperado para a previsão $(0.5, 0.5)$ não será igualado por qualquer outra previsão.

Na definição acima, cabe observar que o valor esperado é tomado em relação aos diferentes resultados possíveis, conforme sua distribuição, e que a

previsão é fixa para o cálculo da esperança. Assim, para o lançamento de duas moedas, a melhor previsão (fixa) em valor esperado é $(0.5, 0.5)$, embora, após os lançamentos ocorridos, a previsão $(1, 0)$ possa receber avaliação melhor, caso os dois resultados observados sejam “cara”. Além disso, após uma série de eventos, se a previsão não estiver fixa, a avaliação pode ser melhor que essa esperança se cada previsão individual (ou uma grande parte delas) estiver “mais próxima” do respectivo resultado.

Das medidas citadas anteriormente, a que considera a média geométrica das probabilidades atribuídas aos resultados ocorridos e as que consideram uma distância entre o ponto que representa a previsão e o ponto que corresponde ao resultado/placar ocorrido são estritamente próprias. A quantidade ou a média de acertos é própria, mas não estritamente. Já a média aritmética das probabilidades atribuídas aos resultados ocorridos não é própria. Por exemplo, num torneio em que 60% das partidas têm vitória do mandante, 20% empate e 20% vitória do visitante, um modelo que preveja $(0.6, 0.2, 0.2)$ para todos os jogos terá avaliação $0.44 (= 0.6^2 + 0.2^2 + 0.2^2)$, pior que $0.52 (= 0.6 \times 0.8 + 0.2 \times 0.1 + 0.2 \times 0.1)$, obtida por um modelo que preveja $(0.8, 0.1, 0.1)$ para todos os jogos.

2.3.3 Medidas Locais e a Medida de Ignorância

Uma medida é dita **local** quando leva em consideração somente a probabilidade prevista para o evento que ocorreu de fato.

Good [1952] escreveu: “(...) $\log p$ (ou $\log (1-p)$) é uma medida do mérito de uma estimativa probabilística (...) e representa a quantidade de informação perdida por não se saber com certeza o que vai ocorrer.”

Winkler and Murphy [1968] trataram a medida definida por Good como medida logarítmica e mostraram se tratar de uma medida própria. Bernardo [1979] demonstrou que toda medida própria e local é uma transformação afim da medida logarítmica. Roulston and Smith [2002] atacaram o problema da avaliação de previsões através da teoria da informação e introduziram a medida de ignorância como “uma medida logarítmica que é uma versão modificada da entropia relativa e pode ser calculada para previsões probabilísticas e os eventos realizados.”. Eles definiram a medida de ignorância utilizando um logaritmo de base 2, o que corresponde, na teoria, a medir a informação em bits.

É interessante observar que a medida de ignorância não se relaciona com

a probabilidade média do evento ocorrido, mas, sim, com a média geométrica dessas probabilidades. A partir da medida de ignorância (em bits), é possível calcular a média geométrica das probabilidades dos eventos ocorridos da seguinte maneira:

$$\begin{aligned} \text{média geométrica } (p_1, \dots, p_N) &= \left(\prod_{n=1}^N p_{n_k} \right)^{\frac{1}{N}} = 2^{\log_2(\prod_{n=1}^N p_{n_k})^{\frac{1}{N}}} = \\ &= 2^{\frac{1}{N} \sum_{n=1}^N \log_2(p_{n_k})} = 2^{\text{ignorância}(p_1, \dots, p_N)}, \end{aligned}$$

onde k denota a k -ésima coordenada do vetor p_n e corresponde à probabilidade atribuída pela previsão ao evento ocorrido $x_n \equiv e_{k_n}$.

Assim, sendo a exponencial uma função estritamente crescente, podemos garantir que, para qualquer conjunto de modelos de previsão, seu ranqueamento baseado na medida de ignorância será o mesmo se baseado na média geométrica das probabilidades dos eventos ocorridos.

2.3.4 Medidas Sensíveis à Distância e a Medida RPS

“Considere previsões probabilísticas para quatro classes de temperatura: $T \leq 0F$; $0F < T \leq 20F$; $20F < T \leq 40F$; e $T > 40F$. Se duas previsões forem $(0.5, 0.3, 0.1, 0.1)$ e $(0.1, 0.3, 0.5, 0.1)$ e a última categoria, $T > 40F$, for observada, todas as medidas próprias existentes avaliariam de forma igual as duas previsões. A maioria concordaria, entretanto, que a segunda previsão teria sido, de certa forma, melhor do que a primeira. Essa conclusão se baseia na noção de que as classes 3 e 4 estão mais ‘próximas’ que as classes 1 e 4. O conceito de ‘distância’, nesse sentido, não existe em qualquer das medidas próprias propostas até hoje.”

[Epstein, 1969]

Uma medida é dita **sensível à distância** quando leva em consideração a proximidade (ou ordenação) entre os diferentes eventos possíveis. Por exemplo, numa partida terminada em 0 x 0, uma medida sensível à distância

avaliará melhor uma previsão com probabilidade concentrada no placar de 1 x 0 do que outra com igual concentração no placar de 2 x 0, respeitando a ordenação em \mathbb{R} , em que o 0 está mais próximo do 1 do que do 2.

Para previsões de resultados de futebol, uma medida sensível à distância considera o empate localizado entre os dois resultados de vitória, i.e., ambos os resultados de vitória estão mais próximos do empate do que entre si. Assim, para um resultado de vitória do mandante, a previsão (0.6, 0.3, 0.1) terá avaliação melhor que a previsão (0.6, 0.1, 0.3).

Não é possível uma medida ser local e sensível à distância, mas existem medidas que não possuem nenhuma das duas características, como a medida 0-1 ou as distâncias no 2-simplexo. Das medidas descritas anteriormente, somente o erro euclidiano em \mathbb{R}^2 (da moda ou da média) da previsão em relação ao placar é sensível à distância.

Epstein [1969] definiu uma medida que é própria e sensível à distância e a denominou 'medida de probabilidade ordenada' ('ranked probability score').

Em notação mais atual e simplificada, Constantinou and Fenton [2012] definiram a medida de probabilidade ordenada (RPS) de uma previsão como:

$$RPS(p, x) = \frac{1}{d-1} \sum_{i=1}^{d-1} \left(\sum_{j=1}^i (p_j - e_{k_j}) \right)^2,$$

onde d é a cardinalidade do espaço de eventos possíveis, ordenado em d classes, p_j é a probabilidade prevista para o evento da j -ésima classe e e_{k_j} é o indicador de ter ocorrido o evento da j -ésima classe (1 para o evento observado, 0 para os demais).

Para N instâncias (N partidas), a RPS é dada como a média aritmética das N avaliações individuais.

A RPS tem orientação negativa e generaliza a distância euclidiana em \mathbb{R} , no seguinte sentido: dada uma previsão pontual, i.e., probabilidade 1 para um número real y e 0 para todos os demais, e um evento ocorrido $x \in \mathbb{R}$, a medida RPS da previsão será $\frac{1}{d-1}|x - y|$.

2.3.5 Localidade x Sensibilidade à Distância

Uma vez que as propriedades de localidade e sensibilidade à distância são mutuamente exclusivas, surge naturalmente a questão: “Qual das duas características é mais relevante para a medida de avaliação?”

Essa pergunta é de difícil resposta e a indecisão tem sido abordada ao longo dos anos nas publicações sobre o assunto.

Wheatcroft [2019] escreveu: *“Diversas medidas de avaliação de previsões foram definidas ao longo dos anos e há muita discussão sobre qual a mais apropriada. Uma abordagem comum para escolher entre diferentes medidas é identificar propriedades desejadas e favorecer as medidas que as possuam. Contudo, há também debate sobre quais propriedades são (mais) importantes e, com isso, a falta de consenso prossegue. Como consequência, em campos como o de previsões climáticas, comumente se apresenta uma grande variedade de medidas de avaliação.”*

Ele fez ainda a seguinte ponderação: *“Murphy [1970] comparou a formulação da RPS e da medida de Brier e recomendou que a RPS fosse utilizada quando o evento de interesse é ordenado. Bernardo [1979], por outro lado, comentou que ‘ao avaliar a validade de uma conclusão científica, apenas a probabilidade atribuída a um pequeno intervalo contendo o valor observado deve ser levada em conta’, apontando localidade como uma propriedade desejada.”*

Constantinou and Fenton [2012], por sua vez, consideraram que *“(...) as diversas medidas utilizadas anteriormente para avaliar modelos (de previsão de resultados de futebol) são inadequadas, uma vez que falham em reconhecer que os resultados de futebol estão numa escala ordenada. Isso gera dúvidas em relação à validade das conclusões desses estudos. Existe uma conhecida medida de avaliação de previsões, a medida de probabilidade ordenada (RPS), não utilizada em estudos anteriores, que avalia adequadamente modelos de previsão de resultados de futebol.”*, dando clara preferência à sensibilidade à distância no contexto que estamos tratando.

2.4 Atributos das Previsões e Decomposições de Medidas Próprias

“Medidas de avaliação de previsões têm sido formuladas com variedade de objetivos e para uma diversidade de situações. (...) Como consequência, medidas de avaliação têm proliferado, e pouco esforço tem sido feito para desenvolver conceitos e princípios gerais, visando a investigar as relações entre medidas ou a examinar seus relativos pontos fortes ou fracos. Esta situação reprime o desenvolvimento de uma ciência de avaliação de previsões e limita a utilidade de muitos dos seus conceitos e métodos.”

[Murphy and Winkler, 1987]

Uma previsão pontual, após a realização do evento, pode ser verificada como certa ou errada. Um conjunto de previsões pontuais e seus respectivos eventos ocorridos, apresenta outros atributos, como, por exemplo, a proporção de acertos quando a previsão corresponde a um ponto particular do espaço de previsões.

Para uma previsão probabilística, a classificação em certa ou errada é menos óbvia. Uma previsão de $(1/3, 1/3, 1/3)$ para o resultado de uma partida de futebol dificilmente poderia ser considerada certa ou errada qualquer que fosse o resultado ocorrido.

Um conjunto de previsões probabilísticas, entretanto, apresenta atributos compatíveis aos de um conjunto de previsões pontuais. Dada uma determinada previsão p , i.e., um ponto particular do espaço de previsões \mathcal{P} , é possível verificar, dentre as partidas que tiveram aquela previsão, o quanto os eventos observados corresponderam a essa previsão. Por exemplo, para um conjunto de previsões de resultados de futebol, vamos supor que a previsão $(0.8, 0.1, 0.1)$ tenha ocorrido dez vezes. Se, nessas dez partidas, o resultado de vitória do mandante ocorreu apenas duas vezes, essa previsão em particular (e possivelmente o conjunto de previsões) não pode ser considerada muito confiável.

De maneira análoga, se olharmos apenas o subconjunto de partidas que tiveram um determinado resultado, por exemplo, vitória do mandante, podemos verificar o quanto as previsões para essas partidas estiveram próximas a esse resultado.

2.4.1 Distribuição Conjunta de Previsões e Observações e suas Decomposições em Distribuições Condicionais e Marginais

“É necessário adotar um modelo geral para a avaliação de previsões. (...) A base para o modelo descrito aqui é a distribuição conjunta de previsões e observações, que contém toda a informação que é relevante. (...) e se torna mais acessível através de sua decomposição. Toda distribuição conjunta pode ser decomposta em distribuições condicional e marginal de duas maneiras diferentes. Assim, considerando ambas as decomposições, obtemos dois tipos de distribuições condicionais e dois tipos de distribuições marginais, cada tipo relacionado a aspectos particulares da avaliação.”

[Murphy and Winkler, 1987]

Sejam P e X variáveis aleatórias representando, respectivamente, previsões e eventos possíveis. Dado um conjunto de partidas, com seus respectivos resultados (placares) e previsões, a distribuição conjunta de previsões e resultados (placares) é uma função $f(P, X)$ que atribui, a cada previsão $P = p$ e evento observado $X = x$, a frequência relativa de sua ocorrência conjunta, i.e., a razão entre o número de ocorrências de (p, x) e a cardinalidade N do conjunto de partidas.

A distribuição conjunta $f(P, X)$ pode ser decomposta de duas maneiras, considerando suas distribuições condicionais e marginais:

$$i. \quad f(P, X) = f(X|P) f(P)$$

$$ii. \quad f(P, X) = f(P|X) f(X)$$

A distribuição condicional $f(X|P=p)$ é uma função de X , definida como a frequência relativa de cada evento x quando a previsão é igual a p (fixa). De maneira análoga, $f(P|X=x)$ é a frequência relativa de cada previsão p quando o evento é igual a x (fixo) e, portanto, uma função de P .

As distribuições marginais $f(P)$ e $f(X)$ correspondem às frequências relativas individuais de cada variável (P ou X) no conjunto de partidas.

2.4.2 Atributos de um Conjunto de Previsões

“A perspectiva oferecida pela abordagem orientada a distribuições revela que a qualidade de uma previsão é inerentemente multifacetada. Por exemplo, aspectos da qualidade geralmente referenciados como confiabilidade e resolução podem ser avaliados pelo exame das distribuições condicional $f(X|P)$ e marginal $f(P)$ (da distribuição conjunta $f(P, X)$). Confiabilidade está relacionada à correspondência entre a média das observações associadas a uma previsão particular ($\bar{X}|P=p$) e essa previsão (p), ponderada por todas as previsões. (...) Resolução está relacionada à diferença entre a mesma observação média condicional ($\bar{X}|P=p$) e a observação média incondicional (\bar{X}), novamente ponderada por todas as previsões.”

[Murphy, 1993]

A distribuição conjunta de previsões e eventos ocorridos e suas decomposições em distribuições condicionais e marginais podem ser utilizadas para caracterizar atributos relevantes das previsões.

Murphy [1993] apresentou uma lista de atributos, com suas respectivas definições e distribuições associadas, dos quais queremos destacar os constantes da Tabela 1. Vale destacar que alguns termos na tabela - como refinamento e correção - são diferentes daqueles utilizados por Murphy na definição dos mesmos atributos, refletindo uma terminologia mais atual para esses conceitos.

Murphy utilizou termos como ‘*diferença*’, que nos permite calcular diretamente esses atributos através das médias e ocorrências individuais das previsões e observações, mas também termos como ‘*correspondência*’ e ‘*variabilidade*’, que não são definidos matematicamente de maneira única. Assim, embora possamos inferir maneiras de calcular cada atributo individualmente, pouco poderíamos afirmar a respeito da comparação entre grandezas relativas a cada atributo.

Tabela 1: Atributos de um conjunto de pares previsão-observação

Atributo	Definição	Distribuições Relevantes
Acurácia	Correspondência média entre pares individuais de previsões e observações	$f(P, X)$
Incerteza	Variabilidade das observações	$f(X)$
Refinamento	Variabilidade das previsões	$f(P)$
Resolução	Diferença entre observação média condicional e observação média incondicional, ponderada por todas as previsões	$f(X P)$ e $f(P)$
Confiabilidade	Correspondência entre observação média condicional e previsão condicionante, ponderada por todas as previsões	$f(X P)$ e $f(P)$
Discriminação	Diferença entre previsão média condicional e previsão média incondicional, ponderada por todas as observações	$f(P X)$ e $f(X)$
Correção	Correspondência entre previsão média condicional e observação condicionante, ponderada por todas as observações	$f(P X)$ e $f(X)$

Murphy [1973] exibiu uma decomposição da medida de Brier - uma estimativa da *acurácia* das previsões - como soma de três outros atributos: *incerteza*, *(-)resolução* e *confiabilidade*. Bröcker [2009] mostrou que essa decomposição pode ser estendida a qualquer medida própria de avaliação de previsões. Uma segunda decomposição da medida de Brier foi apresentada por Murphy and Winkler [1987] como soma de três outros atributos das previsões: *refinamento*, *(-)discriminação* e *correção*.

2.4.3 Um exemplo ilustrativo

Vamos começar com um sistema simplificado e efetuar os cálculos de acurácia e suas decomposições.

Três modelos - A, B e C - fazem previsões para o resultado – *cara* ou *coroa* – do lançamento de quatro moedas. As previsões e os resultados observados

são mostrados na tabela a seguir, onde o valor da previsão pode ser entendido como a probabilidade prevista para o resultado *cara*.

Tabela 2: Sistema simplificado de previsões e resultados

Lançamento	Previsão A	Previsão B	Previsão C	Observação
moeda 1	0.5	1	1	1
moeda 2	0.5	0	0	0
moeda 3	0.5	1	0.8	0
moeda 4	0.5	0	0.2	1

A medida própria utilizada para comparar o desempenho dos modelos é a medida de Brier (que neste caso binário é equivalente à medida RPS), cujas decomposições URR e RDC foram estabelecidas por [Murphy \[1973\]](#) e [Murphy and Winkler \[1987\]](#).

Para o modelo A, temos:

$$\begin{aligned}\mathbb{E}_{P,X}[S(P, X)] &= (2((1 - 0.5)^2 + (0 - 0.5)^2 + (0 - 0.5)^2 + (1 - 0.5)^2))/4 = \\ &= (0.25 + 0.25 + 0.25 + 0.25)/2 = 0.5\end{aligned}$$

Para o modelo B:

$$\begin{aligned}\mathbb{E}_{P,X}[S(P, X)] &= (2((1 - 1)^2 + (0 - 0)^2 + (0 - 1)^2 + (1 - 0)^2))/4 = \\ &= (0 + 0 + 1 + 1)/2 = 1\end{aligned}$$

Para o modelo C:

$$\begin{aligned}\mathbb{E}_{P,X}[S(P, X)] &= (2((1 - 1)^2 + (0 - 0)^2 + (0 - 0.8)^2 + (1 - 0.2)^2))/4 = \\ &= (0 + 0 + 0.64 + 0.64)/2 = 0.64\end{aligned}$$

Comparando os modelos A e B, vemos que ambos apresentam a mesma probabilidade média (aritmética) atribuída ao resultado observado (0.5), sendo o modelo A constante (sempre apresenta a mesma previsão) e o modelo B totalmente assertivo (sempre apresenta previsões pontuais). A média geométrica das probabilidades atribuídas aos resultados observados é 0.5 para o modelo A e 0 para o B, enquanto a acurácia, pela medida de Brier, é 0.5 para o A e 1 para o B.

Quando adicionamos o modelo C às comparações, temos uma probabilidade média aritmética maior (0.6), mas a média geométrica (0.447) e a acurácia (0.64) ficam entre os outros dois modelos, sendo melhores que o B e piores que o A. Para entender por que o ranqueamento é diferente ao considerar essas medidas, vamos estimar os atributos das previsões dos modelos, através das decomposições da medida de Brier.

A decomposição URR da medida de Brier se dá pela seguinte fórmula:

$$\begin{aligned}\mathbb{E}_{P,X}[S(P, X)] &= UNC-RES + REL = \\ &= \mathbb{E}_X[(X - E[X])^2] - \mathbb{E}_{P,X}[(E[X|P] - E[X])^2] + \mathbb{E}_{P,X}[(P - E[X|P])^2]\end{aligned}$$

Procedendo às contas, temos:

- modelo A:

$$\begin{aligned}UNC &= (2((1 - 0.5)^2 + (0 - 0.5)^2 + (0 - 0.5)^2 + (1 - 0.5)^2))/4 = \\ &= (0.25 + 0.25 + 0.25 + 0.25)/2 = 0.5\end{aligned}$$

(a incerteza, relativa apenas aos resultados observados, é comum a todos os modelos)

$$\begin{aligned}RES &= (2((0.5 - 0.5)^2 + (0.5 - 0.5)^2 + (0.5 - 0.5)^2 + (0.5 - 0.5)^2))/4 = \\ &= (0 + 0 + 0 + 0)/2 = 0 \\ REL &= (2((0.5 - 0.5)^2 + (0.5 - 0.5)^2 + (0.5 - 0.5)^2 + (0.5 - 0.5)^2))/4 = \\ &= (0 + 0 + 0 + 0)/2 = 0\end{aligned}$$

-modelo B:

$$\begin{aligned}RES &= (2((0.5 - 0.5)^2 + (0.5 - 0.5)^2 + (0.5 - 0.5)^2 + (0.5 - 0.5)^2))/4 = \\ &= (0 + 0 + 0 + 0)/2 = 0 \\ REL &= (2((1 - 0.5)^2 + (0 - 0.5)^2 + (1 - 0.5)^2 + (0 - 0.5)^2))/4 = \\ &= (0.25 + 0.25 + 0.25 + 0.25)/2 = 0.5\end{aligned}$$

-modelo C:

$$\begin{aligned}
RES &= (2((1 - 0.5)^2 + (0 - 0.5)^2 + (0 - 0.5)^2 + (1 - 0.5)^2)/4 = \\
&= (0.25 + 0.25 + 0.25 + 0.25)/2 = 0.5 \\
REL &= (2((1 - 1)^2 + (0 - 0)^2 + (0.8 - 0)^2 + (0.2 - 1)^2)/4 = \\
&= (0 + 0 + 0.64 + 0.64)/2 = 0.64
\end{aligned}$$

A incerteza dos resultados (0.5) corresponde ao valor da acurácia do modelo A. Isso ocorre para qualquer modelo constante que apresente, como previsão, a frequência média de cada resultado observado, pois modelos constantes sempre têm resolução nula ($E[X|P] - E[X] = 0$) e a frequência média como previsão garante (falta de) confiabilidade nula ($P - E[X|P] = 0$).

Os modelos A e B têm resolução nula. No caso do modelo B, apesar das previsões não serem constantes, o resultado médio dada qualquer das duas previsões (1 ou 0) é 0.5 (com 50% de frequência condicional para o resultado *cara*), sendo, portanto, igual ao resultado médio incondicional.

A confiabilidade é o que difere os modelos A e B. Enquanto A é perfeitamente confiável, a (falta de) confiabilidade calculada para B é 0.5. A cada previsão assertiva de B (1 ou 0) se seguiu um resultado que não correspondeu, em média, à previsão. Por exemplo, quando B previu probabilidade 1 para *cara*, em 50% das vezes o resultado foi *coroa*. Essa falta de confiabilidade é o que gerou a pior acurácia avaliada para o modelo B.

A decomposição URR do modelo C sofreu efeitos do problema da esparsidade, que será descrito no item 2.4.6. Cada previsão do modelo C foi evocada uma única vez. Assim, suas observações médias condicionais a cada previsão coincidem com a própria observação. Isso tende a aumentar (superestimar) os valores calculados para a resolução e a (falta de) confiabilidade.

Considerando os valores calculados, o modelo C teve resolução máxima (igual à incerteza dos resultados) e a pior confiabilidade entre os três modelos. Como a penalização na confiabilidade (0.64) foi superior ao bônus na resolução (0.5), o modelo C teve acurácia inferior ao modelo A, ficando ranqueado em segundo lugar.

Seguindo com a análise, vamos calcular a decomposição RDC da medida de Brier, que se dá pela seguinte fórmula:

$$\begin{aligned}\mathbb{E}_{P,X}[S(P,X)] &= REF-DIS + COR = \\ &= \mathbb{E}_P[(P - E[P])^2] - \mathbb{E}_{P,X}[(E[P|X] - E[P])^2] + \mathbb{E}_{P,X}[(X - E[P|X])^2]\end{aligned}$$

- modelo A:

$$\begin{aligned}REF &= (2((0.5 - 0.5)^2 + (0.5 - 0.5)^2 + (0.5 - 0.5)^2 + (0.5 - 0.5)^2))/4 = \\ &= (0 + 0 + 0 + 0)/2 = 0 \\ DIS &= (2((0.5 - 0.5)^2 + (0.5 - 0.5)^2 + (0.5 - 0.5)^2 + (0.5 - 0.5)^2))/4 = \\ &= (0 + 0 + 0 + 0)/2 = 0 \\ COR &= (2((1 - 0.5)^2 + (0 - 0.5)^2 + (0 - 0.5)^2 + (1 - 0.5)^2))/4 = \\ &= (0.25 + 0.25 + 0.25 + 0.25)/2 = 0.5\end{aligned}$$

- modelo B:

$$\begin{aligned}REF &= (2((1 - 0.5)^2 + (0 - 0.5)^2 + (1 - 0.5)^2 + (0 - 0.5)^2))/4 = \\ &= (0.25 + 0.25 + 0.25 + 0.25)/2 = 0.5 \\ DIS &= (2((0.5 - 0.5)^2 + (0.5 - 0.5)^2 + (0.5 - 0.5)^2 + (0.5 - 0.5)^2))/4 = \\ &= (0 + 0 + 0 + 0)/2 = 0 \\ COR &= (2((1 - 0.5)^2 + (0 - 0.5)^2 + (0 - 0.5)^2 + (1 - 0.5)^2))/4 = \\ &= (0.25 + 0.25 + 0.25 + 0.25)/2 = 0.5\end{aligned}$$

- modelo C:

$$\begin{aligned}REF &= (2((1 - 0.5)^2 + (0 - 0.5)^2 + (0.8 - 0.5)^2 + (0.2 - 0.5)^2))/4 = \\ &= (0.25 + 0.25 + 0.09 + 0.09)/2 = 0.34 \\ DIS &= (2((0.6 - 0.5)^2 + (0.4 - 0.5)^2 + (0.4 - 0.5)^2 + (0.6 - 0.5)^2))/4 = \\ &= (0.01 + 0.01 + 0.01 + 0.01)/2 = 0.02 \\ COR &= (2((1 - 0.6)^2 + (0 - 0.4)^2 + (0 - 0.4)^2 + (1 - 0.6)^2))/4 =\end{aligned}$$

$$= (0.16 + 0.16 + 0.16 + 0.16)/2 = 0.32$$

Todo modelo constante, como o modelo A, tem refinamento e discriminação nulos. Assim, sua acurácia é igual ao valor do atributo correção (e, neste caso, ambos são iguais à incerteza dos resultados).

O modelo B tem o maior refinamento (0.5), sendo o que apresenta maior variabilidade nas previsões, mas tem discriminação nula, pois, dado qualquer resultado (*cara* ou *coroa*) a previsão média condicional ao resultado é igual à previsão média incondicional (0.5). Pelo mesmo motivo, sua (falta de) correção (0.5) é a mesma do modelo A.

O modelo C tem refinamento intermediário (0.34), pois suas previsões variam menos que as do modelo B, e uma pequena discriminação (0.02), uma vez que as previsões médias condicionais aos resultados (0.6 e 0.4) dão um leve indicativo da ocorrência do respectivo resultado. Pelo mesmo motivo, a correção é a melhor entre os três modelos.

No comparativo com o modelo B, o C foi superior nos três atributos, o que corrobora sua melhor acurácia (0.64 contra 1). Já em relação ao modelo A, o C foi superior em discriminação e correção, tendo sido inferior somente no refinamento. Ainda assim, o modelo A – menos correto, menos refinado e sem discriminação – teve melhor acurácia. Isso merece uma reflexão.

O refinamento pode ser um atributo desejado quando acompanhado de uma boa discriminação. Um baixo refinamento (como o modelo A) significa um modelo pouco informativo, que não varia suas previsões e não é capaz de indicar a ocorrência individual de cada evento. Um modelo que forneça bons indicativos individuais, i.e., atribui probabilidade alta, em média, a cada observação individual, tem boa discriminação e, necessariamente, um alto refinamento, que é um limitador para a discriminação. Por outro lado, um alto refinamento sem uma boa discriminação (como os modelos B e C) significa um modelo que varia suas previsões em excesso, sem que cada variação seja um bom indicativo do resultado individual correspondente. Essa variação excessiva nas previsões pode até resultar em melhora na correção do modelo (como para o modelo C), mas vai prejudicar sua confiabilidade, que compara cada previsão individual com o resultado médio condicional à previsão e, em última instância, sua acurácia.

Cabe destacar que a decomposição RDC do modelo C não sofreu os efeitos do problema da esparsidade. Na realidade, como na decomposição RDC as médias condicionais são condicionadas nos resultados, o problema só poderia

ocorrer se os resultados observados fossem esparsos e, nesse caso, seria comum a todos os modelos.

2.4.4 Primeira Decomposição (Primária) de Medidas Próprias: Decomposição URR

“(...) medidas estritamente próprias de avaliação de previsões podem ser decompostas em termos relacionados à resolução e à confiabilidade da previsão. (...) Mostra-se que tanto a resolução como a confiabilidade têm um efeito positivo na medida. Demonstramos que resolução e confiabilidade estão diretamente relacionadas a atributos da previsão que são considerados positivos independentemente da noção de medida. Essa descoberta pode ser considerada uma justificativa epistemológica para utilizarmos medidas próprias ao avaliar previsões.”

[Bröcker, 2009]

Sejam $S = S(P, X)$ uma medida própria de avaliação de previsões e $q \in \mathcal{P}$ uma previsão para os eventos $x \in \mathcal{E}$. Bröcker [2009] definiu:

$$s(P, q) = \sum_{x \in \mathcal{E}} S(P, x) q_x = \mathbb{E}_{X \sim q} [S(P, X)],$$

a função de avaliação associada a S (onde q_x é a probabilidade atribuída pela previsão q ao evento x);

$$d(P, q) = s(P, q) - s(q, q),$$

a divergência de P e q associada a S (nota-se que d não é uma distância, pois, em geral, não é simétrica nem obedece à desigualdade triangular); e

$$e(P) = s(P, P),$$

a entropia de P associada a S .

A **decomposição URR** (uncertainty-resolution-reliability) generalizada para medidas próprias foi apresentada por Bröcker [2009] como:

$$\mathbb{E}_{P,X} [S(P, X)] = e(\bar{X}) - \mathbb{E}_P [d(\bar{X}, \bar{X}|P)] + \mathbb{E}_P [d(P, \bar{X}|P)],$$

onde \bar{X} e $\bar{X}|P$ são abusos de notação, e podem ser interpretados, respectivamente, como a previsão correspondente à observação média incondicional (ou $\frac{1}{N} \sum_{n=1}^N e_{k_n}$) e a previsão correspondente à observação média condicional a P .

O primeiro termo da decomposição, $e(\bar{X})$, corresponde à **incerteza** (UNC) associada às observações. O segundo termo $\mathbb{E}_P [d(\bar{X}, \bar{X}|P)]$ corresponde à **resolução** (RES) do conjunto de previsões, e o terceiro, $\mathbb{E}_P [d(P, \bar{X}|P)]$, à sua **confiabilidade** (REL).

Cabe notar que esses termos, enquanto expressões dos atributos, estão de acordo com as definições de [Murphy \[1993\]](#) apresentadas na tabela do item anterior.

Além disso, as esperanças sempre existem, em nosso contexto, uma vez que a base de jogos é finita e, portanto, os subconjuntos de valores efetivamente assumidos nos conjuntos \mathcal{P} e \mathcal{E} também o são. Em particular, embora \mathcal{P} nunca seja um conjunto discreto, as esperanças em \mathcal{P} serão calculadas como um somatório finito, e vamos utilizar o abuso de notação $\sum_{p \in \mathcal{P}}$.

É importante ressaltar também o significado do valor numérico dos atributos, sempre não-negativos, considerando medidas de avaliação próprias com orientação negativa.

Um conjunto de previsões perfeitas (i.e., que atribuem probabilidade 1 aos eventos que ocorrem), teria avaliação 0. Já um conjunto de previsões de baixa qualidade teria, como avaliação, um número alto, relativamente à escala da medida.

A **incerteza**, definida como a entropia da observação média \bar{X} (ou, equivalentemente, entropia da distribuição incondicional das observações), depende apenas dos eventos ocorridos, podendo ser alta ou baixa mesmo para previsões perfeitas.

A **confiabilidade**, valor esperado da divergência entre P e $\bar{X}|P$, se anula para um conjunto de previsões perfeitas (pois $(\bar{X}|P=p) = p \quad \forall p$) e cresce à medida que $\bar{X}|P=p$ e p se distanciam. Dessa forma, o nome do atributo induz uma interpretação errada do valor aferido, que é, na realidade, uma medida de falta de confiabilidade.

A **resolução**, valor esperado da divergência entre \bar{X} e $\bar{X}|P$, é um atributo

com orientação positiva, que varia entre 0 (previsões sem qualquer resolução) e o valor atribuído à incerteza das observações. Assim, um conjunto de previsões perfeitas teria resolução igual à incerteza, garantindo a igualdade $S = 0 = UNC-RES + REL$.

Por outro lado, um conjunto de previsões de baixa qualidade teria, tipicamente, pouca resolução (número próximo a 0) e muita (falta de) confiabilidade.

Finalmente, cabe observar que um conjunto de previsões constantes não deve ser considerado como tendo boa qualidade, uma vez que não agrega qualquer informação sobre a ocorrência de cada evento individual.

Assim, a incerteza das observações, medida de avaliação de um conjunto de previsões constantes e iguais à média das observações, pode ser considerada como um valor de referência para a qualidade das previsões.

Modelos de previsão com avaliação maior que a incerteza não devem ser considerados modelos de boa qualidade. Esses modelos têm a resolução menor que a (falta de) confiabilidade.

Ao contrário, quanto maior a qualidade de um modelo, maior a diferença entre sua resolução e sua (falta de) confiabilidade, maior a distância entre sua avaliação e o valor da incerteza, e menor sua distância para o valor nulo.

2.4.5 Segunda Decomposição (Dual) de Medidas Próprias: Decomposição RDC

“Para medidas próprias de avaliação de previsões, a acurácia de um conjunto de previsões é definida como o valor esperado de sua avaliação em relação a todas as possíveis previsões e observações. Diferentes qualidades do desempenho das previsões podem ser obtidas quando expressamos a acurácia como uma soma de outros termos, um processo conhecido como decomposição da acurácia. Cada termo da decomposição se refere a uma qualidade do conjunto de previsões; aquelas consideradas importantes são as que representam atributos da distribuição conjunta de previsões e observações. Na principal decomposição, que chamamos ‘decomposição URR’, os termos são incerteza (uncertainty), resolução (resolution) e confiabilidade (reliability). (...) na segunda forma, chamada ‘decomposição RDC’, os termos se referem a refinamento (refinement), discriminação (discrimination) e correção (correctness) das previsões.”

[[Mitchell, 2020](#)]

Dadas uma previsão P e uma medida de avaliação S , [Mitchell \[2020\]](#) definiu o valor esperado de S para P em relação aos eventos possíveis X (com distribuição q), como:

$$S((P, q)) := \mathbb{E}_{X \sim q} [S(P, X)].$$

Mitchell definiu também \mathcal{P}^* como o conjunto das distribuições das previsões p sobre \mathcal{P} , i.e., $\mathcal{P}^* = \{(\dots, f(p), \dots); p \in \mathcal{P}\}$, onde $f(p)$ representa a frequência da ocorrência da previsão p no conjunto de eventos (ou distribuição marginal de P no ponto p). E ainda uma função $S^* : \mathcal{P}^* \times \mathcal{P} \rightarrow \mathbb{R}$ como medida de avaliação estendida de S , no sentido que:

$$S^*(\delta_X^*, P) = S(P, X),$$

onde $\delta_X^* \in \mathcal{P}^*$ corresponde à distribuição em que a previsão que atribui $\text{prob}(X) = 1$ é evento certo em \mathcal{P} e S é uma medida de avaliação própria em $\mathcal{P} \times \mathcal{E}$.

Mitchell [2020] define a **decomposição RDC** (refinement-discrimination-correctness) de uma medida de avaliação própria como:

$$\begin{aligned} \mathbb{E}_{P,X} [S(P, X)] &= \mathbb{E}_{P,X} [S^*(\delta_X^*, P)] = \\ &= S^*((q^*, q^*)) - \left(S^*((q^*, q^*)) - \mathbb{E}_X [S^*((q_X^*, q_X^*))] \right) + \left(\mathbb{E}_X [S^*((\delta_X^*, q_X^*))] - \mathbb{E}_X [S^*((q_X^*, q_X^*))] \right), \end{aligned}$$

onde $q^* = f(P)$, $q_X^* = f(P|X)$ e $S^*((q^*, q^*)) = \mathbb{E}_P [S^*(q^*, P)]$.

Aqui também cabem observações análogas às da seção anterior. Os termos são sempre não-negativos e, enquanto expressões dos atributos, estão de acordo com as definições de Murphy [1993].

O **refinamento**, entropia das previsões, é uma medida da variabilidade das previsões e não tem qualquer relação com os eventos ocorridos. Um conjunto de previsões constantes tem refinamento nulo. Um conjunto de previsões de boa qualidade pode ter refinamento alto ou baixo, dependendo da incerteza das observações. Um conjunto de previsões perfeitas tem refinamento igual à incerteza.

A **correção**, valor esperado da divergência entre X e $\bar{P}|X$, se anula para um conjunto de previsões perfeitas (pois $(\bar{P}|X = x) = x$ para todo x) e cresce à medida que $\bar{P}|X = x$ e x se distanciam. De modo semelhante à confiabilidade, o nome do atributo induz uma interpretação errada do valor aferido, que é, na realidade, uma medida de falta de correção.

A **discriminação**, valor esperado da divergência entre \bar{P} e $\bar{P}|X$, é um atributo com orientação positiva, que varia entre 0 (previsões sem qualquer discriminação) e o valor atribuído ao refinamento das previsões. Assim, um conjunto de previsões perfeitas teria discriminação igual ao refinamento, garantindo a igualdade $S = 0 = REF-DIS + COR$.

Por outro lado, um conjunto de previsões de baixa qualidade teria, tipicamente, discriminação bem inferior ao refinamento e alta (falta de) correção. Um conjunto de previsões constantes teria refinamento e discriminação nulos e (falta de) correção igual ao valor da medida de avaliação, necessariamente maior ou igual à incerteza das observações.

De modo geral, quanto maior a qualidade de um modelo, menor a diferença entre seu refinamento e sua discriminação e menor a sua (falta de) correção.

2.4.6 O problema da esparsidade na distribuição conjunta de previsões e observações

“(...) uma larga variedade de resultados e uma grande diversidade de previsões probabilísticas têm a consequência de que, em qualquer amostra de pares previsão-resultado, é raro que os valores se repitam, tanto de previsões como de resultados. Isso gera uma instabilidade no cálculo dos atributos das decomposições.(...) Esparsidade afeta aqueles atributos que têm esperanças condicionais em suas expressões. (...) Se uma instância particular de uma previsão não aparece numa amostra, nenhum resultado será computado para aquele estrato de previsão e a observação média condicionada àquela previsão terá um valor ilegítimamente nulo; mesmo se a previsão é evocada, mas apenas em poucas ocasiões, isso reduz o tamanho da amostra dos resultados correspondentes a um valor pequeno demais para que a esperança condicional seja estimada com alguma estabilidade.”

[[Mitchell, 2020](#)]

Um conjunto de observações ou de previsões esparso gera problemas no cálculo dos atributos.

A título de ilustração, vamos supor um modelo com três previsões de resultado - $P1 = (0.34, 0.33, 0.33)$, $P2 = (0.33, 0.34, 0.33)$ e $P3 = (0.33, 0.33, 0.34)$ - e ainda que, para cada uma delas, tenha ocorrido um resultado distinto – vitória do mandante, empate e vitória do visitante. Vendo que as três previsões são todas muito próximas entre si, se estivessem consideradas num mesmo estrato condicional, a confiabilidade do modelo teria boa avaliação, pois a frequência média do resultado nesse estrato seria $(1/3, 1/3, 1/3)$. Por outro lado, se cada uma das três previsões for alocada num estrato diferente, a confiabilidade relativa a cada um dos estratos será péssima, o que irá comprometer a avaliação da confiabilidade do modelo como um todo.

[Mitchell \[2020\]](#) apresenta uma solução interessante para o problema: a redução de dimensionalidade do espaço de previsões (ou do espaço de eventos possíveis), que passaria a conter somente alguns pontos. Cada previsão (ou resultado) teria que ser aproximada a algum desses pontos representativos do

espaço reduzido. Isso reduziria a estratificação das previsões (ou resultados), aumentaria o número de resultados (ou previsões) em cada estrato e daria maior estabilidade ao cálculo dos atributos baseados em esperanças condicionais - resolução e confiabilidade, condicionais à previsão; discriminação e correção, condicionais ao resultado.

O método de redução de dimensionalidade traz algumas questões que precisam ser consideradas.

A primeira é a dificuldade em escolher os pontos representativos de maneira efetiva. Por exemplo, se reduzirmos o espaço de previsões de resultados de futebol ao conjunto de pontos em que as coordenadas tenham duas casas decimais (ou, equivalentemente, números inteiros percentuais), ainda teremos um espaço com 5151 pontos. Se consideramos somente os pontos em que as coordenadas sejam múltiplas de 0.05, ainda nos sobrarão 231 pontos. Para uma base de jogos com pouco mais de 1000 partidas (como é o caso da que vamos usar), não parece redução suficiente. (E para o conjunto de previsões de placares esse problema é ainda mais grave.)

Definida a redução do espaço de previsões, a segunda questão é como identificar cada previsão a um desses pontos representativos no novo espaço. Por exemplo, supondo a redução anterior, em que os pontos têm coordenadas múltiplas de 0.05, não fica claro a qual ponto devemos identificar a previsão (0.56, 0.22, 0.22). Fazendo por aproximação, coordenada a coordenada, teríamos (0.55, 0.2, 0.2), que não faz parte do espaço. Escolhas plausíveis parecem ser (0.5, 0.25, 0.25), (0.6, 0.2, 0.2), (0.55, 0.2, 0.25) e (0.55, 0.25, 0.2), mas devemos ter um critério objetivo para escolher entre as opções.

Reduzido o espaço e identificada cada previsão a um ponto do novo espaço, surgem mais duas questões. Ao calcularmos os atributos de resolução e confiabilidade, encontraremos valores relativos às novas previsões. Esses valores são dependentes das escolhas de redução de espaço e de identificação de previsões que fizemos anteriormente; para cada escolha, um valor diferente para os atributos. Além disso, perdemos a igualdade na expressão da decomposição da medida. O valor da medida não é mais igual à soma dos valores dos atributos, uma vez que a medida é referente às previsões originais e os atributos às previsões no espaço reduzido.

Essas questões foram levantadas por Mitchell, que propôs uma abordagem para lidar com as duas últimas, utilizando fatores de correção relativos

à variação das previsões dentro de cada agrupamento (variação dos valores das previsões originais que foram identificadas com cada ponto do espaço reduzido) e à covariação entre previsões e resultados dentro de cada agrupamento. Esses dois fatores, somados aos atributos, garantem a igualdade na equação da decomposição.

O detalhamento dessa abordagem ocuparia espaço demasiado nesta dissertação (e, temo, tomaria um tempo proibitivo do autor), mas acredito que uma investigação nesse sentido possa ser interessante no futuro.

3 Metodologia

Nesta seção descrevemos os modelos, os dados e as medidas de avaliação utilizados no trabalho.

3.1 Primeira Abordagem: Previsões da EMap para o Brasileirão 2019

A inspiração para este trabalho surgiu a partir das tentativas do nosso grupo de pesquisa na EMap – [Esporte em Números](#) - de prever os resultados do Campeonato Brasileiro de 2019. À medida em que as previsões eram publicadas e os resultados das partidas se materializavam, ficava clara a necessidade de avaliar o desempenho do modelo de previsões.

Em um primeiro esforço nesse sentido, selecionamos algumas medidas de avaliação e procuramos utilizar técnicas de visualização da informação para buscar um melhor entendimento a respeito do desempenho do modelo. As figuras a seguir (3, 4, 5, e 6) mostram a avaliação de cada partida nas diferentes medidas utilizadas na ocasião e deixam claro que o desempenho do modelo apresentou grande variação de jogo para jogo.

Todas essas avaliações e visualizações compuseram o artigo [[Fontanella et al., 2020](#)], apresentado na conferência SIBGRAPI 2020. Embora o escopo do artigo fosse distinto, muitas das pesquisas realizadas, técnicas aprimoradas e questões levantadas tiveram grande impacto e aproveitamento na concepção e no desenvolvimento desta dissertação.

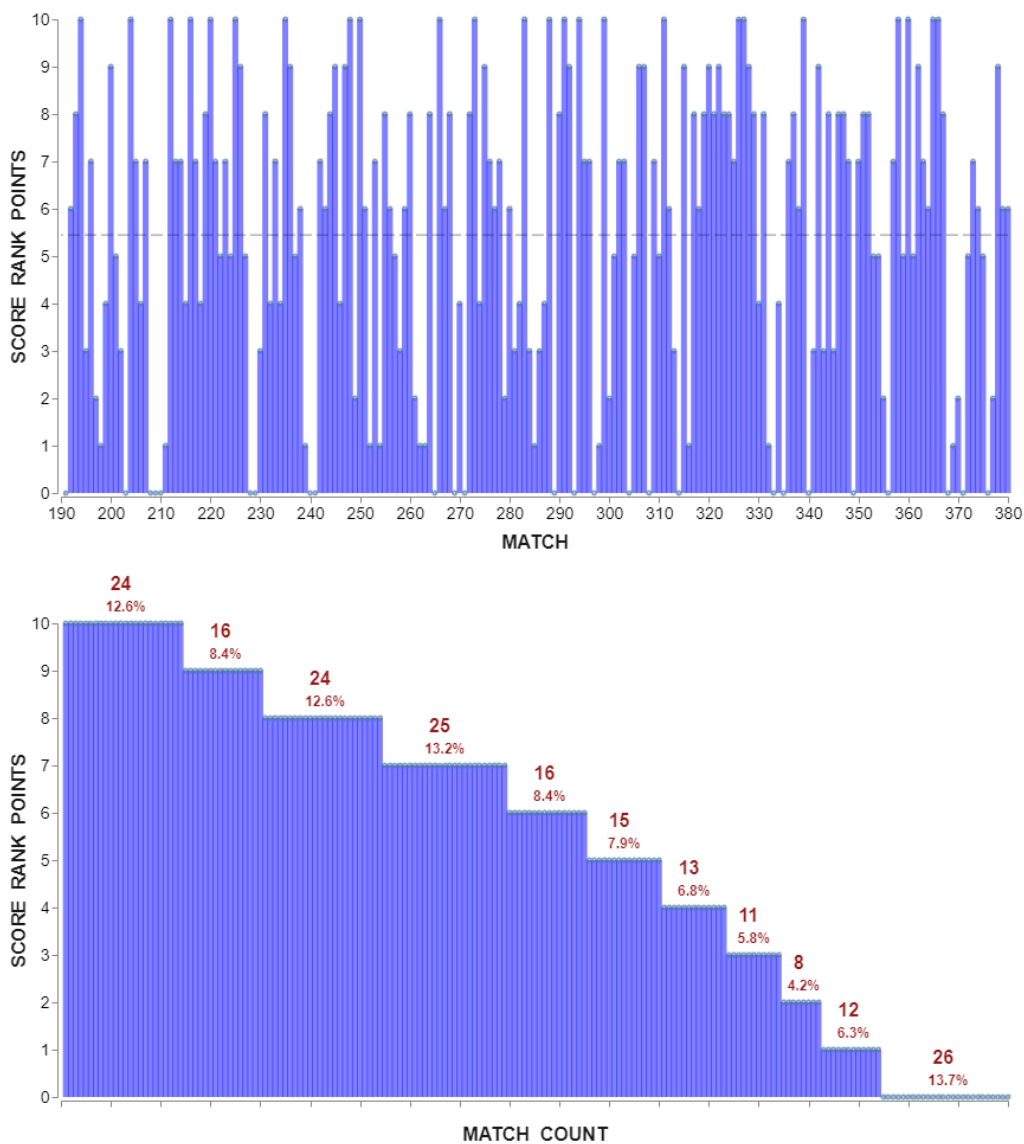


Figura 3: Pontos no ranking de placares previstos, sendo conferidos 10 pontos para o placar mais provável, 9 para o segundo, e assim sucessivamente até o décimo, que recebe 1 ponto. Todos os demais placares recebem 0 ponto. *Em cima*: ordem da tabela do campeonato. *Embaixo*: mesmos dados agrupados pelos pontos no ranking.

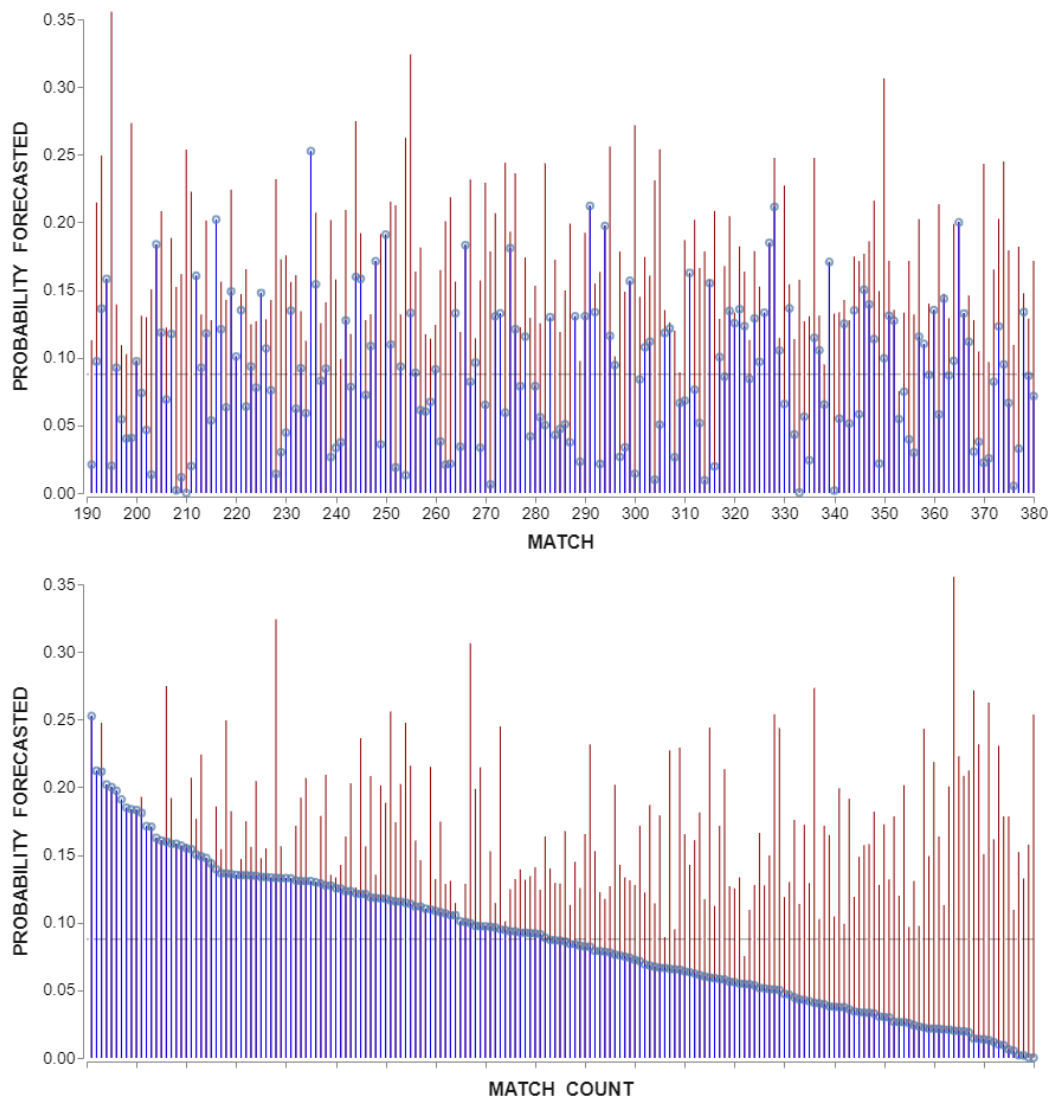


Figura 4: Probabilidades previstas para os placares ocorridos plotados em barras azuis e círculos. As barras vermelhas representam as maiores probabilidades previstas para um placar. *Em cima*: ordem da tabela do campeonato. *Embaixo*: mesmos dados com a probabilidade prevista para o placar ocorrido em ordem decrescente.

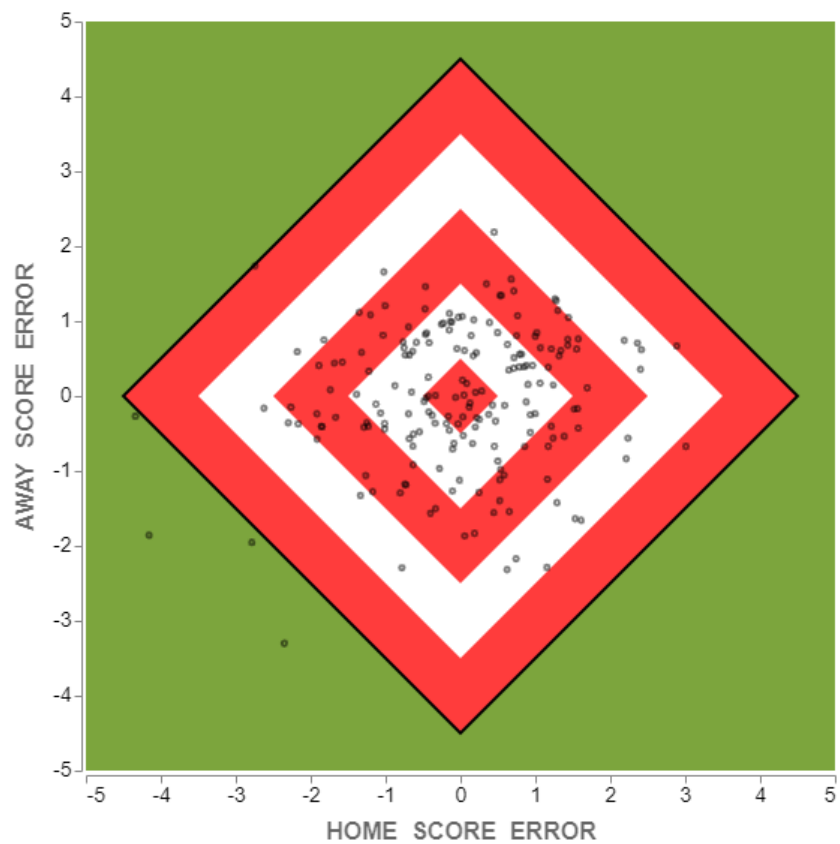


Figura 5: Alvo tendo o placar ocorrido como centro e pontos representando o erro relativo à média da previsão.

Infelizmente, poucas previsões probabilísticas para jogos de futebol são publicadas, e nenhuma delas (até onde encontramos) diz respeito às probabilidades dos placares das partidas. Dessa forma, a base de comparação para o nosso modelo é pequena para previsões de resultados e inexistente para previsões de placares.

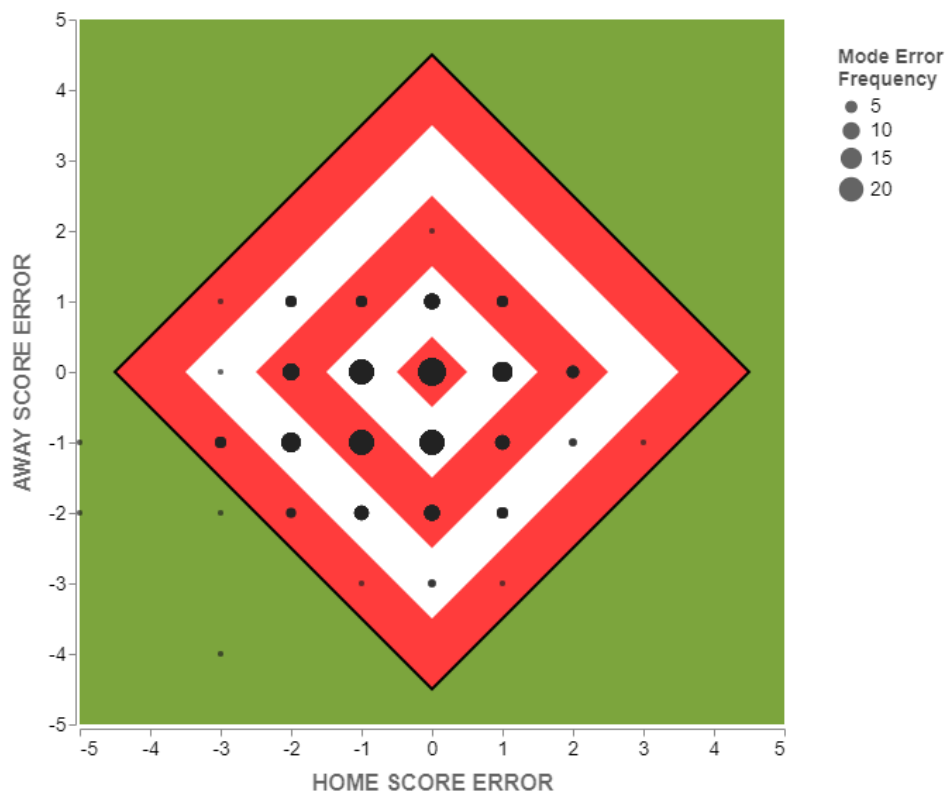


Figura 6: Alvo tendo o placar ocorrido como centro e pontos representando o erro relativo à moda da previsão.

3.2 Outros Modelos de Previsão

Para fazer melhor uso das medidas de avaliação e possibilitar uma análise do desempenho do nosso modelo a partir da comparação com outros, o próximo passo será definir e implementar novos modelos de previsão de placares e resultados de partidas de futebol.

3.2.1 Modelos Poisson Independentes Baseados em Retrospecto de Partidas e/ou Dados Econômicos

Nossa abordagem inicial para a construção de novos modelos de previsão será utilizar o modelo já em prática como ponto de partida e modificá-lo através da variação de alguns parâmetros. Em primeiro lugar, vamos caracterizar o modelo inicial.

Nosso modelo em uso é baseado em duas variáveis aleatórias com distribuição de Poisson, independentes, determinadas a partir dos parâmetros individuais dos times. Cada time tem quatro parâmetros, um representando o poder ofensivo quando mandante, outro o poder defensivo quando mandante, outro o poder ofensivo quando visitante e o último o poder defensivo quando visitante. Esses parâmetros são estimados em função dos placares das partidas disputadas anteriormente naquele campeonato. Para que partidas recentes tenham maior influência na estimação dos parâmetros do que aquelas realizadas no início do campeonato, é utilizada uma função de esquecimento. Em geral, essa função é uma exponencial com base fixa, maior que 1. Seu expoente é uma função crescente da data ou da rodada de cada partida considerada.

Para gerar nosso segundo modelo, o parâmetro que vamos variar é a base da exponencial da função de esquecimento. Quanto maior o número, maior o esquecimento, no sentido de jogos antigos terem um peso bem menor em relação aos recentes. Quando a base diminui e se aproxima de 1, o peso de cada partida tende à igualdade. De fato, quando a base é igual a 1, todas as partidas têm o mesmo peso e não há esquecimento.

Antes de definir o fator de esquecimento do novo modelo, vamos testar diferentes valores para a base da exponencial, implementar cada modelo, rodar as previsões e as respectivas avaliações. Ao fim dessa etapa de testes, vamos selecionar o modelo com o melhor desempenho.

Definido nosso segundo modelo, o próximo passo será gerar o terceiro. Para isso, voltaremos à origem desses modelos, o artigo de [Maher \[1982\]](#). Maher questionou a necessidade de usar tantos parâmetros: “(...) *seria realmente necessário ter parâmetros separados para a qualidade de ataque de um time em casa e fora?*” Após alguns testes, ele decidiu adotar um modelo com dois parâmetros por time, como “*sendo o mais apropriado*”.

[Lee \[1997\]](#) também utilizou um modelo com dois parâmetros por time – um representando a força ofensiva e outro representando a força defensiva – e mais um parâmetro geral relacionado ao fator-casa, correspondendo a um bônus ao parâmetro da variável Poisson relativa ao time mandante.

Com base nessas ideias, nosso terceiro modelo seguirá esse formato, com apenas dois parâmetros por time e mais um parâmetro geral representando o fator-casa.

Em procedimento semelhante ao adotado anteriormente, vamos testar diferentes valores para o fator-casa e, ao fim, selecionar o modelo que obtiver o melhor desempenho. Para todos os modelos de teste, vamos usar a função de esquecimento igual à do último modelo que selecionamos.

Para definir nosso quarto modelo, vamos utilizar dados econômicos, em complemento aos placares de partidas anteriores, para estimar os parâmetros de cada time. O número de parâmetros para cada time – dois ou quatro – será igual ao daquele modelo – segundo ou terceiro – que tiver obtido melhor desempenho.

Os dados econômicos correspondem ao valor do elenco de cada time, no início de cada campeonato, estimado pelo site [Transfermarkt](#). Vamos assumir uma correlação positiva no sentido de, quanto maior o valor do elenco de um time, maiores tenderão a ser seus parâmetros. Essa tendência poderá não se verificar, de fato, em função da combinação dos dados econômicos com os dados de placares anteriores que já vínhamos utilizando.

Mais uma vez, vamos implementar diferentes modelos de teste, cada um dando um peso diferente à variação dos valores dos elencos para a estimação dos parâmetros, e aquele que obtiver o melhor desempenho será selecionado.

Nesse ponto, teremos três novos modelos de previsão, cujas avaliações poderão ser comparadas ao modelo inicial.

3.2.2 Modelos Idealizados para Benchmarking: PIP, PIS e FREQ

Definidos os novos modelos, uma vez avaliados, é imediato proceder à comparação entre eles e ranqueá-los quanto ao desempenho na previsão das partidas em nossa base. Isso não quer dizer que, de alguma forma, será possível aferir se o desempenho foi bom ou não. Em teoria, é possível que todos os modelos tenham bons desempenhos ou que nenhum deles apresente desempenho satisfatório. Para tal, vamos precisar definir maneiras de aferir esses níveis de desempenho. Isso será feito através da implementação de modelos ideais, que chamaremos modelos de *benchmarking*. Esses modelos farão “previsão a posteriori”, isto é, utilizarão nas previsões os dados das próprias partidas a serem previstas. Assim, o nível de desempenho das previsões poderá ser controlado.

O primeiro modelo de *benchmarking* será denominado Poisson Independen-

dentes Perfeito (PIP). Para cada partida, o modelo PIP usará, como parâmetros das variáveis Poisson, o número de gols que cada equipe realmente marcou. A partir daí, as probabilidades da previsão serão calculadas como em todos os modelos anteriores.

O modelo PIP tem a peculiaridade de ser um limitador maximal para o desempenho dos modelos de Poisson Independentes na previsão de placares de partidas de futebol. Isso quer dizer que, qualquer que seja a medida de avaliação adotada relativa ao placar, o desempenho do modelo PIP será superior ou igual ao de qualquer outro modelo desse tipo.

Ao segundo modelo de *benchmarking* chamaremos Poisson Independentes Satisfatório (PIS). O modelo PIS foi concebido como uma perturbação do modelo PIP. Para cada partida, o modelo usará, como parâmetro da variável Poisson, um número aleatório advindo de uma distribuição normal em torno do número de gols que a equipe realmente marcou.

Por exemplo, se uma partida terminou com placar de 3 x 1, o modelo PIS usará, para prever o número de gols do mandante, uma variável aleatória com distribuição de Poisson com parâmetro sorteado de uma distribuição normal em torno de 3. De maneira análoga, para prever o número de gols do visitante, o modelo terá uma variável aleatória Poisson com parâmetro sorteado de uma distribuição normal em torno de 1.

A distribuição normal utilizada no modelo PIS terá desvio-padrão igual a 1.071. Assim, o parâmetro da variável Poisson estará a menos de 1 unidade de distância do número real de gols em cerca de 65% dos casos e a menos de 2.5 unidades de distância em cerca de 98% dos casos.

É claro que, para a variável de Poisson fazer sentido e o desempenho da previsão não ser alterado, toda vez que a distribuição normal gerar um número negativo, o parâmetro utilizado será o simétrico desse número, relativo ao centro da distribuição. Por exemplo, se o número de gols marcados por um time for igual a 1 e a distribuição normal gerar -0.6, o parâmetro da variável Poisson correspondente será 2.6. Isso vai gerar um viés de erro para as previsões do modelo PIS no sentido positivo, mas não vai alterar o erro absoluto.

O modelo PIS funcionará como um *benchmarking* para a avaliação de modelos no seguinte sentido: sabemos que, independentemente do placar de uma partida, o PIS será capaz de, com quase dois terços de chances para cada time, prever como mais provável o número exato de gols marcados ou 1 a mais ou a menos. Além disso, em apenas 2% das vezes o número de gols

indicado como mais provável estará fora de um intervalo de 2.5 em torno do número de gols realmente marcado.

Nosso terceiro modelo de *benchmarking* será um modelo ingênuo, que acompanha a frequência dos placares ocorridos nas partidas a serem previstas. Vamos denominá-lo Ingênuo-Frequência (FREQ).

No modelo FREQ a previsão será a mesma em todo o conjunto de partidas a serem previstas. Se, por exemplo, o placar de 1 x 1 tiver ocorrido em 11% das partidas desse conjunto, para cada uma delas, a previsão indicará uma probabilidade de 11% para esse placar.

O modelo FREQ servirá como *benchmarking* no sentido de ser uma espécie de limitador inferior de desempenho. Podemos considerar que as previsões de um modelo agregam alguma informação relevante à medida que elas apresentam desempenho superior às do modelo FREQ. Caso contrário, se um modelo apresenta desempenho inferior, poderia simplesmente ser substituído por um modelo que tentasse acompanhar as frequências, independentemente dos times envolvidos na partida.

Com isso, obtemos três modelos de *benchmarking*, onde um é um limitador (superior) absoluto de desempenho, outro, entendemos, deve apresentar um desempenho empiricamente satisfatório e o último representa um desempenho mínimo a ser alcançado.

3.3 Base de Dados Utilizada Para Avaliação dos Modelos

Os principais dados com que vamos trabalhar correspondem aos resultados e placares de partidas da série A do campeonato brasileiro de futebol masculino. Os dados foram obtidos como um arquivo em formato .csv na base do projeto Esporte em Números. Constam no arquivo todas as partidas do Brasileirão Série A desde a temporada de 2006.

Uma segunda base de dados corresponde ao valor total do elenco de cada time participante do Brasileirão Série A, obtido no site [Transfermarkt](#). Essa informação se encontra disponível desde o ano de 2006, mas com pouca precisão para as temporadas mais antigas. Por exemplo, na temporada de 2006, o número de jogadores com valor avaliado em times como Cruzeiro, Grêmio e Santos é inferior a 11, o que certamente acarreta que o valor total do elenco desses times esteja subestimado.

Além disso, para incorporar essa informação em nosso modelo, precisamos que ela esteja disponível por clube e por data. Infelizmente, esse recorte só está acessível para os 40 clubes que disputam, na temporada de 2020, as séries A ou B do Brasileirão. Assim, para clubes como Santa Cruz (série A em 2016), Joinville (em 2015), Criciúma (em 2013-14) e Portuguesa (em 2008 e 2012-13), Grêmio Prudente (em 2009-10), Santo André (em 2009) e Ipatinga (em 2008) não conseguimos obter os dados de valor do elenco nas datas desejadas.

Com tudo isso, para viabilizar a utilização dos dados econômicos, se tornou necessário fazer algumas escolhas:

Entre 2017 e 2019 não houve qualquer tipo de problema, e os dados obtidos no site puderam ser usados em sua totalidade.

Para 2016, os dados econômicos do Santa Cruz não estavam disponíveis. Após uma busca pela rede, encontrei uma reportagem de [Zirpoli \[2015\]](#), que apresentava os dados do Transfermarkt naquela ocasião. Pelo que pude perceber, os dados seriam referentes à quinzena anterior àqueles que eu coletei e, portanto, o valor do elenco do Santa Cruz poderia ser aproveitado sem perda considerável.

Para 2015, os dados econômicos do Joinville não estavam disponíveis. Mesmo realizando buscas semelhantes pela rede, não consegui obter a informação desejada. A partir daí, considerando o histórico do Joinville, único dos 20 clubes a não ter participação recente na Série A, e as projeções da época, que o apontavam, antes do início do campeonato, como um dos favoritos ao rebaixamento [[Goal, 2015](#)], optei por igualar o valor de mercado do Joinville ao menor valor dentre os demais 19 times do campeonato.

Para 2014, os dados econômicos começaram a apresentar aparente inconsistência. Enquanto no período de 2015 a 2019 os valores de elenco no início da série A variaram entre 11.9 milhões de euros (América-MG de 2016) e 112.7 milhões de euros (Palmeiras de 2019), os dados do início do campeonato de 2014 mostravam três elencos (Figueirense, Sport e Chapecoense) com valor inferior a 2 milhões de euros. Dessa forma, decidi aumentar o valor atribuído a esses times para o menor valor plausível dentre os demais dezesseis elencos: 8.45 milhões de euros (Goiás). Ao Criciúma, cuja avaliação da época estava inacessível no site, foi atribuído o mesmo valor de 8.45 milhões, que ficou como uma espécie de piso, comum a cinco elencos do início daquele campeonato.

De 2013 para trás as inconsistências aumentaram, tornando difícil apurar e justificar as decisões a respeito dos dados econômicos. As distorções seriam

tais que poderiam causar grande impacto no modelo e nas previsões a serem avaliadas. Decidi, pois, fazer o recorte nesse ponto, desprezando todos os dados até 2013.

Assim, nossa base de dados final considera somente as partidas e os dados econômicos referentes ao período de 2014 a 2019, perfazendo um total de seis temporadas.

3.4 Medidas de Avaliação Mais Adequadas a Previsões de Placares e Resultados de Futebol

Pegando a esteira da discussão apresentada na seção 2 sobre diferentes propriedades das medidas de avaliação e sua relevância, tomamos as seguintes orientações para definir que medida(s) usar para avaliar previsões de resultados e placares de futebol:

- i. Considerando que as previsões devem ser consistentes, a escolha estará restrita às opções de medidas próprias.
- ii. O espaço de eventos é ordenado no caso de previsões de resultados e parcialmente ordenado no caso de previsões de placares. Dessa forma, a sensibilidade a distância é uma propriedade desejada para a medida.
- iii. A localidade é uma propriedade intuitiva e única com relação a medidas próprias, o que facilita seu entendimento e sua comparabilidade com outros trabalhos. Dessa forma, não deve ser ignorada.

Assim, uma vez que uma medida não pode ser ao mesmo tempo sensível a distância e local, vamos utilizar duas medidas, uma com cada propriedade. A sensibilidade a distância, entretanto, é considerada a propriedade mais importante em nosso contexto e, portanto, todos os ranqueamentos de modelos serão efetuados utilizando medidas que têm essa propriedade.

Para previsões de resultado, a medida de probabilidade ordenada parece adequada, pois além de possuir as características desejadas (sendo estritamente própria e sensível a distância), já é amplamente utilizada em trabalhos sobre o assunto. Contudo, proponho utilizar o dobro de seu valor, que generaliza a distância de maneira mais compatível com o espaço de resultados e com as demais medidas adotadas. Para evitar confusão, chamarei essa medida de dupla probabilidade ordenada (2RPS). É claro que, para comparar as avaliações dos modelos com as efetuadas em outros trabalhos que utilizem a RPS, basta dividir por 2 o valor aferido pela medida 2RPS.

Além dela, vamos utilizar também a medida de ignorância em bans (IGN10),

que é a unidade de informação em que o logaritmo tem base 10 [MacKay, 2003]. Essa medida é estritamente própria, local e de orientação negativa.

O motivo da escolha de medir a ignorância em bans, em vez de bits ou nats, é que a escala da medida se torna mais próxima da escala da 2RPS, facilitando tanto a comparação como as plotagens gráficas.

Vale ressaltar, com relação à medida de ignorância, que uma única previsão que aponte probabilidade 0 para um resultado que venha a ocorrer tornará a avaliação infinita, independentemente de quantos outros resultados e previsões estejam sendo avaliados.

Quando consideramos previsões de placares, a medida de probabilidade ordenada (RPS) não está bem definida, uma vez que não existe a ordenação natural do espaço de placares. Dessa forma, precisamos pesquisar ou desenvolver uma outra medida, própria e sensível a distância, que possa ser utilizada nesse espaço.

Gneiting and Raftery [2007] escreveram que “A medida de probabilidade ordenada se aplica a previsões probabilísticas que tomam a forma de funções de distribuição cumulativa das previsões. Ela generaliza o erro absoluto e se trata de um caso especial de uma medida nova e mais geral, a medida de energia.”.

A medida de energia é definida, no mesmo artigo, da seguinte maneira:

$$\text{Energia}(p, x) = \frac{1}{2} \mathbb{E}_{X, X'} [\|X - X'\|^\beta] - \mathbb{E}_X [\|X - x\|^\beta],$$

onde p é a previsão, x é o evento observado, X e X' são cópias independentes de um vetor aleatório com distribuição p , $\|\cdot\|$ é a norma euclidiana em \mathbb{R}^m , \mathbb{E}_X denota o valor esperado conforme a distribuição p de X e β é um número real no intervalo $(0, 2)$. Quando $\beta = 1$ e os vetores X e X' estão em \mathbb{R} ($m = 1$), a medida de energia se iguala à 2RPS.

Ainda no mesmo artigo, Gneiting and Raftery [2007] definem as medidas de energia não-euclidianas, que têm a mesma expressão, mas onde a norma $\|\cdot\|$ não é a euclidiana. Em particular, estamos interessados na medida de energia com norma-1 ou “norma da soma”, a qual chamaremos de medida de energia norma-1 (ENERG1).

O espaço de placares de futebol é composto pelos placares $x = (x_1, x_2)$, que são elementos de \mathbb{R}^2 . Aplicando a medida de energia norma-1, com orientação negativa, ao nosso contexto, obtemos:

$$ENERG1(p, (x_1, x_2)) = \mathbb{E}_X \left[|X_1 - x_1| + |X_2 - x_2| \right] - \frac{1}{2} \mathbb{E}_{X, X'} \left[|X_1 - X'_1| + |X_2 - X'_2| \right]$$

Considerando o espaço de placares restrito a qualquer número finito de gols, as esperanças acima existem e são finitas. A medida de energia norma-1 fica, então, bem definida para qualquer placar x e qualquer previsão p do espaço de previsões. [Gneiting and Raftery \[2007\]](#) mostram que a medida de energia norma-1 é própria.

A medida de energia norma-1 generaliza distância da seguinte maneira: para uma previsão pontual, i.e., que aponte probabilidade 1 para um placar específico e 0 para os demais placares, a medida $ENERG1$ será igual à soma das distâncias em \mathbb{R} entre o número de gols marcados por cada time e o respectivo número de gols previsto. Por exemplo, se a previsão é de probabilidade 1 para o placar 2 x 1, e o resultado da partida é 1 x 3, a medida $ENERG1$ da previsão será 3 ($= |1 - 2| + |3 - 1|$), podendo ser identificada como a diferença (absoluta) em número de gols entre o placar ocorrido e o placar previsto (somando as duas dimensões).

Além da medida de energia norma-1, vamos utilizar, para as previsões de placares, a medida de ignorância em bans ($IGN10$), como medida estritamente própria e local.

Vale aqui a mesma ressalva observada para a utilização da $IGN10$ nas previsões de resultados, com o agravante de que, na maioria das previsões, a probabilidade prevista para alguns placares é bem próxima de zero.

Finalmente, vamos também exibir nas tabelas, apenas por curiosidade, outras medidas de avaliação, como a 0-1, a probabilidade média e a de Brier.

3.5 Decomposições das Medidas de Avaliação Seleccionadas

Nesta seção vamos explicitar as equações de decomposição para a medida 2RPS, proposta no item anterior. É composta exclusivamente por longas

sequências de contas relativas a esperanças, esperanças condicionais, somatórios, logaritmos e probabilidades. O leitor que deseja apenas ver a aplicação das decomposições URR e RDC aos resultados efetivamente calculados nas medidas selecionadas, pode ignorar esta parte da dissertação e avançar direto para a leitura da seção 4, sem qualquer perda relevante.

3.5.1 Convenções e métodos adotados para o cálculo das esperanças

Na seção 2 apresentamos as decomposições URR e RDC generalizadas desenvolvidas por Bröcker [2009] e Mitchell [2020]. Agora, vamos calcular essas decomposições e explicitar os atributos para o caso específico da medida 2RPS que selecionamos para avaliar previsões de resultados.

Formalmente, as esperanças em P e em X se referem, respectivamente, a todo o espaço de previsões e a todo o espaço de eventos possíveis. Na prática, queremos calcular cada esperança em P relativamente ao conjunto de previsões que realmente foram evocadas e cada esperança em X relativamente ao conjunto de resultados que realmente ocorreram na base de jogos. Assim, ambos os conjuntos a serem considerados são discretos, finitos e de cardinalidade inferior ou igual a N (o número de partidas na base).

À exceção dos modelos FREQ e PIP, todos os demais apresentam esparsidade total no espaço de previsões, com cada previsão sendo evocada uma única vez. Isso gera o problema da esparsidade, discutido na seção 2.4.6, que acarreta a superestimação dos atributos resolução e (falta de) confiabilidade.

O conjunto de resultados observados não é esparsa, com cada um dos três resultados ocorrendo em mais de 20% das partidas. Isso permite que os atributos condicionados em resultados - discriminação e correção - possam ser estimados com estabilidade.

Todas as esperanças relativas às previsões podem ser calculadas a partir de um somatório no conjunto de previsões efetivamente evocadas ponderado pela frequência com que cada previsão foi evocada – $1/N$ para todos os modelos à exceção de FREQ e PIP.

As esperanças relativas ao resultado podem ser calculadas como um somatório no conjunto de três resultados possíveis, ponderado pela frequência relativa de cada um.

Assim, para calcular as esperanças, vamos adotar a seguinte metodologia:

$$\mathbb{E}_P[\cdot] = \sum_{i=1}^N \frac{Y_i}{N}$$

$$\mathbb{E}_X[\cdot] = \sum_{k=1}^3 Y_k f(x_k) = Y_H f(x_H) + Y_D f(x_D) + Y_A f(x_A).$$

3.5.2 Expressões das decomposições da medida 2RPS

A medida 2RPS é definida por:

$$\begin{aligned} 2RPS(p, x) &= \sum_{i=1}^2 \left(\sum_{j=1}^i (p_j - x_j) \right)^2 = (p_H - x_H)^2 + (p_H - x_H + p_D - x_D)^2 = \\ &= (p_H - x_H)^2 + ((1 - p_A) - (1 - x_A))^2 = (p_H - x_H)^2 + (p_A - x_A)^2 \end{aligned}$$

Decomposição URR:

$$\begin{aligned} \mathbb{E}_{P,X} [S(P, X)] &= e(\bar{X}) - \mathbb{E}_P [d(\bar{X}, \bar{X}|P)] + \mathbb{E}_P [d(P, \bar{X}|P)] = \\ &= \sum_{k=1}^3 S(\bar{X}, x_k) \bar{X}_k + \\ &- \sum_{i=1}^N \left[\sum_{k=1}^3 S(\bar{X}, x_k) (\bar{X}|P=p_i)_k - \sum_{k=1}^3 S(\bar{X}|P=p_i, x_k) (\bar{X}|P=p_i)_k \right] \frac{1}{N} + \\ &+ \sum_{i=1}^N \left[\sum_{k=1}^3 S(P_i, x_k) (\bar{X}|P=p_i)_k - \sum_{k=1}^3 S(\bar{X}|P=p_i, x_k) (\bar{X}|P=p_i)_k \right] \frac{1}{N} \end{aligned}$$

A incerteza é dada pela expressão:

$$e(\bar{X}) = \sum_{k=1}^3 S(\bar{X}, x_k) \bar{X}_k =$$

$$\begin{aligned}
&= \bar{X}_H ((\bar{X}_H - 1)^2 + (\bar{X}_A - 0)^2) + \bar{X}_D ((\bar{X}_H - 0)^2 + (\bar{X}_A - 0)^2) + \\
&+ \bar{X}_A ((\bar{X}_H - 0)^2 + (\bar{X}_A - 1)^2) = \bar{X}_H^3 - 2\bar{X}_H^2 + \bar{X}_H + \bar{X}_H^2 \bar{X}_D + \bar{X}_H^2 \bar{X}_A + \\
&+ \bar{X}_A^2 \bar{X}_H + \bar{X}_A^2 \bar{X}_D + \bar{X}_A^3 - 2\bar{X}_A^2 + \bar{X}_A = \bar{X}_H^2 (\bar{X}_H + \bar{X}_D + \bar{X}_A - 2) + \bar{X}_H + \\
&\bar{X}_A^2 (\bar{X}_H + \bar{X}_D + \bar{X}_A - 2) + \bar{X}_A = \bar{X}_H (1 - \bar{X}_H) + \bar{X}_A (1 - \bar{X}_A),
\end{aligned}$$

onde \bar{X}_H , \bar{X}_D e \bar{X}_A representam, respectivamente, as frequências dos resultados de vitória do mandante, empate e vitória do visitante na base de jogos.

A **resolução** é dada pela expressão:

$$\begin{aligned}
\mathbb{E}_P [d(\bar{X}, \bar{X}|P)] &= \mathbb{E}_P \left[\sum_{k=1}^3 S(\bar{X}, x_k) (\bar{X}|P)_k - \sum_{k=1}^3 S(\bar{X}|P, x_k) (\bar{X}|P)_k \right] = \\
&= \mathbb{E}_P \left[(\bar{X}|P)_H ((\bar{X}_H - 1)^2 + (\bar{X}_A - 0)^2) - ((\bar{X}|P)_{H-1})^2 - ((\bar{X}|P)_A - 0)^2) + \right. \\
&\quad + (\bar{X}|P)_D ((\bar{X}_H - 0)^2 + (\bar{X}_A - 0)^2) - ((\bar{X}|P)_{H-0})^2 - ((\bar{X}|P)_A - 0)^2) + \\
&\quad \left. + (\bar{X}|P)_A ((\bar{X}_H - 0)^2 + (\bar{X}_A - 1)^2) - ((\bar{X}|P)_{H-0})^2 - ((\bar{X}|P)_A - 1)^2) \right] = \\
&= \mathbb{E}_P \left[(\bar{X}|P)_H \left((\bar{X}_H^2 - 2\bar{X}_H + 1) - ((\bar{X}|P)_H^2 - 2(\bar{X}|P)_H + 1) \right) + (\bar{X}|P)_D (\bar{X}_H^2 - (\bar{X}|P)_H^2) + \right. \\
&\quad + (\bar{X}|P)_A (\bar{X}_H^2 - (\bar{X}|P)_H^2) + (\bar{X}|P)_H (\bar{X}_A^2 - (\bar{X}|P)_A^2) + (\bar{X}|P)_D (\bar{X}_A^2 - (\bar{X}|P)_A^2) + \\
&\quad \left. + (\bar{X}|P)_A \left((\bar{X}_A^2 - 2\bar{X}_A + 1) - ((\bar{X}|P)_A^2 - 2(\bar{X}|P)_A + 1) \right) \right] = \\
&= \mathbb{E}_P \left[(\bar{X}_H^2 - (\bar{X}|P)_H^2) ((\bar{X}|P)_H + (\bar{X}|P)_D + (\bar{X}|P)_A) + (\bar{X}|P)_H (1 - 2\bar{X}_H - 1 + 2(\bar{X}|P)_H) + \right. \\
&\quad \left. + (\bar{X}_A^2 - (\bar{X}|P)_A^2) ((\bar{X}|P)_H + (\bar{X}|P)_D + (\bar{X}|P)_A) + (\bar{X}|P)_A (1 - 2\bar{X}_A - 1 + 2(\bar{X}|P)_A) \right] = \\
&= \mathbb{E}_P \left[\bar{X}_H^2 - (\bar{X}|P)_H^2 - 2(\bar{X}|P)_H \bar{X}_H + 2(\bar{X}|P)_H^2 + \bar{X}_A^2 + (\bar{X}|P)_A^2 - 2(\bar{X}|P)_A \bar{X}_A + 2(\bar{X}|P)_A^2 \right] = \\
&= \mathbb{E}_P \left[(\bar{X}_H - (\bar{X}|P)_H)^2 + (\bar{X}_A - (\bar{X}|P)_A)^2 \right] =
\end{aligned}$$

$$= \sum_{i=1}^N \left[\left(\bar{X}_H - (\bar{X}|P=p_i)_H \right)^2 + \left(\bar{X}_A - (\bar{X}|P=p_i)_A \right)^2 \right] \frac{1}{N},$$

A **confiabilidade** é dada pela expressão:

$$\begin{aligned} \mathbb{E}_P [d(P, \bar{X}|P)] &= \mathbb{E}_P \left[\sum_{k=1}^3 S(P, x_k) (\bar{X}|P)_k - \sum_{k=1}^3 S((\bar{X}|P), x_k) (\bar{X}|P)_k \right] = \\ & \quad (\dots) \\ &= \sum_{i=1}^N \left[\left(p_{i_H} - (\bar{X}|P=p_i)_H \right)^2 + \left(p_{i_A} - (\bar{X}|P=p_i)_A \right)^2 \right] \frac{1}{N}, \end{aligned}$$

Decomposição RDC:

Seja $S^* : \mathcal{P}^* \times \mathcal{P} \rightarrow \mathbb{R}$ definida por:

$$S^*(r^*, p) = \left(\left(\mathbb{E}_{P \sim r^*} [P] \right)_H - p_H \right)^2 + \left(\left(\mathbb{E}_{P \sim r^*} [P] \right)_A - p_A \right)^2$$

S^* é própria, pois:

$$\begin{aligned} S^*((r^*, q^*)) &= \mathbb{E}_{Q \sim q^*} [S^*(r^*, Q)] = \mathbb{E}_{Q \sim q^*} \left[\left(\left(\mathbb{E}_{P \sim r^*} [P] \right)_H - Q_H \right)^2 + \left(\left(\mathbb{E}_{P \sim r^*} [P] \right)_A - Q_A \right)^2 \right] = \\ &= \mathbb{E}_{Q \sim q^*} \left[\left(\left(\mathbb{E}_{P \sim r^*} [P] \right)_H^2 - 2 \left(\mathbb{E}_{P \sim r^*} [P] \right)_H Q_H + Q_H^2 \right) + \left(\left(\mathbb{E}_{P \sim r^*} [P] \right)_A^2 - 2 \left(\mathbb{E}_{P \sim r^*} [P] \right)_A Q_A + Q_A^2 \right) \right] = \\ &= \left(\left(\mathbb{E}_{P \sim r^*} [P] \right)_H^2 - 2 \left(\mathbb{E}_{P \sim r^*} [P] \right)_H \left(\mathbb{E}_{Q \sim q^*} [Q] \right)_H + \left(\mathbb{E}_{Q \sim q^*} [Q^2] \right)_H \right) + \\ &+ \left(\left(\mathbb{E}_{P \sim r^*} [P] \right)_A^2 - 2 \left(\mathbb{E}_{P \sim r^*} [P] \right)_A \left(\mathbb{E}_{Q \sim q^*} [Q] \right)_A + \left(\mathbb{E}_{Q \sim q^*} [Q^2] \right)_A \right) = \\ &= \left(\left(\mathbb{E}_{P \sim r^*} [P] \right)_H - \left(\mathbb{E}_{Q \sim q^*} [Q] \right)_H \right)^2 - \left(\mathbb{E}_{Q \sim q^*} [Q] \right)_H^2 + \left(\mathbb{E}_{Q \sim q^*} [Q^2] \right)_H + \\ &= \left(\left(\mathbb{E}_{P \sim r^*} [P] \right)_A - \left(\mathbb{E}_{Q \sim q^*} [Q] \right)_A \right)^2 - \left(\mathbb{E}_{Q \sim q^*} [Q] \right)_A^2 + \left(\mathbb{E}_{Q \sim q^*} [Q^2] \right)_A \end{aligned}$$

é minimizada para $r^* = q^*$.

Além disso, S^* é uma extensão de S , pois, $\forall x \in \mathcal{E}$:

$$\begin{aligned} S^*(\delta_x^*, p) &= \left(\left(\mathbb{E}_{P \sim \delta_x^*} [P] \right)_H - p_H \right)^2 + \left(\left(\mathbb{E}_{P \sim \delta_x^*} [P] \right)_A - p_A \right)^2 = \\ &= \left(x_H - p_H \right)^2 + \left(x_A - p_A \right)^2 = S(p, x). \end{aligned}$$

A decomposição RDC da medida 2RPS é:

$$\begin{aligned} \mathbb{E}_{P, X} [S(P, X)] &= \mathbb{E}_{P, X} [S^*(\delta_X^*, P)] = \\ &= S^*((q^*, q^*)) - \left(S^*((q^*, q^*)) - \mathbb{E}_X [S^*((q_X^*, q_X^*))] \right) + \left(\mathbb{E}_X [S^*((\delta_X^*, q_X^*))] - \mathbb{E}_X [S^*((q_X^*, q_X^*))] \right). \end{aligned}$$

O **refinamento** é dado pela expressão:

$$\begin{aligned} S^*((q^*, q^*)) &= \mathbb{E}_P [S^*(q^*, P)] = \\ &= \mathbb{E}_P \left[\left(\left(\mathbb{E}_{Q \sim q^*} [Q] \right)_H - P_H \right)^2 + \left(\left(\mathbb{E}_{Q \sim q^*} [Q] \right)_A - P_A \right)^2 \right] = \\ &= \sum_{i=1}^N \left[(\bar{P}_H - p_{i_H})^2 + (\bar{P}_A - p_{i_A})^2 \right] \frac{1}{N}, \end{aligned}$$

onde \bar{P}_H e \bar{P}_A representam as previsões médias para os resultados de, respectivamente, vitória do mandante e vitória do visitante na base de jogos.

A **discriminação** é dada pela expressão:

$$S^*((q^*, q^*)) - \mathbb{E}_X [S^*((q_X^*, q_X^*))] = REF - \mathbb{E}_X \left[\mathbb{E}_{P \sim q_X^*} [S^*(q_X^*, P)] \right] =$$

$$\begin{aligned}
&= REF - \mathbb{E}_X \left[\mathbb{E}_{P \sim q_X^*} \left[\left(\left(\mathbb{E}_{Q \sim q_X^*} [Q] \right)_H - P_H \right)^2 + \left(\left(\mathbb{E}_{Q \sim q_X^*} [Q] \right)_A - P_A \right)^2 \right] \right] = \\
&= REF - \mathbb{E}_X \left[\mathbb{E}_{P \sim q_X^*} \left[\left((\bar{P}|X)_H - P_H \right)^2 + \left((\bar{P}|X)_A - P_A \right)^2 \right] \right] = \\
&= REF - \sum_{l=1}^3 \left[\sum_{i=1}^{N_l} \left[\left((\bar{P}|X=x_l)_H - p_{i_H} \right)^2 + \left((\bar{P}|X=x_l)_A - p_{i_A} \right)^2 \right] \frac{1}{N_l} \right] f(x_l) = \\
&= REF - \sum_{l=1}^3 \left[\sum_{i=1}^{N_l} \left[\left((\bar{P}|X=x_l)_H - p_{i_H} \right)^2 + \left((\bar{P}|X=x_l)_A - p_{i_A} \right)^2 \right] \frac{1}{N_l} \right] \frac{N_l}{N} = \\
&= REF - \sum_{i=1}^N \left[\left((\bar{P}|X=x_i)_H - p_{i_H} \right)^2 + \left((\bar{P}|X=x_i)_A - p_{i_A} \right)^2 \right] \frac{1}{N} = \\
&= \sum_{i=1}^N \left[(\bar{P}_H - p_{i_H})^2 + (\bar{P}_A - p_{i_A})^2 - \left((\bar{P}|X=x_i)_H - p_{i_H} \right)^2 - \left((\bar{P}|X=x_i)_A - p_{i_A} \right)^2 \right] \frac{1}{N} = \\
&= \sum_{i=1}^N \left[(\bar{P}_H^2 - 2\bar{P}_H p_{i_H} + p_{i_H}^2) + (\bar{P}_A^2 - 2\bar{P}_A p_{i_A} + p_{i_A}^2) + \right. \\
&\quad \left. - \left((\bar{P}|X=x_i)_H^2 - 2(\bar{P}|X=x_i)_H p_{i_H} + p_{i_H}^2 \right) - \left((\bar{P}|X=x_i)_A^2 - 2(\bar{P}|X=x_i)_A p_{i_A} + p_{i_A}^2 \right) \right] \frac{1}{N} = \\
&= \sum_{i=1}^N \left[\bar{P}_H^2 - 2\bar{P}_H p_{i_H} + \bar{P}_A^2 - 2\bar{P}_A p_{i_A} - (\bar{P}|X=x_i)_H^2 + 2(\bar{P}|X=x_i)_H p_{i_H} + \right. \\
&\quad \left. - (\bar{P}|X=x_i)_A^2 + 2(\bar{P}|X=x_i)_A p_{i_A} \right] \frac{1}{N} = \left[\sum_{i=1}^N [\bar{P}_H^2 - 2\bar{P}_H p_{i_H}] + \sum_{i=1}^N [\bar{P}_A^2 - 2\bar{P}_A p_{i_A}] + \right. \\
&\quad \left. - \sum_{i=1}^N [(\bar{P}|X=x_i)_H^2 - 2(\bar{P}|X=x_i)_H p_{i_H}] - \sum_{i=1}^N [(\bar{P}|X=x_i)_A^2 + 2(\bar{P}|X=x_i)_A p_{i_A}] \right] \frac{1}{N} = \\
&= \left[\left[N\bar{P}_H^2 - 2\bar{P}_H \sum_{i=1}^N p_{i_H} \right] + \left[N\bar{P}_A^2 - 2\bar{P}_A \sum_{i=1}^N p_{i_A} \right] + \right. \\
&\quad \left. - \sum_{l=1}^3 \left[\sum_{i=1}^{N_l} [(\bar{P}|X=x_i)_H^2 - 2(\bar{P}|X=x_i)_H p_{i_H}] + \sum_{i=1}^{N_l} [(\bar{P}|X=x_i)_A^2 + 2(\bar{P}|X=x_i)_A p_{i_A}] \right] \right] \frac{1}{N} =
\end{aligned}$$

$$\begin{aligned}
&= \left[\left[N\bar{P}_H^2 - 2\bar{P}_H N\bar{P}_H \right] + \left[N\bar{P}_A^2 - 2\bar{P}_A N\bar{P}_A \right] - \sum_{l=1}^3 \left[\left[N_l(\bar{P}|X=x_l)_H^2 + \right. \right. \\
&\quad \left. \left. - 2(\bar{P}|X=x_l)_H \sum_{i=1}^{N_l} p_{i_H} \right] + \left[N_l(\bar{P}|X=x_l)_A^2 - 2(\bar{P}|X=x_l)_A \sum_{i=1}^{N_l} p_{i_A} \right] \right] \frac{1}{N} = \\
&= \left[\left[-N\bar{P}_H^2 \right] + \left[-N\bar{P}_A^2 \right] - \sum_{l=1}^3 \left[\left[N_l(\bar{P}|X=x_l)_H^2 - 2(\bar{P}|X=x_l)_H N_l(\bar{P}|X=x_l)_H \right] + \right. \right. \\
&\quad \left. \left. + \left[N_l(\bar{P}|X=x_l)_A^2 - 2(\bar{P}|X=x_l)_A N_l(\bar{P}|X=x_l)_A \right] \right] \right] \frac{1}{N} = \\
&= \left[\left[-N\bar{P}_H^2 \right] + \left[-N\bar{P}_A^2 \right] - \sum_{l=1}^3 \left[\left[-N_l(\bar{P}|X=x_l)_H^2 \right] + \left[-N_l(\bar{P}|X=x_l)_A^2 \right] \right] \right] \frac{1}{N} = \\
&= \left[-\sum_{i=1}^N \left[\bar{P}_H^2 + \bar{P}_A^2 \right] + \sum_{i=1}^N \left[(\bar{P}|X=x_i)_H^2 + (\bar{P}|X=x_i)_A^2 \right] \right] \frac{1}{N} = \\
&\quad \sum_{i=1}^N \left[\bar{P}_H^2 - (\bar{P}|X=x_i)_H^2 + \bar{P}_A^2 - (\bar{P}|X=x_i)_A^2 \right] \frac{1}{N}
\end{aligned}$$

A **correção** é dada pela expressão:

$$\begin{aligned}
&\mathbb{E}_X [S^*((\delta_X^*, q_X^*))] - \mathbb{E}_X [S^*((q_X^*, q_X^*))] = \mathbb{E}_X \left[\mathbb{E}_{P \sim q_X^*} \left[S^*(\delta_X^*, P) - S^*(q_X^*, P) \right] \right] = \\
&= \mathbb{E}_X \left[\mathbb{E}_{P \sim q_X^*} \left[(X_H - P_H)^2 + (X_A - P_A)^2 - ((\bar{P}|X)_H - P_H)^2 - ((\bar{P}|X)_A - P_A)^2 \right] \right] = \\
&= \mathbb{E}_X \left[\mathbb{E}_{P \sim q_X^*} \left[X_H^2 - 2X_H P_H + X_A^2 - 2X_A P_A - (\bar{P}|X)_H^2 + 2P_H(\bar{P}|X)_H + \right. \right. \\
&\quad \left. \left. - (\bar{P}|X)_A^2 + 2P_A(\bar{P}|X)_A \right] \right] = \sum_{l=1}^3 \frac{N_l}{N} \left[\sum_{i=1}^{N_l} \frac{1}{N_l} \left[x_{l_H}^2 - 2x_{l_H} p_{i_H} + x_{l_A}^2 - 2x_{l_A} p_{i_A} + \right. \right. \\
&\quad \left. \left. - (\bar{P}|X=x_l)_H^2 + 2p_{i_H}(\bar{P}|X=x_l)_H - (\bar{P}|X=x_l)_A^2 + 2p_{i_A}(\bar{P}|X=x_l)_A \right] \right] =
\end{aligned}$$

$$\begin{aligned}
&= \sum_{l=1}^3 \frac{N_l}{N} \left[x_{l_H}^2 + x_{l_A}^2 - (\bar{P}|X=x_l)_H^2 - (\bar{P}|X=x_l)_A^2 + \right. \\
&+ \sum_{i=1}^{N_l} \frac{1}{N_l} \left[-2x_{l_H}p_{i_H} - 2x_{l_A}p_{i_A} + 2p_{i_H}(\bar{P}|X=x_l)_H + 2p_{i_A}(\bar{P}|X=x_l)_A \right] \Big] = \\
&= \sum_{l=1}^3 \frac{N_l}{N} \left[x_{l_H}^2 + x_{l_A}^2 - (\bar{P}|X=x_l)_H^2 - (\bar{P}|X=x_l)_A^2 - 2x_{l_H}(\bar{P}|X=x_l)_H + \right. \\
&- 2x_{l_A}(\bar{P}|X=x_l)_A + 2(\bar{P}|X=x_l)_H(\bar{P}|X=x_l)_H + 2(\bar{P}|X=x_l)_A(\bar{P}|X=x_l)_A \Big] = \\
&= \sum_{i=1}^N \frac{1}{N} \left[x_{i_H}^2 - 2x_{i_H}(\bar{P}|X=x_i)_H - (\bar{P}|X=x_i)_H^2 + 2(\bar{P}|X=x_i)_H(\bar{P}|X=x_i)_H + \right. \\
&+ x_{i_A}^2 - 2x_{i_A}(\bar{P}|X=x_i)_A - (\bar{P}|X=x_i)_A^2 + 2(\bar{P}|X=x_i)_A(\bar{P}|X=x_i)_A \Big] = \\
&= \sum_{i=1}^N \left[\left(x_{i_H} - (\bar{P}|X=x_i)_H \right)^2 + \left(x_{i_A} - (\bar{P}|X=x_i)_A \right)^2 \right] \frac{1}{N}.
\end{aligned}$$

4 Resultados e Análise

Nesta seção vamos apresentar os resultados, comparar os modelos e decompor as avaliações.

4.1 Resultados dos Diferentes Modelos

Vamos começar apresentando os resultados dos diferentes modelos descritos na seção anterior.

4.1.1 Modelos Poisson Independentes com 4 Parâmetros por Time e 1 Fator de Esquecimento

Esses modelos apresentam quatro parâmetros por time – um para força ofensiva como mandante, um para força defensiva como mandante, um para força ofensiva como visitante e um para força defensiva como visitante – e mais um parâmetro geral, correspondente ao fator de esquecimento do modelo.

A função de esquecimento é do tipo:

$$base^{-1 + \frac{\text{rodada da partida}}{\text{última rodada jogada}}},$$

onde a base é um número positivo maior que 1 e o expoente varia entre -1 (partidas mais antigas) e 0 (partidas mais recentes). Dessa forma, o modelo “lembra” de cada partida com um peso que varia entre $1/base$ e 1.

Os modelos de teste foram implementados com diferentes valores para a base da função de esquecimento. Os valores testados foram: 1 (sem esquecimento), 1.5, 2, 3, 5, 10 e 100. Quanto maior o fator, maior o esquecimento das partidas iniciais do campeonato. A medida utilizada para a comparação foi a energia norma-1 do placar.

A tabela 3 mostra o desempenho de cada um dos modelos de teste sob diversas medidas.

É possível notar que diferentes medidas podem levar a conclusões distintas sobre o ranqueamento de desempenho dos modelos. Contudo, na maioria das medidas, o modelo com fator de esquecimento 1.5 obteve o melhor desempenho. Esse é o caso nas quatro medidas com as quais escolhemos trabalhar:

Tabela 3: Comparação dos modelos de teste de 4 parâmetros por time

Modelo	0-1 placar	Prob.m. placar	IGN10 placar	Brier placar	ENERG1 placar
fator 1	0.15713	0.08759	1.21184	0.92428	1.13981
fator 1.5	0.15099	0.08764	1.21070	0.92444	1.13813
fator 2	0.14835	0.08764	1.21088	0.92479	1.13833
fator 3	0.13958	0.08759	1.21251	0.92560	1.14054
fator 5	0.13169	0.08745	1.21688	0.92714	1.14652
fator 10	0.12820	0.08713	1.22685	0.93012	1.16014
fator 100	0.10977	0.08498	1.29157	0.94596	1.24562
Modelo	0-1 result.	Prob.m. result.	IGN10 result.	Brier result.	2RPS result.
fator 1	0.49958	0.40650	0.44865	0.61659	0.42779
fator 1.5	0.50223	0.40720	0.44860	0.61626	0.42738
fator 2	0.50135	0.40768	0.44892	0.61648	0.42747
fator 3	0.50573	0.40833	0.44987	0.61739	0.42814
fator 5	0.49783	0.40911	0.45187	0.61950	0.42982
fator 10	0.49696	0.41005	0.45597	0.62396	0.43350
fator 100	0.47850	0.41169	0.48038	0.64945	0.45487

energia norma-1 do placar, ignorância do placar, duplo RPS do resultado e ignorância do resultado.

Outra observação pertinente é o bom desempenho obtido por modelos com pouco esquecimento. O fator 1.5 corresponde a um peso acima de 65% para as partidas iniciais, o que significa que o conjunto de resultados das três primeiras rodadas vale mais para o modelo que o conjunto de resultados das duas últimas. Mesmo o modelo sem qualquer esquecimento, que dá o mesmo peso para todas as partidas, obteve um desempenho razoável, melhor do que modelos que deem pesos abaixo de 40% para as rodadas iniciais. Isso pode parecer contraintuitivo, na maneira como os seres humanos parecem fazer suas previsões, e talvez indique que os modelos se beneficiariam em utilizar mais dados.

A figura 7 mostra o desempenho dos modelos relativo à energia norma-1 do placar (ENERG1).

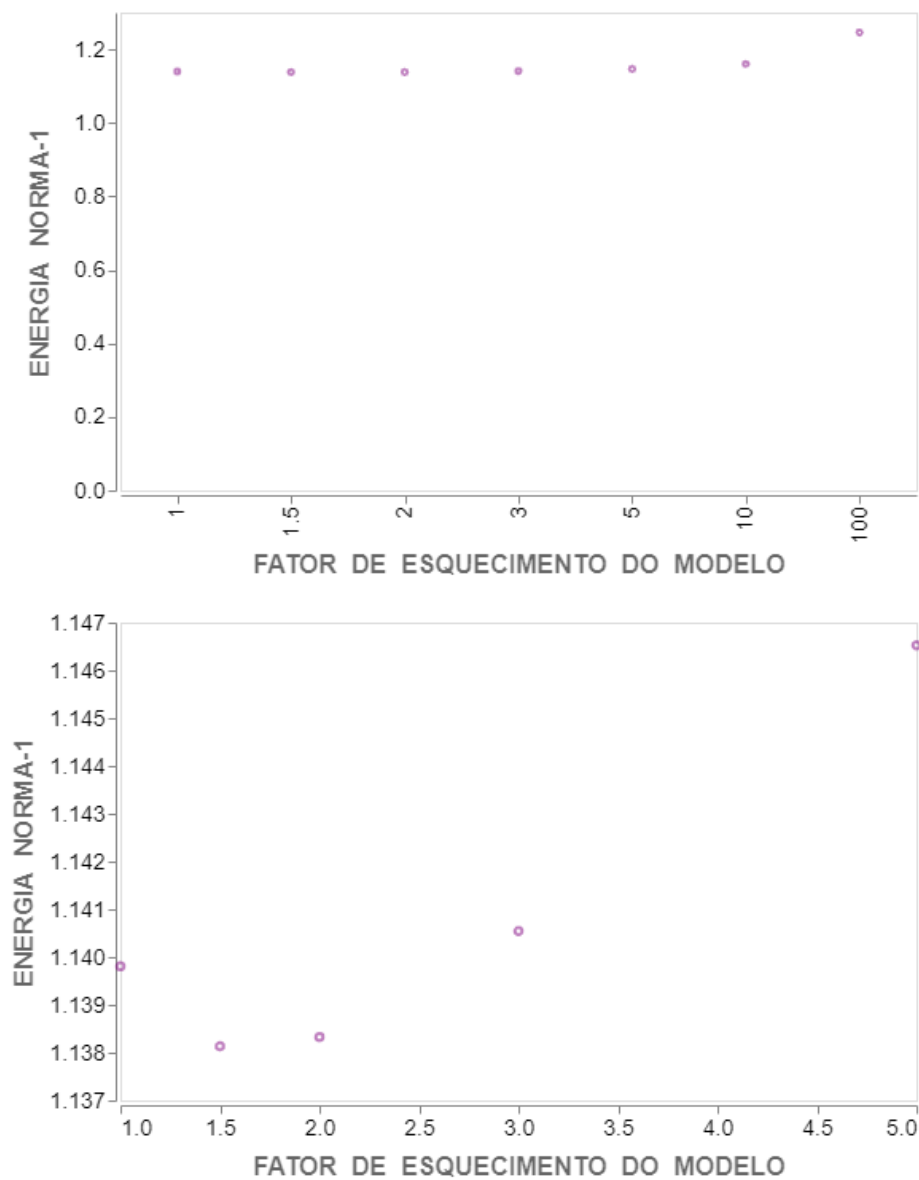


Figura 7: Modelos de 4 parâmetros por time. No eixo x, o fator de esquecimento do modelo. Na figura de cima, todos os modelos de teste. Na figura de baixo, zoom no eixo y e um recorte com os modelos que tiveram desempenho próximo.

4.1.2 Modelos Poisson Independentes com 2 Parâmetros por Time, 1 Fator de Esquecimento e 1 Parâmetro Geral para o Fator Casa

Esses modelos têm apenas dois parâmetros por time – a força ofensiva e a força defensiva – utilizados tanto nas partidas como mandante como quando visitante. O fator de esquecimento é o mesmo para todos os modelos: 1.5, que obteve melhor desempenho no teste anterior de modelos.

O fator casa, não contemplado no número de parâmetros de cada time, é, em vez disso, um parâmetro geral, comum a todos os times, e corresponde a um percentual aplicado como bônus ao parâmetro da variável Poisson referente ao time mandante e como ônus ao parâmetro da variável do time visitante, da seguinte maneira:

$$\begin{aligned}\Lambda_{mandante} &= \lambda_{mandante} (1 + fator\ casa) \\ \Lambda_{visitante} &= \lambda_{visitante} (1 - fator\ casa)\end{aligned}$$

Em nossa base de dados, no período observado de 2014 a 2019, o total de gols marcados pelos times mandantes foi de 3236, enquanto os visitantes marcaram 2060. O número médio de gols marcados pelos mandantes por partida foi de aproximadamente 1.42, cerca de 22.2% acima da média de gols por time por partida, o que pode servir como uma referência aos valores de teste.

Os valores testados para o fator casa foram: 0 (sem fator casa), .15, .20, .22, .23, .25, .30 e .35. A tabela 4 mostra o desempenho de alguns dos modelos de teste sob diversas medidas.

Tabela 4: Comparação dos modelos de teste de 2 parâmetros por time

Modelo	0-1 placar	Prob.m. placar	IGN10 placar	Brier placar	ENERG1 placar
casa 0	0.12908	0.08284	1.21329	0.92484	1.14717
casa 0.15	0.14134	0.08646	1.18625	0.91881	1.10428
casa 0.20	0.13694	0.08744	1.18225	0.91783	1.09797
casa 0.25	0.14221	0.08831	1.18086	0.91739	1.09574
casa 0.30	0.14572	0.08905	1.18218	0.91755	1.09767
casa 0.35	0.14135	0.08967	1.18641	0.91833	1.10384
Modelo	0-1 result.	Prob.m. result.	IGN10 result.	Brier result.	2RPS result.
casa 0	0.44338	0.37053	0.46855	0.65015	0.45964
casa 0.15	0.48907	0.39215	0.44695	0.61635	0.42695
casa 0.20	0.49346	0.39916	0.44339	0.61046	0.42124
casa 0.25	0.49784	0.40604	0.44166	0.60725	0.41810
casa 0.30	0.50574	0.41277	0.44181	0.60668	0.41748
casa 0.35	0.51714	0.41935	0.44388	0.60869	0.41932

O modelo com fator casa 0.25 obteve o melhor desempenho em nossa medida de interesse, a ENERG1, bem como em outras medidas, como as de ignorância (IGN10), tanto para o placar como para o resultado. O modelo com fator 0.3 foi o melhor na 2RPS do resultado.

Cabe destacar aqui que, à exceção do modelo com ausência de fator casa, todos os modelos tiveram desempenho melhor que os modelos com 4 parâmetros por time, o que sugere uma prevalência dos modelos com 2 parâmetros por time e um fator casa único. Curiosamente, o desempenho variou pouco em todo o intervalo $[0.15, 0.35]$ de fatores casa testados, o que fortalece a sugestão anterior.

A figura 8 mostra o desempenho dos modelos relativo à medida ENERG1.

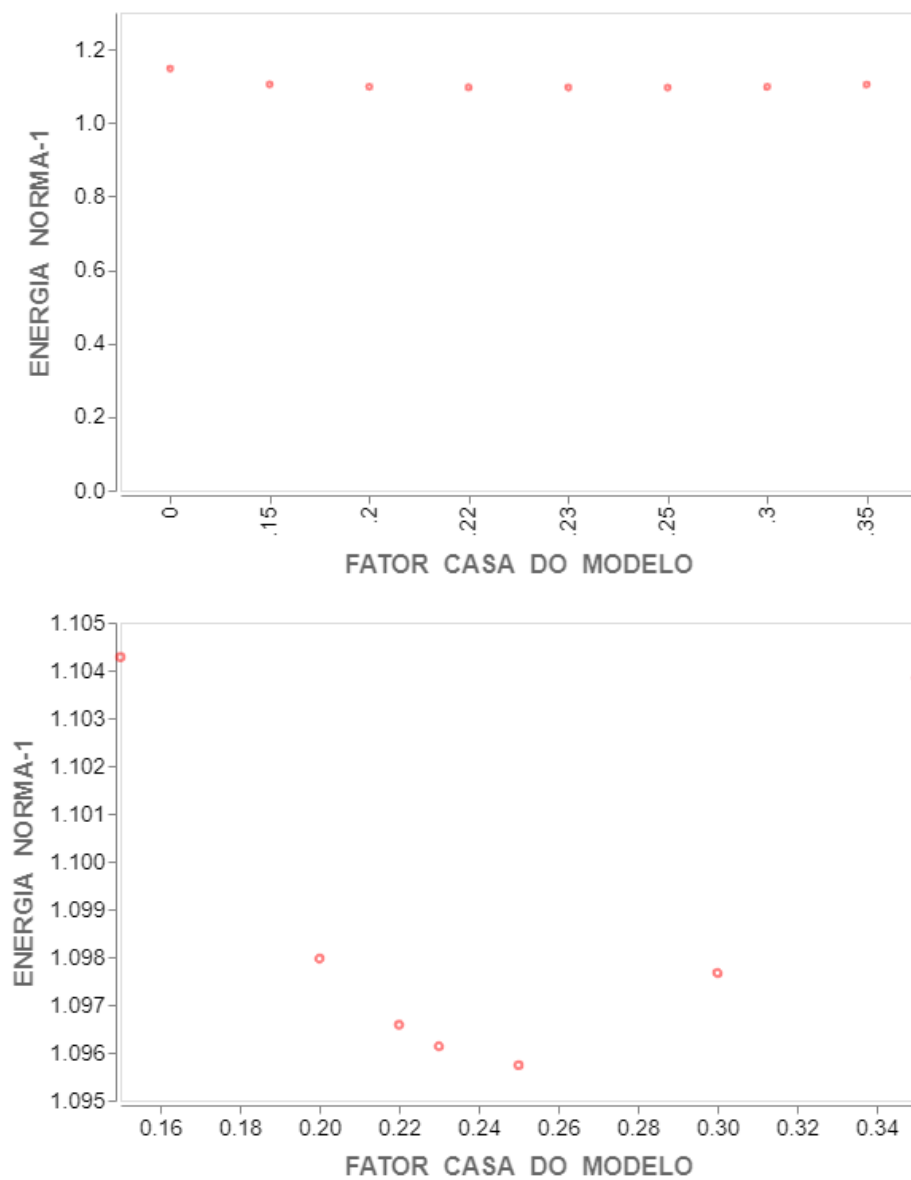


Figura 8: Modelos de 2 parâmetros por time. No eixo x, o fator casa do modelo. Na figura de cima, todos os modelos de teste. Na figura de baixo, zoom no eixo y e um recorte com os modelos que tiveram desempenho próximo.

4.1.3 Modelos Poisson Independentes com 2 Parâmetros por Time, 1 Fator de Esquecimento, 1 Parâmetro Geral para o Fator Casa e Utilização de Dados Econômicos

Dados os testes que efetuamos anteriormente e os resultados aferidos em termos de desempenho dos modelos, os novos modelos serão implementados com 2 parâmetros por time – representando forças de ataque e de defesa – um fator de esquecimento (1.5), um fator casa único para todos os times (0.25) e uma variável econômica.

O dado de entrada da variável econômica é o valor de mercado do elenco de cada time, avaliado e publicado pelo site [Transfermarkt](#). A tabela 5 mostra os valores dos elencos dos times no início do campeonato brasileiro de 2019.

Tabela 5: Valores dos elencos no início do campeonato brasileiro de 2019

Time	Valor em R\$ milhões		Time	Valor em R\$ milhões
Palmeiras	112.7		Athletico-PR	37
Flamengo	96.25		Vasco	34.6
Santos	94		Bahia	33.1
Grêmio	78.4		Botafogo	32.2
São Paulo	74.6		Chapecoense	25.2
Corinthians	73.65		Goiás	22.4
Internacional	62.45		Ceará	20.15
Cruzeiro	52.5		CSA	18.75
Atlético-MG	46.8		Fortaleza	17.85
Fluminense	42.05		Avaí	10.05

O fator que variamos, chamado fator econômico, é um dimensionador de grandeza dos dados de entrada. Cada time tem seu valor elevado ao quadrado do fator econômico e depois dividido pela média desses valores modificados. Esse último valor corresponde a uma estimativa inicial para o parâmetro de defesa do time. A estimativa inicial para o parâmetro de ataque é uma função do parâmetro de defesa e do número total de gols marcados na edição anterior do campeonato brasileiro. Esses parâmetros iniciais têm uma influência na estimativa dos parâmetros, que diminui a cada rodada, à medida que os resultados do campeonato vão sendo incorporados.

Os valores de teste para o fator econômico são: 0 (sem variável econômica), 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 1 e 2. A tabela 6 mostra o desempenho de cada um dos modelos de teste sob diversas medidas.

Tabela 6: Comparação dos modelos de teste com fator econômico

Modelo	0-1 placar	Prob.m. placar	IGN10 placar	Brier placar	ENERG1 placar
econom 0	0.14398	0.08755	1.16329	0.91154	1.06925
econom 0.1	0.14486	0.08759	1.16316	0.91150	1.06901
econom 0.2	0.14573	0.08771	1.16285	0.91141	1.06839
econom 0.3	0.14748	0.08790	1.16255	0.91129	1.06770
econom 0.4	0.15012	0.08813	1.16261	0.91124	1.06744
econom 0.5	0.15276	0.08837	1.16348	0.91135	1.06828
econom 0.6	0.15188	0.08856	1.16573	0.91176	1.07109
econom 0.7	0.14311	0.08867	1.16997	0.91262	1.07682
econom 1	0.13435	0.08784	1.20033	0.91917	1.12190
econom 2	0.09746	0.07778	1.68436	0.97899	1.50534
Modelo	0-1 result.	Prob.m. result.	IGN10 result.	Brier result.	2RPS result.
econom 0	0.51978	0.39767	0.43573	0.59949	0.41103
econom 0.1	0.51889	0.39801	0.43555	0.59922	0.41078
econom 0.2	0.51539	0.39902	0.43508	0.59850	0.41012
econom 0.3	0.51540	0.40067	0.43446	0.59754	0.40925
econom 0.4	0.51541	0.40292	0.43393	0.59674	0.40855
econom 0.5	0.51803	0.40571	0.43380	0.59659	0.40848
econom 0.6	0.51978	0.40892	0.43447	0.59763	0.40955
econom 0.7	0.51627	0.41240	0.43633	0.60037	0.41220
econom 1	0.49431	0.42265	0.45196	0.62081	0.43104
econom 2	0.47060	0.43106	0.60750	0.70995	0.50050

O modelo com fator econômico 0.4 obteve o melhor desempenho na medida ENERG1. Na medida IGN10 do placar, o modelo com fator 0.3 se saiu melhor. Já para as medidas de resultado IGN10 e 2RPS, o melhor desempenho foi do modelo com fator 0.5. De modo geral, os desempenhos foram bastante próximos para todos os fatores até 0.6, começando a piorar a partir desse ponto.

Aqui, temos muito a destacar. Em primeiro lugar, com relação aos modelos, a maneira pouco usual com que foram implementados abre uma possibilidade que não era considerada para os modelos anteriores: a de fazer previsões para as partidas do primeiro turno do campeonato. Os parâmetros iniciais de cada time, ponderados exclusivamente como função do valor do elenco, servem para prever a primeira rodada. A partir daí, os dados das partidas realizadas são incorporados e passam a fazer parte das estimativas para prever as rodadas posteriores. As previsões do primeiro turno não são objetos dessa dissertação, uma vez que não têm comparabilidade com os modelos anteriores.

Com relação aos dados econômicos, conforme relatado na seção anterior, utilizamos os valores referentes à quinzena imediatamente anterior ao início do campeonato, o que possibilita a previsão da primeira rodada e de qualquer outra subsequente. Contudo, o site disponibiliza uma atualização dos valores de elenco a cada quinzena. A utilização dos dados econômicos mais recentes disponíveis para cada rodada pode melhorar o desempenho dos modelos o que, possivelmente, aumentaria o “valor ótimo” do fator econômico encontrado nos testes. Esse valor ótimo também pode ser diferente quando comparados os desempenhos das previsões das rodadas iniciais do campeonato, em vez de apenas das partidas do retorno, como em nosso caso.

Com relação ao desempenho dos modelos de teste, vemos que alcançaram melhores resultados que os modelos dos testes anteriores para todos os fatores no intervalo $[0, 0.7]$.

A inclusão do fator 0 nesse contexto requer uma reflexão quanto à importância real dos dados econômicos nesses modelos. O fator 0 faz com que o valor de cada elenco seja elevado a 0, o que significa que todos os times têm os mesmos parâmetros iniciais. Na prática, é um modelo baseado somente nos resultados anteriores, que alcançou desempenho semelhante aos que possuem a variável econômica e melhor que os modelos anteriores. Essa melhoria só pode ser explicada pela utilização do número de gols marcados no campeonato anterior na inferência dos parâmetros iniciais de ataque. Trata-se de mais uma evidência de que os modelos podem se beneficiar com o uso de mais dados.

A figura 9 mostra o desempenho dos modelos relativo à medida ENERGI1.

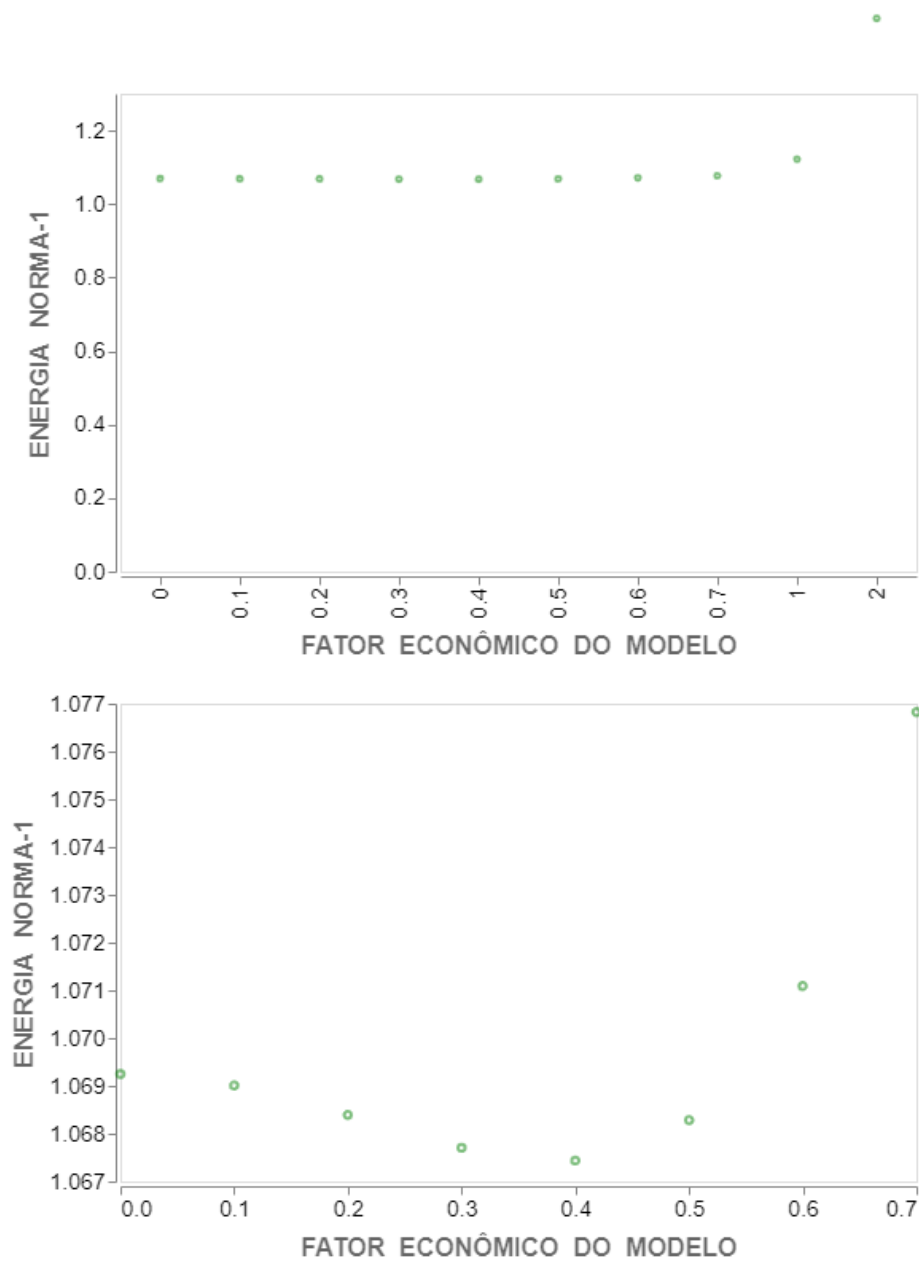


Figura 9: Modelos com variável econômica. No eixo x, o fator econômico do modelo. Na figura de cima, todos os modelos de teste. Na figura de baixo, zoom no eixo y e um recorte com os modelos que tiveram desempenho próximo.

4.1.4 Modelo PIP e as limitações dos Modelos Poisson Independentes

Todos os modelos implementados até aqui utilizam variáveis aleatórias independentes com distribuição de Poisson para estimar o número de gols de cada equipe numa partida. Cada variável Poisson tem um parâmetro λ , usado no cálculo das probabilidades de cada número específico de gols.

A distribuição de Poisson possui a peculiaridade de que o parâmetro λ é, ao mesmo tempo, a moda, a média e a variância da distribuição. Assim, para maximizar a probabilidade de um placar na previsão, basta que tomemos os λ s idênticos ao número de gols marcados por cada time.

Por exemplo, se o placar de uma partida foi 1 x 0, a maior probabilidade possível que um modelo Poisson independentes pode atribuir é:

$$Prob(placar\ mandante = 1) \times Prob(placar\ visitante = 0) \approx 36.8\%.$$

Nenhum modelo Poisson independentes poderá prever uma probabilidade maior que essa, mesmo fazendo “previsão a posteriori”. Para placares com mais gols essa probabilidade fica ainda menor. Por exemplo, para um 2 x 2, a probabilidade prevista nunca será superior a 7.33%, o que é um limitador bastante severo.

A figura 10 exibe o limite máximo da probabilidade que um modelo baseado em Poisson independentes pode prever para cada placar.

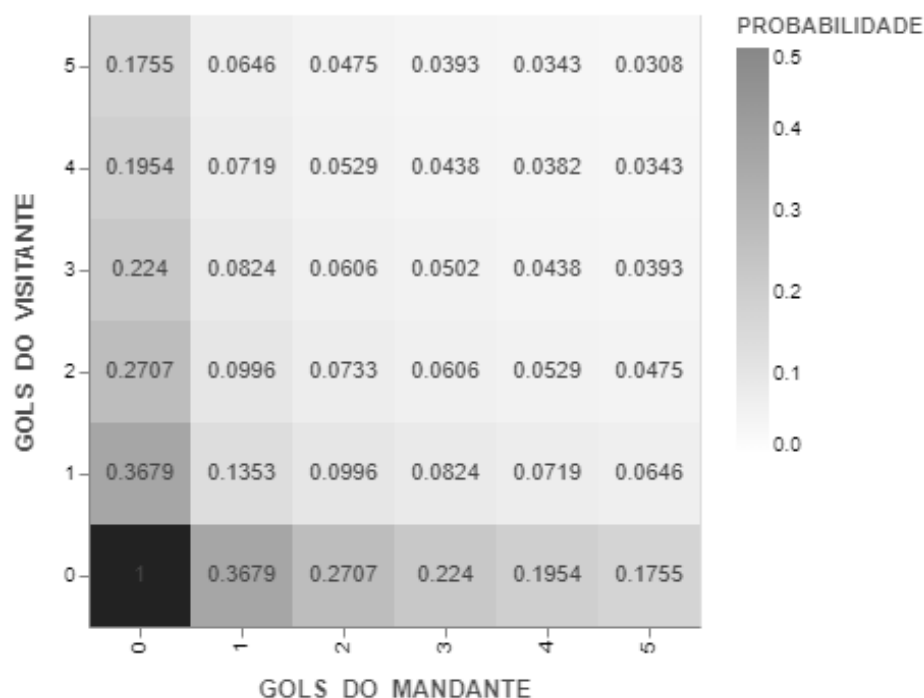


Figura 10: Limites máximos para as probabilidades num modelo baseado em Poisson independentes.

O modelo Poisson Independentes Perfeito (PIP) foi implementado para prever a probabilidade máxima para cada placar ocorrido em cada partida prevista. Ele utiliza como λ s o número de gols marcados por cada time.

Dessa maneira, nenhum modelo baseado em Poisson independentes pode obter desempenho melhor que o PIP, o que o torna uma boa referência na comparação com os demais.

É claro que o desempenho do modelo PIP é exageradamente superior ao dos modelos que implementamos, sobretudo, pelas altas probabilidades previstas nas partidas em que um dos times, ou ambos, não marca gols.

A tabela 7 apresenta o desempenho do modelo PIP em cada temporada entre 2014 e 2019.

Cabe ressaltar que a variação no desempenho se deve, exclusivamente, à distribuição dos placares na temporada. Quanto mais placares baixos (ou placares onde ao menos um time não marque gols) e quanto menos placares altos (ou onde ambos os times marquem gols), melhor o desempenho.

Tabela 7: Desempenho do modelo PIP ao longo das temporadas 2014-2019

Temporada	0-1 placar	Prob.m. placar	IGN10 placar	Brier placar	ENERG1 placar
2014	1	0.24449	0.73155	0.71075	0.39242
2015	1	0.26997	0.72086	0.68819	0.39214
2016	1	0.27520	0.68591	0.67960	0.36712
2017	1	0.23204	0.76267	0.72498	0.40397
2018	1	0.30457	0.66023	0.65301	0.34933
2019	1	0.2673	0.71138	0.68961	0.38002
Temporada	0-1 result.	Prob.m. result.	IGN10 result.	Brier result.	2RPS result.
2014	0.83684	0.66165	0.20815	0.2588	0.12675
2015	0.86842	0.70996	0.17835	0.21477	0.10677
2016	0.83069	0.66009	0.21419	0.27117	0.12750
2017	0.75789	0.61057	0.25077	0.31951	0.14755
2018	0.81579	0.66481	0.21058	0.26725	0.12645
2019	0.82105	0.66509	0.20973	0.26244	0.12441

Outro ponto importante a destacar é que, embora o modelo PIP obtenha índice 1 na medida 0-1 de placares, ou seja, sempre atribua maior probabilidade para o placar que, de fato, ocorre, o mesmo não acontece na medida 0-1 de resultados.

Isso é outra peculiaridade dos modelos Poisson independentes. Para uma partida que termine em 1 x 1, por exemplo, a probabilidade que o modelo PIP atribui ao resultado de empate é de cerca 30.8%, enquanto a probabilidade para cada resultado de vitória é aproximadamente 34.6%. Para empates com mais gols, essa diferença é ainda maior.

A figura 11 mostra a avaliação do PIP, por temporada, segundo a medida ENERG1.

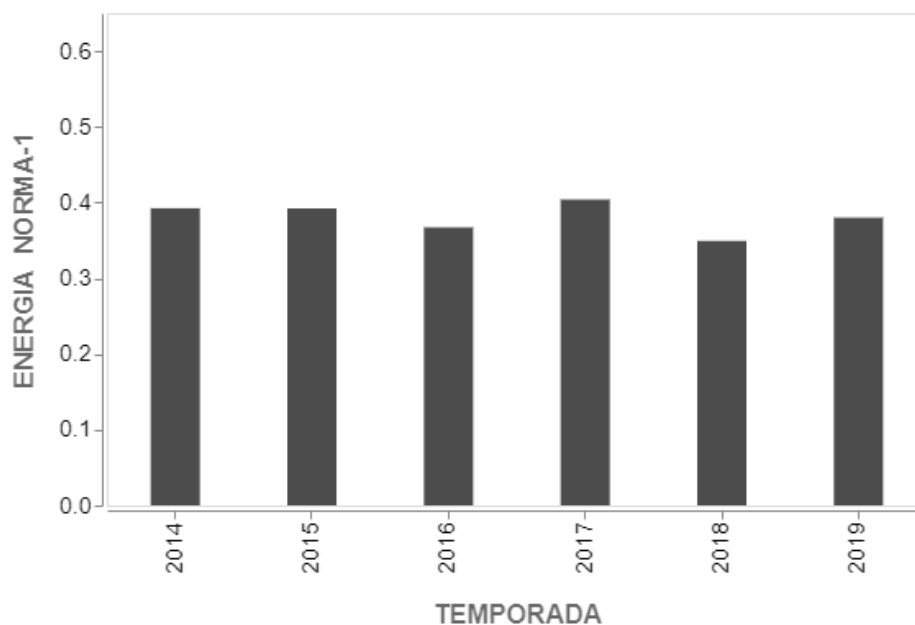


Figura 11: Desempenho do modelo PIP a cada temporada na medida ENERGA1.

4.2 Comparativo entre Modelos

Nesta seção vamos apresentar a comparação entre os quatro modelos que implementamos – aquele usado pela EMap; o de 4 parâmetros por time com fator de esquecimento 1.5; o de 2 parâmetros por time com fator casa 0.25; e o que utiliza fator econômico 0.4 – e os modelos de *benchmarking* – PIP, PIS e FREQ.

Como apresentados na seção 3, o modelo PIS é uma variação do modelo PIP, onde o parâmetro de cada Poisson é sorteado de uma distribuição normal, enquanto o FREQ é um modelo ingênuo, em que todas as previsões são iguais e correspondem à frequência relativa de cada placar na base de jogos.

O modelo PIS é considerado uma referência de desempenho satisfatório a ser alcançado pelos modelos. Já o modelo FREQ pode ser interpretado como uma referência de desempenho mínimo, uma vez que ignora completamente a qualidade dos times envolvidos na partida.

A figura 12 é um recorte de duas rodadas do campeonato brasileiro de 2019, com os placares das partidas e a avaliação de cada modelo na medida ENERGA1.

ano	rodada	mandante	placar	visitante	EMAp	4param	2param	Elenco	FREQ	PIS	PIP
2019	21	Flamengo	3x1	Internacional	0.710	0.734	0.838	0.992	1.289	0.737	0.600
2019	21	Ceará	0x0	Cruzeiro	1.776	1.798	1.235	1.077	1.241	0.919	0.000
2019	21	Athletico-PR	4x1	Fortaleza	2.249	2.262	1.752	1.957	2.190	0.693	0.667
2019	21	Chapecoense	0x1	Corinthians	0.500	0.502	0.516	0.568	1.038	2.307	0.212
2019	21	Fluminense	1x1	Santos	0.548	0.523	0.547	0.456	0.462	1.922	0.424
2019	21	Atlético-MG	1x2	Vasco	0.886	1.034	1.068	1.175	1.017	0.968	0.523
2019	21	Bahia	2x0	Botafogo	0.776	0.843	0.598	0.724	0.829	1.290	0.311
2019	21	Grêmio	6x1	Avaí	3.893	3.802	3.657	3.642	4.146	1.240	0.799
2019	21	São Paulo	0x1	Goiás	2.545	2.358	1.681	1.450	1.038	0.406	0.212
2019	21	Palmeiras	6x2	CSA	5.113	4.937	4.933	4.764	4.701	2.062	0.909
2019	30	Fluminense	0x0	Vasco	0.665	0.759	0.812	1.045	1.241	0.524	0.000
2019	30	Palmeiras	1x0	Ceará	0.852	0.776	0.668	0.605	0.666	0.532	0.212
2019	30	Flamengo	4x1	Corinthians	1.939	1.894	1.877	2.008	2.190	1.412	0.667
2019	30	Athletico-PR	1x0	CSA	0.540	0.528	0.924	0.703	0.666	0.513	0.212
2019	30	Cruzeiro	1x1	Bahia	0.574	0.548	0.447	0.450	0.462	1.069	0.424
2019	30	Grêmio	2x0	Internacional	0.833	0.934	0.663	0.716	0.829	0.354	0.311
2019	30	Fortaleza	2x2	Atlético-MG	1.502	1.412	1.000	1.051	1.180	1.505	0.622
2019	30	Santos	4x1	Botafogo	1.571	1.614	1.911	1.990	2.190	0.704	0.667
2019	30	Goiás	2x0	Avaí	0.500	0.486	0.528	0.562	0.829	1.239	0.311
2019	30	Chapecoense	0x3	São Paulo	2.156	1.998	1.885	1.946	2.489	0.678	0.388

Figura 12: Duas rodadas da temporada 2019 com placares e avaliações dos modelos na medida ENERGI.

As comparações seguirão o seguinte roteiro: primeiro vamos considerar o desempenho acumulado em todo o período de 2014 a 2019, verificando diversas medidas e efetuando a comparação segundo as quatro medidas de interesse – ENERGI e IGN10 para placares, 2RPS e IGN10 para resultados. Depois, vamos comparar o desempenho dos modelos em cada uma das seis temporadas, tentando observar e entender as variações.

A tabela 8 mostra o desempenho em todo o período de cada um dos modelos sob diversas medidas, nos moldes já apresentados em seções anteriores.

Tabela 8: Comparação entre os sete modelos finais

Modelo	0-1 placar	Prob.m. placar	IGN10 placar	Brier placar	ENERG1 placar
EMAp	0.14046	0.08754	1.21404	0.92602	1.14277
4param	0.15099	0.08764	1.21070	0.92444	1.13813
2param	0.14221	0.08813	1.18086	0.91739	1.09574
Elenco	0.15012	0.08813	1.16261	0.91124	1.06744
FREQ	0.15632	0.08713	1.17018	0.91286	1.09111
PIS	0.18526	0.10602	1.13341	0.88974	1.02774
PIP	1	0.26559	0.71210	0.69102	0.38083
Modelo	0-1 result.	Prob.m. result.	IGN10 result.	Brier result.	2RPS result.
EMAp	0.50134	0.40846	0.45038	0.61774	0.42866
4param	0.50223	0.40720	0.44860	0.61626	0.42738
2param	0.49784	0.40604	0.44166	0.60725	0.41810
Elenco	0.51541	0.40292	0.43393	0.59674	0.40855
FREQ	0.50925	0.37977	0.44856	0.62011	0.43077
PIS	0.61813	0.50354	0.39307	0.52147	0.33173
PIP	0.82178	0.66203	0.21196	0.26566	0.12657

Analisando os resultados, chama a atenção a disparidade de desempenho do modelo PIP para todos os outros e ainda, surpreendentemente, a proximidade de desempenho do modelo FREQ para os demais, chegando a ser ranqueado em terceiro em algumas medidas.

Se levarmos em conta a medida ENERG1, o modelo com variável econômica é o único, dos quatro que implementamos, que conseguiu superar o FREQ no período analisado. Os outros três modelos ficaram abaixo da nossa referência mínima, ou seja, tiveram desempenho abaixo de um modelo que desconsidera a qualidade dos times que disputam cada partida.

A figura 13 mostra os resultados de todos os modelos na medida ENERG1.

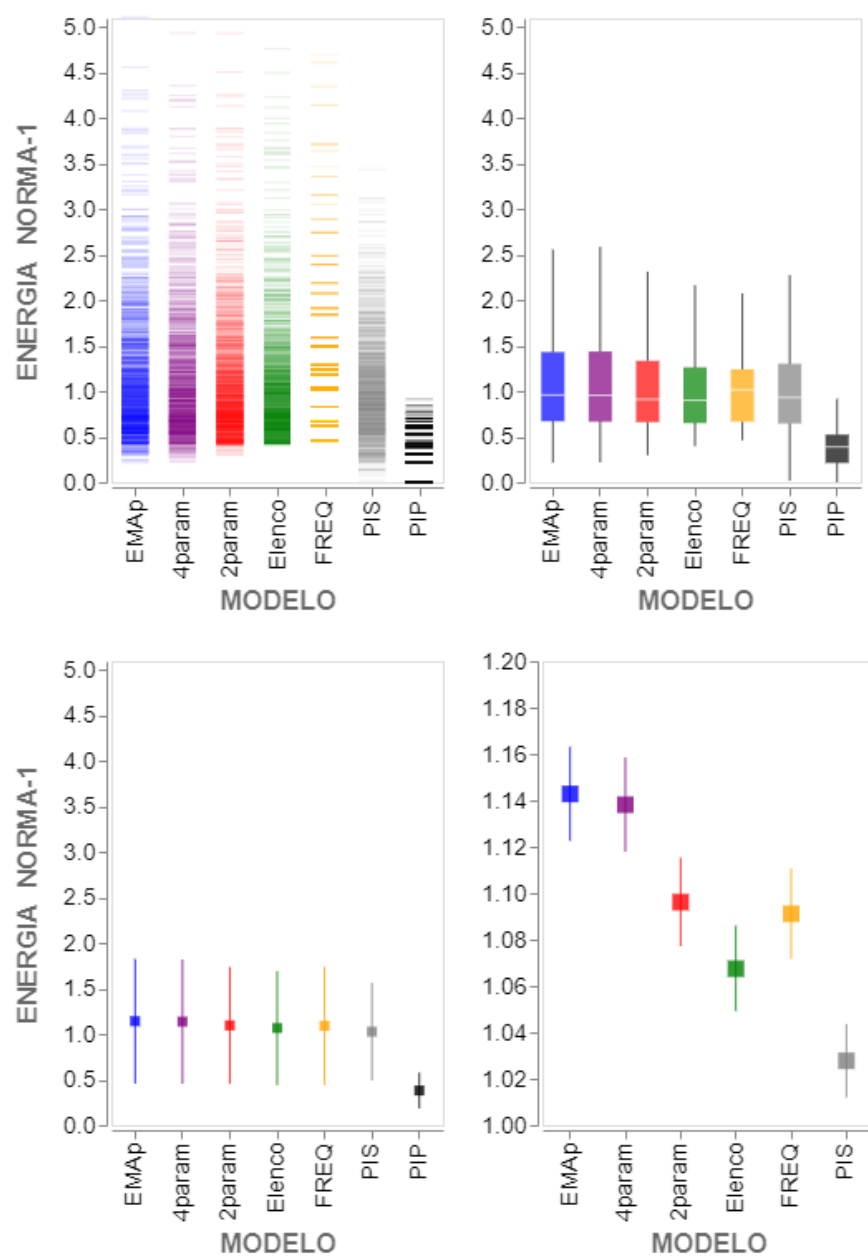


Figura 13: Desempenho dos sete modelos na medida ENERGA1. Em cima, à esquerda, a avaliação de todas as partidas; à direita, o box-plot dessa avaliação. Embaixo, à esquerda, média e desvio-padrão; à direita, zoom no eixo y, média e erro-padrão das seis medidas com avaliação no intervalo.

Quando avaliados pela medida IGN10 de placares, os modelos foram ranqueados na mesma ordem da medida ENERGI1, sendo notadas apenas pequenas variações na proximidade entre as avaliações, como, por exemplo, o modelo FREQ se distanciando do modelo 2param.

A figura 14 apresenta, para cada modelo, a avaliação nas duas medidas plotadas num gráfico bidimensional. É possível perceber uma forte correlação entre as medidas, que parece diferente somente nos modelos PIS e PIP.

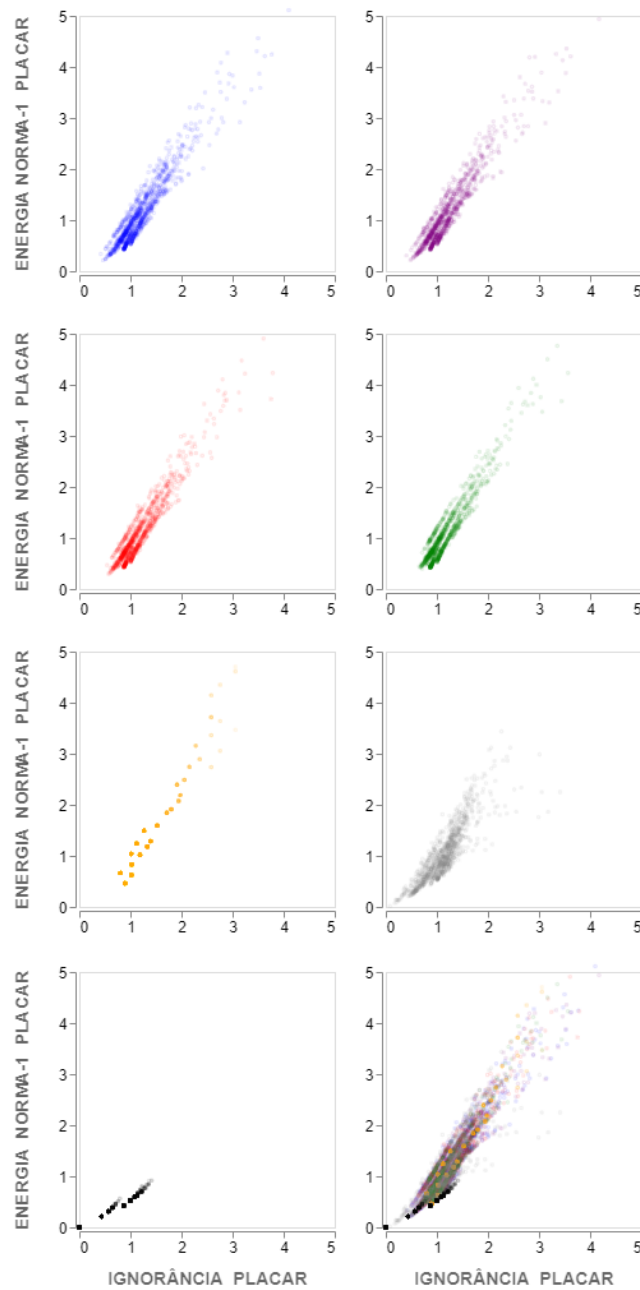


Figura 14: Gráficos IGN10 (placares) x ENERG1 de todas as partidas para cada um dos sete modelos e para todos os modelos na mesma plotagem. EMap e 4param na linha de cima. Nas demais linhas temos, de cima para baixo: 2param e Elenco, FREQ e PIS, PIP e todos.

Para previsões de resultados, o modelo PIS obteve um segundo lugar bem consolidado, alcançando um desempenho muito superior ao dos modelos com previsão *'a priori'*. O modelo com 2 parâmetros por time se mostrou superior ao FREQ, que teve desempenho nivelado aos dois modelos com 4 parâmetros por time, sendo inferior no 2RPS, mas superior no IGN10.

As figuras 15 e 16 seguem os moldes das duas figuras anteriores, apenas trocando as medidas de avaliação por 2RPS e IGN10 de resultados.

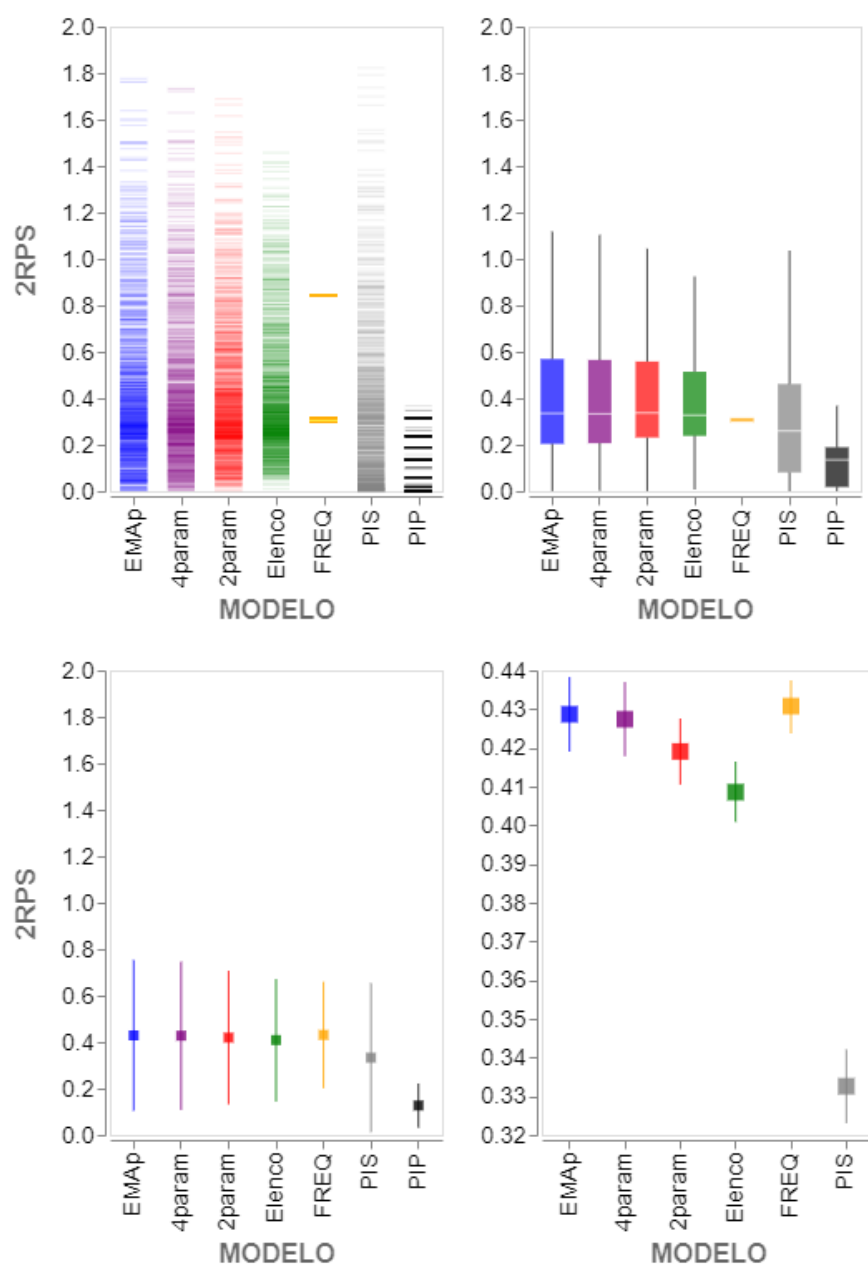


Figura 15: Desempenho dos sete modelos na medida 2RPS. Em cima, à esquerda, a avaliação de todas as partidas; à direita, o box-plot dessa avaliação. Embaixo, à esquerda, média e desvio-padrão; à direita, zoom no eixo y, média e erro-padrão das seis medidas com avaliação no intervalo.

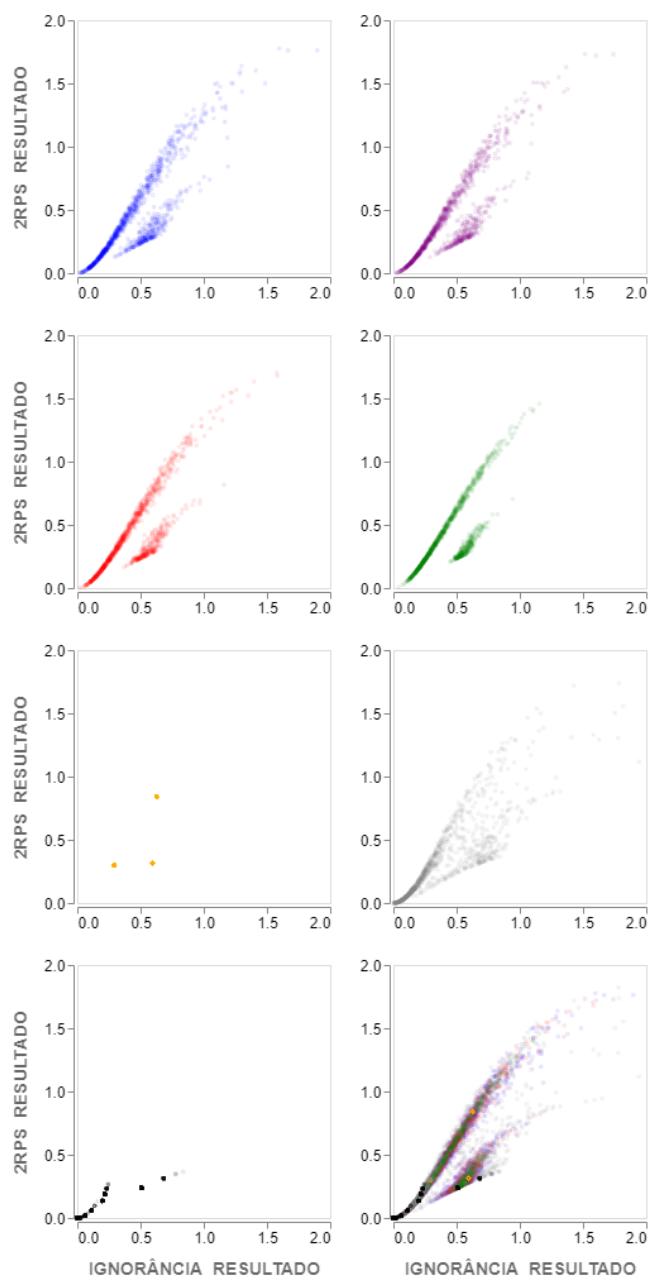


Figura 16: Gráficos IGN10 (resultados) x 2RPS de todas as partidas para cada um dos sete modelos e para todos os modelos na mesma plotagem. EMaP e 4param na linha de cima. Nas demais linhas temos, de cima para baixo: 2param e Elenco, FREQ e PIS, PIP e todos.

É interessante notar a diferença no formato dos gráficos bidimensionais da figura 16 em relação aos da figura 14.

Em primeiro lugar, a correlação entre as medidas deixa de parecer uma linha reta e passa a se assemelhar a uma sigmoide de pequena curvatura.

Em segundo, os gráficos parecem conter duas componentes conexas. Uma, mais comprida e estreita, referente aos resultados de vitória e outra, mais curta e larga, correspondendo aos empates. Isso se deve, principalmente, ao fato de a 2RPS ser sensível a distâncias, penalizando menos a avaliação de um empate, ao passo que a IGN10 penaliza de acordo com a probabilidade atribuída ao resultado ocorrido, independentemente de qual seja. Assim, uma previsão de $(1/3, 1/3, 1/3)$ teria avaliação de 0.477 segundo a IGN10, independentemente do resultado, mas sua avaliação na 2RPS mudaria de 0.556 em caso de vitória de um dos times para 0.222 em caso de empate.

Esse efeito também está presente na previsão de placares, mas é bem menos perceptível, uma vez que o espaço de placares possíveis tem cardinalidade muito maior e as previsões são menos concentradas, com probabilidades menores previstas para cada evento.

Um terceiro ponto a observar na figura 16 é a diferença dos gráficos entre os modelos. Os de 4 parâmetros por time apresentam dispersão bem maior que o de 2 parâmetros por time, que, por sua vez, apresenta mais dispersão que o modelo com variável econômica.

Essa característica também estava presente e menos perceptível na figura 14, o que pode sugerir que os modelos de previsão '*a priori*' que obtiveram melhores avaliações tenham menor variabilidade em suas previsões, estando, possivelmente, um pouco mais próximos do modelo FREQ de previsões constantes. A última hipótese precisa ser investigada com mais cuidado. Um bom caminho pode ser através da decomposição das avaliações, explicitando atributos como o refinamento das previsões.

O ranqueamento dos quatro modelos com previsão '*a priori*' é o mesmo em todas as quatro medidas de interesse, bem como, em quase todas as outras medidas. O mesmo ocorre em relação ao ranqueamento dos três modelos com previsão '*a posteriori*'. Se consideramos somente as duas medidas de interesse relativas a previsão de placares, o ranqueamento de todos os sete modelos é o mesmo.

Nesta segunda etapa, vamos efetuar a comparação entre os modelos temporada a temporada. A tabela 9 exibe a avaliação média em cada temporada de três dos sete modelos nas quatro medidas de interesse.

Tabela 9: Comparação ano a ano dos modelos EMap, Elenco e PIS

Ano	Modelo	ENERG1 placar	IGN10 placar	2RPS result.	IGN10 result.
2014	EMAp	1.2187	1.2820	0.4500	0.4560
2014	Elenco	1.1024	1.1913	0.4075	0.4218
2014	PIS	1.0702	1.1575	0.3302	0.3780
2015	EMAp	1.2458	1.2735	0.4420	0.4518
2015	Elenco	1.1926	1.2318	0.4315	0.4427
2015	PIS	1.0228	1.1277	0.3009	0.3488
2016	EMAp	1.0899	1.1683	0.4022	0.4329
2016	Elenco	1.0493	1.1430	0.3964	0.4214
2016	PIS	1.0576	1.1524	0.3223	0.3888
2017	EMAp	1.1495	1.2346	0.4938	0.5019
2017	Elenco	1.0615	1.1804	0.4684	0.4838
2017	PIS	1.0208	1.1510	0.3441	0.4162
2018	EMAp	1.0338	1.1376	0.3780	0.4241
2018	Elenco	0.9595	1.0882	0.3531	0.4091
2018	PIS	0.9953	1.1018	0.3668	0.4378
2019	EMAp	1.1189	1.1882	0.4060	0.4357
2019	Elenco	1.0394	1.1410	0.3944	0.4247
2019	PIS	0.9996	1.1101	0.3312	0.3978

As figuras 17 e 18 exibem a evolução das medidas ENERG1 e 2RPS a cada temporada para todos os sete modelos.

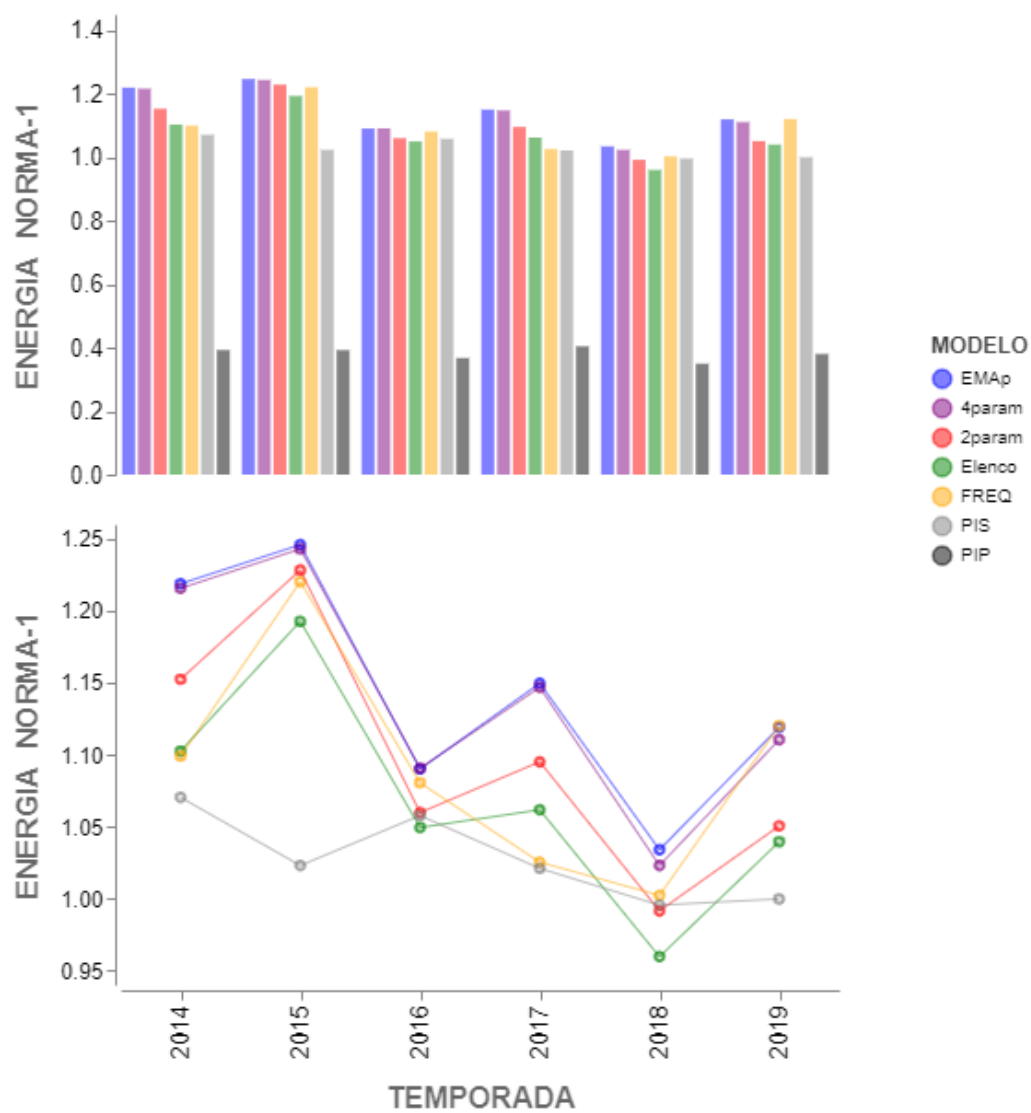


Figura 17: Desempenho médio, a cada temporada, dos sete modelos segundo a medida ENERGI1. Embaixo, zoom no eixo y e linhas de desempenho dos seis modelos com médias no intervalo.

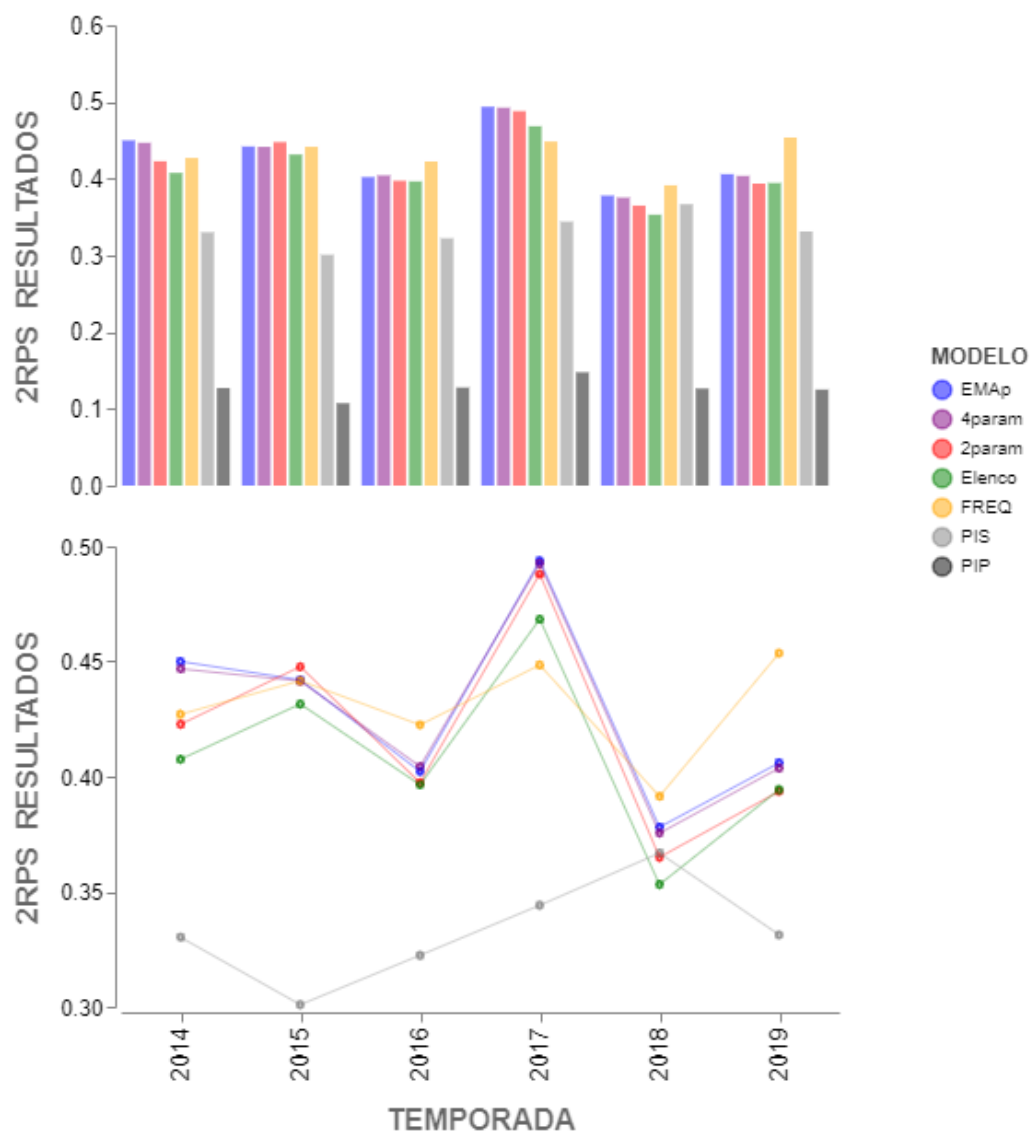


Figura 18: Desempenho médio, a cada temporada, dos sete modelos segundo a medida 2RPS para previsão de resultados. Embaixo, zoom no eixo y e linhas de desempenho dos seis modelos com médias no intervalo.

É interessante notar que os quatro modelos de previsão ‘*a priori*’ mantêm o mesmo ranqueamento para quase todas as medidas e temporadas, embora exista certa variação na distância entre eles.

No comparativo com os outros modelos, o ranqueamento muda bastante em relação ao FREQ – ora no meio, ora melhor, ora pior que todos os quatro – e ocasionalmente em relação ao PIS, que em 2018 foi superado pelo Elenco e se mostrou nivelado ao 2param.

A temporada de 2016 também tem números curiosos, com os seis modelos muito próximos nas medidas de previsões de placares, mas com certo distanciamento, nas medidas de previsões de resultados, dos quatro ‘*a priori*’ (todos muito próximos) para o FREQ (pior) e o PIS (muito melhor). Já na temporada de 2017, os quatro foram superados pelo FREQ, tanto nas previsões de placares como nas de resultados.

É possível perceber ainda uma tendência compartilhada entre os quatro modelos ‘*a priori*’, sobretudo entre EMap e 4param, que caminham quase juntos, e entre 2param e Elenco, que caminham em paralelo mantendo certa distância.

O modelo FREQ parece acompanhar a tendência dos outros quatro, mas com quedas e subidas ligeiramente mais suaves. Já o PIS aponta uma tendência inversa, piorando sua avaliação nos anos em que os outros cinco melhoram e vice-versa. Isso ocorreu na sequência de todos os anos, exceto de 2017 para 2018 na ENERGI e de 2016 para 2017 na 2RPS.

4.3 Decomposições das Avaliações

Neste item, vamos decompor a medida 2RPS em somas de atributos, conforme descrito em seções anteriores.

4.3.1 Modelo FREQ e a Incerteza dos Eventos Ocorridos

A incerteza é um atributo exclusivo dos resultados e, uma vez que depende das previsões, deve ser a mesma para todos os modelos.

Como observamos anteriormente, todo modelo com previsões constantes tem resolução nula. É o caso do modelo FREQ, que, além disso, prevê, para

todas as partidas, probabilidades correspondentes à frequência relativa de cada resultado no conjunto de partidas. Isso faz com que o modelo *FREQ* seja perfeitamente confiável, e, portanto, seu atributo (falta de) confiabilidade deve ser nulo.

Finalmente, usando a decomposição *URR* para a avaliação do modelo *FREQ*, temos:

$$\mathbb{E}_{P,X}[S(P, X)] = UNC - RES + REF = UNC - 0 + 0 = UNC.$$

Isso quer dizer que, independentemente da medida própria utilizada, a avaliação média do modelo *FREQ* deve ser igual ao valor do atributo incerteza dos resultados.

O mesmo vale para previsões de placares, já que o modelo *FREQ* também é constante em prever a frequência relativa de cada placar no conjunto de partidas.

Quando avaliamos os atributos da decomposição *RDC*, fica claro que as previsões de um modelo constante não têm variabilidade, e portanto, seu refinamento é nulo. Da mesma forma, a discriminação, que compara previsões condicionais aos resultados à previsão incondicional, também é nula. Temos então:

$$\mathbb{E}_{P,X}[S(P, X)] = REF - DIS + COR = 0 - 0 + COR = COR.$$

E a (falta de) correção corresponde a todo o valor aferido para a acurácia, o que vale para o *FREQ* e também para qualquer outro modelo constante.

4.3.2 Decomposições da medida *2RPS*

Vamos agora decompor em soma de atributos as avaliações dos modelos, segundo a medida *2RPS*, que apresentamos no item anterior.

Utilizando as fórmulas para a decomposição *URR* apresentadas no item 3.5.2, encontramos os valores exibidos na tabela 10.

A figura 19 mostra a avaliação dos atributos referentes à decomposição *URR* para cada um dos modelos.

Tabela 10: Atributos URR das previsões dos modelos na medida 2RPS

Modelo	2RPS	UNC	RES	REL
EMAp	0.4287	0.4308	0.4308	0.4287
4param	0.4274	0.4308	0.4308	0.4274
2param	0.4181	0.4308	0.4308	0.4181
Elenco	0.4086	0.4308	0.4308	0.4086
FREQ	0.4308	0.4308	0	0
PIS	0.3317	0.4308	0.4308	0.3317
PIP	0.1266	0.4308	0.4308	0.1266

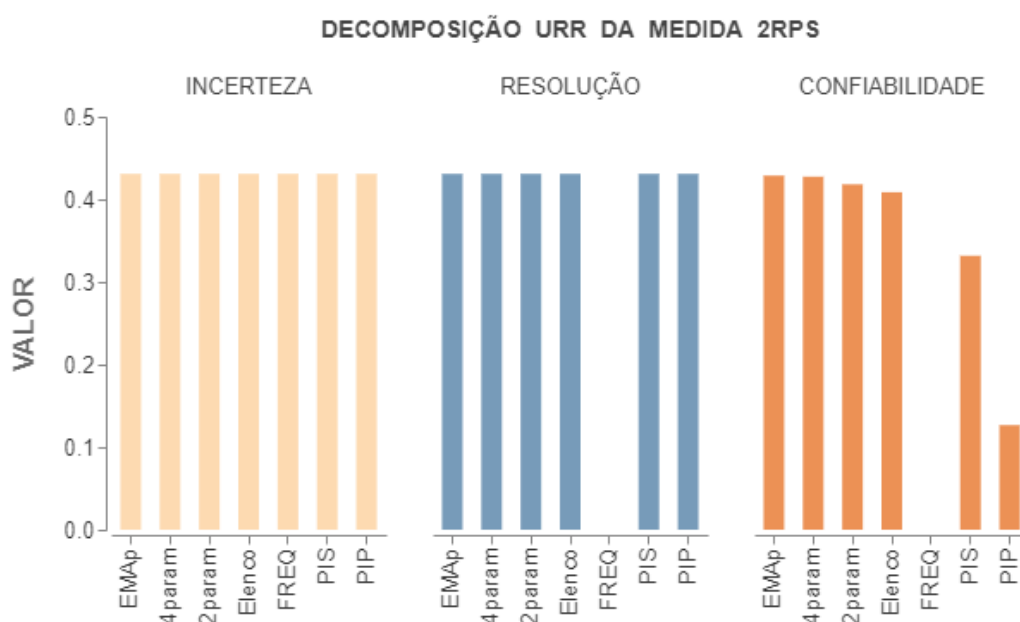


Figura 19: Atributos URR para os sete modelos na medida 2RPS

Cabe aqui ressaltar que o problema da esparsidade, discutido no item 2.4.6, fez com que a resolução ficasse igual à incerteza dos resultados para todos os modelos à exceção do FREQ. Isso ocorre porque a resolução compara a observação média incondicional e as observações médias condicionais a cada previsão e, como cada previsão aparece uma única vez no conjunto de previsões, essas observações médias condicionais se tornam a própria observação (resultado) de cada partida. Problema semelhante ocorre na estima-

tiva da confiabilidade, que também usa as observações médias condicionais no cálculo. Assim, esses dois atributos devem estar superestimados, i.e., os valores encontrados são maiores do que deveriam ser na realidade.

A tabela 11 e a figura 20 apresentam os atributos referentes à decomposição RDC.

Tabela 11: Atributos RDC das previsões dos modelos na medida 2RPS

Modelo	2RPS	REF	DIS	COR
EMAp	0.4287	0.0622	0.0030	0.3696
4param	0.4274	0.0592	0.0029	0.3711
2param	0.4181	0.0454	0.0024	0.3751
Elenco	0.4086	0.0260	0.0017	0.3842
FREQ	0.4308	0	0	0.4308
PIS	0.3317	0.1503	0.0478	0.2292
PIP	0.1266	0.1801	0.1484	0.0948

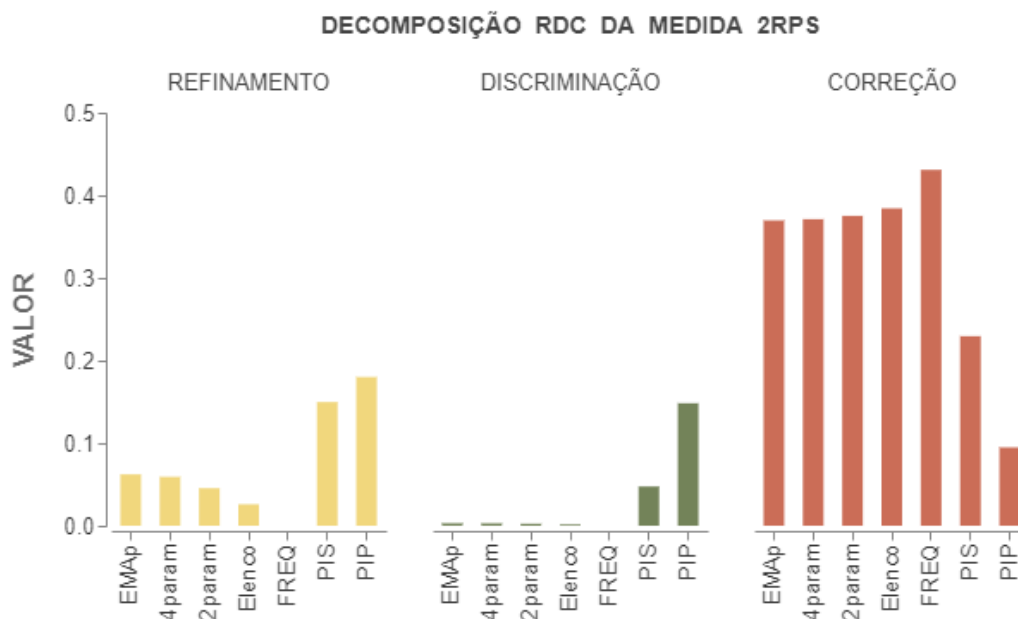


Figura 20: Atributos RDC para os sete modelos na medida 2RPS

Neste caso, o problema da esparsidade não afeta as avaliações, uma vez que o espaço de resultados contém apenas três pontos e todos ocorreram um número considerável de vezes, mais de 20% do total de partidas.

São dignos de nota os pequenos valores de refinamento e discriminação encontrados para os quatro modelos de previsão ‘*a priori*’. O refinamento é uma medida da variabilidade das previsões e a discriminação mede o quanto cada diferente resultado é precedido, em média, por diferentes previsões.

Um baixo refinamento indica que as previsões desses modelos não variaram muito, ficando usualmente próximas à previsão média. Uma baixa discriminação indica que as previsões médias condicionais não são muito diferentes da previsão média incondicional, e, conseqüentemente, também não são muito diferentes entre si.

A título de comparação, mostramos abaixo as matrizes de discriminação dos modelos EMap e PIP, onde cada linha corresponde a um resultado e cada coluna à previsão média dado um resultado. Assim, por exemplo, a entrada 1,2 da matriz corresponde à probabilidade média atribuída ao empate nas partidas com vitória do mandante. Modelos com boa discriminação devem ter valores altos na diagonal principal da matriz e baixos nas demais entradas. Modelos com discriminação ruim têm, nas colunas, valores muito próximos entre si e, em cada linha, valores próximos à distribuição incondicional dos resultados.

$$EMAp : \begin{bmatrix} 0.524 & 0.254 & 0.222 \\ 0.452 & 0.274 & 0.274 \\ 0.428 & 0.268 & 0.304 \end{bmatrix} ; PIP : \begin{bmatrix} 0.736 & 0.206 & 0.057 \\ 0.253 & 0.494 & 0.253 \\ 0.073 & 0.248 & 0.679 \end{bmatrix}$$

Analisando a correção, surge outro destaque bastante curioso: as correções dos quatro modelos *a priori* estão inversamente ordenadas em relação à acurácia desses modelos, i.e., quanto mais incorreto o modelo, melhor sua acurácia, o que não parece fazer sentido. Mais ainda, a mesma ordenação inversa ocorre também na discriminação desses quatro modelos. Quanto menos discriminação, mais acurácia. A explicação para isso está no refinamento, cuja ordenação dos valores acompanha a da acurácia e tem diferenças tais que compensam a ordenação inversa dos outros dois atributos.

É o que ocorre também no exemplo dos modelos de previsão de lançamento de moedas, apresentado no item 2.4.3, e a interpretação também deve ser a mesma. O modelo EMap, ainda que mais correto e com mais discriminação

que os outros três, varia suas previsões em excesso, sem que o resultado correspondente a cada previsão acompanhe essa variação. Isso provavelmente significa ser um modelo menos confiável e, de fato, é o modelo que tem a pior avaliação para esse atributo.

Ao modelo Elenco, ocorre o oposto. Apesar de ser mais incorreto e menos discriminativo que os demais, varia suas previsões “quando necessário”, se tornando um modelo mais confiável e com melhor acurácia. É talvez um modelo “a meio caminho entre o EMap e o FREQ”, tendo menos da metade do refinamento do primeiro, sem perder tanto em correção (e obviamente em discriminação) quanto o modelo constante.

Mudando o foco da análise para os outros dois modelos de referência, PIS e PIP, é possível perceber que têm muito mais refinamento que os demais. O PIP converte a maior parte de seu refinamento em discriminação e alcança um nível baixo de (falta de) correção. O PIS, por outro lado, tem discriminação bem menor que seu refinamento, o que deixa sua correção em nível intermediário entre o PIP e os demais modelos.

Finalmente, a figura 21 exibe, conjuntamente, as duas decomposições URR e RDC da medida 2RPS para todos os modelos.

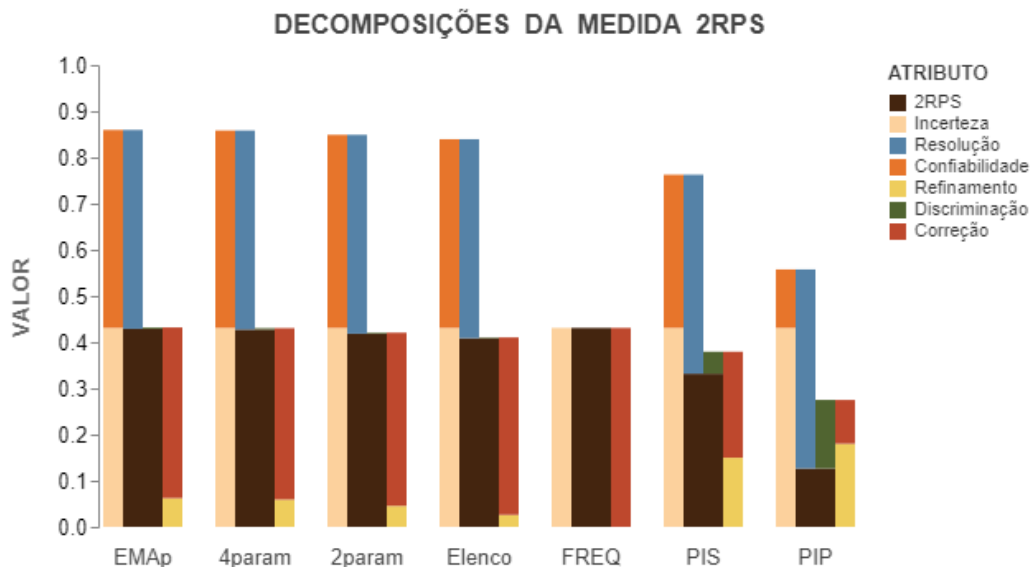


Figura 21: Decomposições da medida 2RPS para os sete modelos

5 Conclusão

Na seção 2, revisamos a literatura sobre previsões, modelos de previsões de futebol, medidas de avaliação de previsões e suas decomposições em somas de atributos.

Na seção 3, apresentamos nossa base de dados, os métodos de implementação dos modelos, nossa seleção de medidas de avaliação e as expressões de suas decomposições URR e RDC.

Na seção 4, definimos quatro modelos de previsão para placares e resultados de futebol e comparamos seu desempenho ao de outros três modelos que consideramos para *benchmarking*. Vimos também as decomposições da avaliação dos modelos em somas de atributos, segundo a medida 2RPS.

De modo geral, os quatro modelos performaram aquém do esperado, ficando muito mais próximos à nossa referência mínima de desempenho do que àquela considerada satisfatória. Esse problema se verificou para previsões de resultados e foi ainda mais grave para previsões de placares, onde alguns modelos tiveram desempenho inferior ao modelo ingênuo frequentista.

Além disso, após analisar as decomposições das medidas, constatamos que os melhores modelos implementados foram os menos assertivos, no sentido de que suas previsões tiveram menos refinamento e discriminação. Isso parece ir de encontro ao que se espera de uma boa previsão. Um bom modelo de previsão deve ser tão confiável e assertivo quanto possível, para que suas previsões sejam bons indicativos dos eventos que virão a ocorrer.

Com tudo isso, é necessário refletir e buscar novas alternativas para melhorar os modelos de previsão. É possível que os modelos que utilizamos, com variáveis Poisson independentes, não sejam os mais adequados para esse tipo de previsão. Também é possível que os dados utilizados – placares de partidas anteriores de um mesmo campeonato – não sejam suficientes para gerar boas previsões.

Há ainda a possibilidade de que a irregularidade dos times participantes do campeonato brasileiro torne os resultados e placares individuais tão imprevisíveis que um modelo ingênuo, que não considere a qualidade de cada time envolvido na partida, possa levar vantagem. Essa última possibilidade, se verdadeira, poderia inviabilizar o bom desempenho de qualquer modelo de previsão, tornando a tarefa de prever resultados dessas partidas algo se-

melhante a prever o resultado do lançamento de moedas, o *cara* ou *coroa*. Esperamos, pois, que ela não seja verdadeira.

No item 2.3, apresentamos as definições de [Murphy \[1993\]](#) para os três tipos que caracterizam uma boa previsão. Ao longo deste trabalho, exploramos somente os dois primeiros tipos - consistência e qualidade, tendo faltado o terceiro: a utilidade ou valor da previsão.

Medidas de utilidade são muitas vezes associadas a algum ganho econômico que as previsões possam gerar. Um bom exemplo seria um o lucro auferido com apostas baseadas no modelo de previsão. Outra possibilidade seria as regras de pontuação de um concurso de placares ou resultados de futebol, o popular *bolão*.

As medidas de utilidade, em sua maioria, não são próprias. Muitas vezes, uma estratégia de previsões consistentes não é a que tem o melhor retorno esperado. Mais ainda, o modelo de previsão com mais acurácia não necessariamente é o que alcança o melhor desempenho numa medida de utilidade. Apesar disso, ainda que o objeto de interesse do preditor seja maximizar/minimizar uma medida de utilidade não própria, a avaliação do modelo por uma medida própria de acurácia e sua decomposição em somas de atributos, possibilita um melhor entendimento de seu desempenho e a identificação de quais atributos estão sobressaindo.

Analisando pela ótica oposta, a identificação antecipada de atributos desejáveis pode servir de norte para a implementação de um modelo destinado a maximizar/minimizar uma medida de utilidade não própria. Nesse caso, a utilização de uma medida própria de acurácia, que possa ser decomposta em atributos, pode ser essencial para o diagnóstico e o aperfeiçoamento do modelo.

Tendo em vista todas as considerações já expostas, consideramos que os principais objetivos deste trabalho foram atingidos: a definição de medidas adequadas para avaliar previsões de placares e resultados de futebol; a aplicação dessas medidas para avaliar e comparar modelos de previsões para as partidas em nossa base; e a decomposição das avaliações em atributos, de modo a auxiliar a interpretação do desempenho dos modelos.

Como trabalho futuro, esperamos desenvolver as expressões das decomposições em atributos para as outras medidas, em especial, para a medida de energia, e aplicá-las às avaliações dos modelos, de modo a obter os atributos

relativos às previsões de placares.

Também gostaríamos de desenvolver o método de redução de dimensionalidade do espaço de previsões, apresentado por [Mitchell \[2020\]](#), que comentamos no item 2.4.6. Caso implementado, ganharemos compreensão de dois atributos – resolução e confiabilidade – que ficaram superestimados em nossa decomposição.

Outro trabalho considerado é estudar medidas de utilidade relacionadas a estratégias de apostas e sua relação com nossas medidas de acurácia e suas decomposições.

Finalmente, esperamos também aplicar este e os demais trabalhos a outras bases de jogos, investigando campeonatos de outras localidades e outras temporadas.

6 Referências

- J. M. Bernardo. Expected information as expected utility. *The Annals of Statistics*, 7(3):686–690, 1979. ISSN 00905364.
- G. W. Brier. Verification of Forecasts in Terms of Probability. *Monthly Weather Review*, 78(1):1–3, 01 1950. ISSN 0027-0644. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
- J. Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, Jul 2009. ISSN 1477-870X. doi: 10.1002/qj.456.
- A. Constantinou and N. Fenton. Evaluating the predictive accuracy of association football forecasting systems. 2011.
- A. Constantinou and N. Fenton. Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports*, 8, Jan 2012. doi: 10.1515/1559-0410.1418.
- A. C. Constantinou. Dolores: a model that predicts football match outcomes from all over the world. *Machine Learning*, 108(1):49–75, 2019.
- M. J. Dixon and S. G. Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280, 1997. doi: 10.1111/1467-9876.00065.
- SportyTell Editors. Top-10 most popular sports in the world 2020 [online], available: <https://sportytell.com/sports/most-popular-sports-world/>, 2018. (accessed Nov.24, 2020).
- E. S. Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology (1962-1982)*, 8(6):985–987, 1969. ISSN 00218952, 2163534X.
- FGV/EMAp Esporte em Números. Modelos matemáticos no campeonato brasileiro 2019 [online], available: <https://www.fgv.br/emap/campeonato-brasileiro/index.html>. (accessed Dec.1, 2020).

- F. Fontanella, A. M. e. Sá, and M. A. H. B. da Silva. Visual analysis of forecasts of football match scores. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 140–147, 2020. doi: 10.1109/SIBGRAPI51738.2020.00027.
- T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477): 359–378, 2007. ISSN 01621459.
- Goal. Brasileirão 2015: Colunistas da goal apresentam suas apostas [online], available: <https://www.goal.com/br/news/619/especiais/2015/05/08/11540792/brasileir2015>. (accessed Dec.1, 2020).
- I. J. Good. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1):107–114, 1952. ISSN 00359246.
- A. J. Lee. Modeling scores in the premier league: Is manchester united really the best? *CHANCE*, 10(1):15–19, 1997. doi: 10.1080/09332480.1997.10554791.
- D. J. C. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
- M. J. Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982. doi: 10.1111/j.1467-9574.1982.tb00782.x.
- K. Mitchell. Score decompositions in forecast verification. 2020.
- A. H. Murphy. A new vector partition of the probability score. *Journal of applied Meteorology*, 12(4):595–600, 1973.
- A. H. Murphy. What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Weather and Forecasting*, 8(2):281–293, 06 1993. ISSN 0882-8156. doi: 10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.
- A. H. Murphy and E. S. Epstein. A note on probability forecasts and hedging. *Journal of Applied Meteorology and Climatology*, 6(6):1002 – 1004, 1967. doi: 10.1175/1520-0450(1967)006<1002:ANOPFA>2.0.CO;2.

- A. H. Murphy and R. L. Winkler. A General Framework for Forecast Verification. *Monthly Weather Review*, 115(7):1330–1338, 07 1987. ISSN 0027-0644. doi: 10.1175/1520-0493(1987)115(1330:AGFFV)2.0.CO;2.
- M. S. Roulston and L. A. Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(6):1653 – 1660, 2002. doi: 10.1175/1520-0493(2002)130(1653:EPFUIT)2.0.CO;2.
- Transfermarkt. Campeonato brasileiro série a [online], available: <https://www.transfermarkt.com.br/campeonato-brasileiro-serie-a/startseite/wettbewerb/bra1>. (accessed Dec.1, 2020).
- E. Wheatcroft. Evaluating probabilistic forecasts of football matches: The case against the ranked probability score. Aug 2019.
- R. L. Winkler and A. H. Murphy. Good probability assessors. *Journal of Applied Meteorology (1962-1982)*, 7(5):751–758, 1968. ISSN 00218952, 2163534X.
- C. Zirpoli. Avaliação do transfermarkt sobre o valor mercado dos 40 elencos das séries a e b [online], available: <http://blogs.diariodepernambuco.com.br/esportes/2016/05/12/transfermarkt-avalia-elencos-de-sport-santa-e-nautico-antes-do-brasileiro-2016/>, 2015. (accessed Dec.1, 2020).