

O Teorema de Perron-Frobenius e Rankings Esportivos

Luís Felipe de Almeida Marques
Jean Fernando Horn

9 de Novembro de 2022

Conteúdo

1	Introdução	1
2	A Motivação	1
3	Método Direto	2
3.1	A intuição direta	2
3.2	O Teorema de Perron-Frobenius	3
3.3	Aplicando o método direto num exemplo real	6
3.4	Especificando a Preferência	10
4	Método de Probabilidades	11
4.1	Achando uma Boa Solução	12
4.2	Prevendo a Copa do Mundo	14
5	Conclusão	15

1 Introdução

Este relatório tem como objetivo apresentar métodos de *rankeamento* de times em competições esportivas pareadas (em que cada partida se dá entre pares de times) e as relações desses métodos com o *Teorema de Perron-Frobenius*. Também são apresentadas aplicações desses métodos com dados reais, através de implementações em *Python*.

2 A Motivação

O interesse geral da população em esportes leva a um interesse em perguntas como "qual o melhor time?", "quem tem mais chances de ganhar?", ou "qual o resultado mais provável?". Essas perguntas podem ser abordadas de várias maneiras, através de várias interpretações e modelagens matemáticas.

Como exemplo, temos o Campeonato Brasileiro de Futebol, discutivelmente a competição esportiva mais popular do Brasil, um campeonato em que todos os competidores enfrentam todos os outros um mesmo número de vezes. Isso é o que chamamos de uma competição balanceada, pela igualdade em quantidades de confrontos entre pares de times.

Em contrapartida, temos a Major League Soccer, liga principal de futebol nos Estados Unidos, que segue o mesmo formato desbalanceado de outras ligas esportivas profissionais e universitárias dos Estados Unidos, em que não há garantia de que todos se enfrentarão um mesmo número de vezes, nem se todos terão a mesma quantidade de jogos ou mesmo se todos os pares de times se enfrentarão.

Surge então a necessidade de achar uma forma de comparar times nesses campeonatos *assimétricos*, e verificar se esse modelo se encaixa também nos campeonatos *mais organizados* como o Brasileirão.

Analisaremos esse problema de ordenação através de dois métodos distintos.

3 Método Direto

Nosso objetivo nesse método, e, de modo geral, em todos os métodos, é criar premiar cada time com uma pontuação condizente com o desempenho em partidas analisadas. Numa situação ideal, essa pontuação será influenciada pelos resultados das partidas e pela “força” dos oponentes enfrentados.

Assim, supondo que N seja a quantidade de times no torneio, n_i seja a quantidade de partidas disputadas pela i -ésima equipe, $\mathbf{f} \in \mathbb{R}^N$ seja um vetor em que \mathbf{f}_i é a força do i -ésimo time, e que $\mathbf{p} \in \mathbb{R}^N$ seja um vetor em que \mathbf{p}_i seja a pontuação dada ao i -ésimo time, podemos fazer uma modelagem inicial da forma

$$\mathbf{p}_i = \frac{1}{n_i} \sum_{j=1}^N a_{ij} \mathbf{f}_j, \quad (1)$$

onde a_{ij} é algum número não-negativo determinado pelo resultado do confronto entre as equipes i e j . Ou seja, resultados melhores somam mais para a pontuação, e oponentes mais fortes somam mais para a pontuação (a não ser que o resultado tenha representação 0). Além disso, podemos notar que o fator $\frac{1}{n_i}$ “tira a média” das partidas, impedindo que um time simplesmente jogue mais jogos para garantir mais pontos.

Esse modelo por si só é bem maleável, já que não definimos como os valores a_{ij} (ou a matriz A de entradas a_{ij} , muitas vezes chamada de *matriz de preferência*) são definidos. Para jogos em que não há empate, como basquete ou vôlei, temos como o exemplo o modelo em que $a_{ij} = 1$ se o time i ganha, e $a_{ij} = 0$ caso i perca. No caso do esporte permitir empates, como é o caso do futebol, podemos ter $a_{ij} = \frac{1}{2}$.

3.1 A intuição direta

Agora, podemos começar a solucionar esse vetor de pontuações \mathbf{p} através de uma suposição bem simples: a pontuação deve ser diretamente proporcional à força, isto é:

$$\mathbf{p} = B\mathbf{f} = \lambda\mathbf{f}, \quad (2)$$

onde B é a matriz de entradas a_{ij}/n_i , ou seja, uma matriz de preferência balanceada para a quantidade de jogos de cada time. Desta forma, o vetor de pontuação \mathbf{p} será um autovetor da matriz positiva B .

3.2 O Teorema de Perron-Frobenius

O que nos garante que o autovetor \mathbf{p} existe é o referido teorema. Vamos enunciá-lo e prová-lo:

Teorema (Perron-Frobenius). *Se A é uma matriz não-trivial e tem entradas não-negativas, então A possui um autovetor \mathbf{v} de entradas não-negativas, correspondente a um autovalor positivo λ tal que nenhum outro autovalor de A o ultrapassa em módulo. Além disso, se a matriz A é irredutível, o autovetor \mathbf{v} tem entradas estritamente positivas, é único e simples, e o autovalor correspondente é estritamente maior que outros autovalores de A em módulo.*

Demonstração. Antes de tudo, vamos fazer algumas definições:

- chamaremos de vetor (ou matriz) positivo (ou positiva) aquele que tiver apenas entradas positivas (e de vetor/matriz não-negativo aquele com entradas não-negativas).
- $\mathbf{p} > \mathbf{q}$ sempre que $\mathbf{p} - \mathbf{q}$ for vetor positivo e $\mathbf{p} \geq \mathbf{q}$ sempre que $\mathbf{p} - \mathbf{q}$ for vetor não-negativo.
- uma matriz A é dita irredutível se não existe uma matriz de permutação P (ou seja, $P^T = P^{-1}$) tal que:

$$PAP^T = \begin{bmatrix} A_{11} & A_{12} \\ \mathbf{0} & A_{22} \end{bmatrix}$$

onde A_{11} , A_{12} , $\mathbf{0}$ e A_{22} são matrizes bloco, e A_{11} e A_{22} são matrizes quadradas.

Assim, como em nossa matriz de preferência B , $b_{ij} = 0$ representa uma derrota, nossa matriz será irredutível se, e só se, não existirem conjuntos de times G e D tal que não há uma vitória de algum time de D sobre algum time de G . Como corolário, não podemos ter um time apenas com derrotas caso queiramos manter a matriz de preferência irredutível.

Começemos provando um truque muito útil: se A é matriz $N \times N$ não-negativa e irredutível, então:

$$(I + A)^{N-1} > \mathbf{0}.$$

Para provar, considere o vetor qualquer $\mathbf{y} \geq \mathbf{0}$ e defina \mathbf{z} como $(I + A)\mathbf{y} = \mathbf{y} + A\mathbf{y}$. Como $A \geq \mathbf{0}$, $A\mathbf{y} \geq \mathbf{0}$ e, assim, \mathbf{z} tem, ao menos, tantas entradas não-nulas quanto \mathbf{y} . Provaremos que, se \mathbf{y} já não for positivo, então certamente \mathbf{z} tem mais entradas não-nulas que \mathbf{y} .

Fazendo uma permutação consistente das entradas (tanto nos vetores, quanto nas colunas e linhas da matriz coordenadamente), podemos rescrever:

$$\mathbf{y} = \begin{bmatrix} \mathbf{u} \\ \mathbf{0} \end{bmatrix}$$

Onde $\mathbf{u} \in \mathbb{R}^k$, com $k < N$, e sendo $\mathbf{u} > 0$. Supondo que \mathbf{z} não tem mais entradas não-nulas que \mathbf{y} , temos $\mathbf{z} = \begin{bmatrix} \mathbf{v} \\ \mathbf{0} \end{bmatrix}$, com $\mathbf{v} \in \mathbb{R}^k$. Além disso, temos:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

sendo A_{11} uma matriz $k \times k$ e A_{22} é $(N - k) \times (N - k)$. Assim, $\mathbf{z} = \mathbf{y} + A\mathbf{y}$ é igual a $\begin{bmatrix} \mathbf{u} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} A_{11}\mathbf{u} + A_{12}\mathbf{0} \\ A_{21}\mathbf{u} + A_{22}\mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{u} + A_{11}\mathbf{u} \\ A_{21}\mathbf{u} \end{bmatrix}$, o que implica que $A_{21}\mathbf{u} = \mathbf{0}$. Como $\mathbf{u} > 0$, temos que $A_{21}\mathbf{u} = \mathbf{0} \iff A_{21} = \mathbf{0}$, o que é absurdo, pela definição de matrizes irredutíveis.

Logo, $\mathbf{z} = (I + A)\mathbf{y}$ tem mais entradas não-nulas que \mathbf{y} . Indutivamente, temos que $(I + A)^{N-1}\mathbf{y} > 0$ para todo $\mathbf{y} \geq 0$. Variando \mathbf{y} sobre $\{\mathbf{e}_1, \dots, \mathbf{e}_N\}$, temos que $(I + A)^{N-1} > 0$.

◆

Agora, considere A como uma matriz $N \times N$ não-negativa irredutível e seja $r(\mathbf{x})$ uma função, definida sobre os vetores não-nulos $\mathbf{x} \geq 0$ e de N entradas, definida como:

$$r(\mathbf{x}) = \min_{\substack{1 \leq i \leq N \\ x_i \neq 0}} \frac{(A\mathbf{x})_i}{x_i}.$$

Note que $x_j r(\mathbf{x}) \leq (A\mathbf{x})_j$ para todo $j = 1, 2, \dots, N$, com igualdade para ao menos um desses j . Assim, $r(\mathbf{x})\mathbf{x} \leq A\mathbf{x}$ e, para todo $k > r(\mathbf{x})$, $k\mathbf{x} \not\leq A\mathbf{x}$ (ou seja, $r(\mathbf{x})$ é o maior λ que torna $A\mathbf{x} - \lambda\mathbf{x}$ não-negativo).

Seja \mathcal{D} o domínio de r , ou seja, o conjunto dos vetores de N entradas, não-negativos e diferentes da identidade aditiva. Vamos definir o número M da seguinte forma:

$$M = \sup_{\mathbf{x} \in \mathcal{D}} r(\mathbf{x}).$$

Como $r(\alpha\mathbf{x}) = r(\mathbf{x})$ para todo $\alpha \in \mathbb{R}_{>0}$, temos que, sendo \mathcal{S} o subconjunto de \mathcal{D} com os elementos de \mathcal{D} com norma euclidiana igual a 1,

$$M = \sup_{\mathbf{x} \in \mathcal{S}} r(\mathbf{x}).$$

Perceba que, caso $r(\mathbf{x})$ fosse uma função contínua sobre \mathcal{S} , teríamos que $M = \max_{\mathbf{x} \in \mathcal{S}} r(\mathbf{x})$. Porém, não temos essa garantia, pois r pode ter discontinuidades quando as entradas de \mathbf{x} se tornam 0.

Assim, vamos nos ater ao conjunto \mathcal{U} dos vetores \mathbf{y} definidos por $\mathbf{y} = (I + A)^{N-1}\mathbf{x} \forall \mathbf{x} \in \mathcal{S}$.

Pelo truque já demonstrado, todos os elementos de \mathcal{U} serão vetores positivos, assim, \mathcal{U} é um subconjunto fechado de \mathcal{D} .

Agora, tome um \mathbf{y} qualquer de \mathcal{U} . Então, existe $\mathbf{x} \in \mathcal{S}$ de tal forma que

$$\mathbf{y}r(\mathbf{x}) = (I + A)^{N-1}\mathbf{x}r(\mathbf{x}) \leq (I + A)^{N-1}A\mathbf{x}$$

já que $r(\mathbf{x})\mathbf{x} \leq A\mathbf{x}$. Como $(I + A)^{N-1} = \sum_{k=0}^{N-1} \binom{N-1}{k} A^k$ (uma soma de potências de A), temos que $(I + A)^{N-1}$ comuta com A , o que implica que

$$\mathbf{y}r(\mathbf{x}) \leq (I + A)^{N-1}A\mathbf{x} = A(I + A)^{N-1}\mathbf{x} = A\mathbf{y}$$

Assim, temos que $r(\mathbf{x})\mathbf{y} \leq A\mathbf{y}$ para todos os pares $(\mathbf{x}, \mathbf{y}) \in \mathcal{S} \times \mathcal{U}$.

Como $r(\mathbf{y})$ é o maior número λ tal que $\lambda\mathbf{y} \leq A\mathbf{y}$, temos que $r(\mathbf{x}) \leq r(\mathbf{y})$. Desta forma,

$$\begin{aligned} M = \sup_{\mathbf{x} \in \mathcal{S}} r(\mathbf{x}) &\leq \sup_{\mathbf{y} \in \mathcal{U}} r(\mathbf{y}) \\ &= \max_{\mathbf{y} \in \mathcal{U}} r(\mathbf{y}) \text{ (já que } \mathcal{U} \text{ é fechado)} \end{aligned}$$

Porém, como $\mathcal{U} \subset \mathcal{S}$,

$$\max_{\mathbf{y} \in \mathcal{U}} r(\mathbf{y}) \leq \sup_{\mathbf{x} \in \mathcal{S}} r(\mathbf{x}) = \sup_{\mathbf{x} \in \mathcal{D}} r(\mathbf{x})$$

Logo, $M = \max_{\mathbf{y} \in \mathcal{U}} r(\mathbf{y})$, e há um vetor positivo \mathbf{v} tal que $M = r(\mathbf{v})$. Mostraremos que M é autovalor de A e que o vetor \mathbf{v} é autovetor de A associado a M .

Sem perda de generalidade, suponhamos que $\mathbf{v} \in \mathcal{S}$. Assim, seja $\mathbf{w} = (I + A)^{N-1}\mathbf{v} \Rightarrow \mathbf{w} \in \mathcal{U}$. Já sabemos que $A\mathbf{v} - M\mathbf{v} \geq 0$. Supondo que não tenhamos autovetor e autovalor nessa situação, temos que $A\mathbf{v} - M\mathbf{v} \neq 0$. Assim,

$$\begin{aligned}
& (I + A)^{N-1}(A\mathbf{v} - M\mathbf{v}) > 0 \\
\Rightarrow & (I + A)^{N-1}A\mathbf{v} - M(I + A)^{N-1}\mathbf{v} > 0 \\
\Rightarrow & A(I + A)^{N-1}\mathbf{v} - M\mathbf{w} > 0 \\
\Rightarrow & A\mathbf{w} - M\mathbf{w} > 0
\end{aligned}$$

Daí, $A\mathbf{w} > M\mathbf{w}$, o que implica que $M < r(\mathbf{w})$, um absurdo pela maximalidade de M . Assim, M e \mathbf{v} são autovalor e autovetor de A .

Por fim, provemos que M é o raio espectral de A (ou seja, qualquer outro autovalor tem módulo menor ou igual a M). Supondo α como autovalor e \mathbf{z} como autovetor de A , temos, pelo fato de $\mathbf{z} \neq 0$ e $A \geq 0$, que

$$\begin{aligned}
& \alpha\mathbf{z} = A\mathbf{z} \\
\Rightarrow & |\alpha||\mathbf{z}| = |A\mathbf{z}| \leq A|\mathbf{z}|,
\end{aligned}$$

onde, se \mathbf{w} é um vetor, $\|\mathbf{w}\|$ é o vetor obtido ao substituir cada entrada de \mathbf{w} pelo módulo dessa entrada. Assim, temos $|\alpha||\mathbf{z}| \leq A|\mathbf{z}| \Rightarrow |\alpha| \leq r(|\mathbf{z}|) \leq M$, como queríamos demonstrar.

□

Note que a demonstração se retém às matrizes não-negativas irredutíveis. No caso de redutibilidade, a dependência contínua entre os autovalores e as entradas da matriz permitem aproximar uma matriz não-negativa redutível por matrizes positivas. Assim, nesse caso mais geral, só não teremos garantia que nosso autovalor definido é o único de módulo igual ao raio espectral de A , e nem se o autovetor correspondente será positivo (mas com certeza será não-negativo).

3.3 Aplicando o método direto num exemplo real

Agora que já temos a base teórica necessária para criar nosso ranking, podemos prosseguir aplicando num exemplo prático. Usaremos o Brasileirão 2021 em seu estado atual (27 de outubro de 2022).

Para efeito demonstrativo, aqui está a tabela de pontuação no momento:

Pos.	Equipes	P	J	V	E	D	GP	GC	SG
1	Palmeiras	74	34	21	11	2	59	22	+37
2	Internacional	64	34	17	13	4	52	30	+22
3	Flamengo	61	34	18	7	9	56	32	+24
4	Fluminense	58	34	17	7	10	55	40	+15
5	Corinthians	57	33	16	9	8	39	32	+7
6	Athletico Paranaense	51	34	14	9	11	41	43	-2
7	Atlético Mineiro	48	33	12	12	9	38	33	+5
8	Botafogo	47	34	13	8	13	36	38	-2
9	São Paulo	47	33	11	14	8	46	35	+11
10	América Mineiro	46	34	13	7	14	33	36	-3
11	Fortaleza	45	33	12	9	12	34	33	+1
12	Santos	43	34	11	10	13	40	33	+7
13	Goiás	42	33	10	12	11	37	43	-6
14	Red Bull Bragantino	41	34	10	11	13	46	47	-1
15	Coritiba	35	33	10	5	18	33	53	-20
16	Ceará	34	34	6	16	12	30	36	-6
17	Atlético Goianiense	33	33	8	9	16	33	49	-16
18	Cuiabá	31	33	7	10	16	25	37	-12
19	Avaí	28	33	7	7	19	28	55	-27
20	Juventude	21	33	3	12	18	26	60	-34

Note que a ordenação é feita pela pontuação (P), que é determinada a partir da fórmula $P = 3V + E$, onde V é a quantidade de vitórias do time, e E a quantidade de empates (ou seja, nesse modelo, três empates equivalem uma vitória). Note também que a quantidade de jogos em que cada time participou (J) varia, já que, apesar do campeonato ser completo, cada time tem seu calendário que pode contemplar outras competições.

Para criar nossa matriz de preferência, usaremos uma tabela que mostra os resultados de cada um dos jogos anteriores:

	AMM	ATP	ATG	ATM	AVA	BOT	CEA	COR	CTB	CUI	FLA	FLU	FOR	GOI	INT	JUV	PAL	RBB	SAN	SPA
América-MG	—	1-0	R-38	1-1	3-1	1-1	0-2	1-0	2-0	2-1	1-2	0-0	1-2	1-0	R-35	4-1	0-1	0-3	1-0	1-2
Athletico-PR	1-1	—	4-1	0-1	2-1	R-38	1-0	1-1	1-0	2-2	1-0	1-0	1-1	R-35	0-0	2-0	1-3	4-2	2-2	1-0
Atlético-GO	0-1	R-37	—	0-2	2-1	1-1	1-0	0-1	2-0	1-1	1-1	3-2	0-1	0-1	1-2	3-1	1-1	2-1	R-35	1-2
Atlético-MG	1-2	2-3	2-0	—	2-1	R-36	0-0	1-2	2-2	R-37	2-0	2-0	3-2	0-1	2-0	R-34	0-1	1-1	1-1	0-0
Avaí	1-0	1-1	1-2	1-0	—	1-2	R-37	1-1	2-1	1-2	1-2	0-3	3-2	3-2	0-1	1-2	2-2	R-35	1-0	1-1
Botafogo	0-0	2-0	0-0	0-1	0-1	—	1-1	1-3	2-0	R-35	0-1	0-1	3-1	1-2	0-1	1-1	1-3	2-1	R-37	1-0
Ceará	1-2	0-0	1-1	0-0	1-0	1-3	—	3-1	1-1	1-1	2-2	R-35	0-1	1-1	1-1	R-38	1-2	0-1	2-1	0-2
Corinthians	1-1	2-1	2-1	R-38	3-0	1-0	R-36	—	3-1	2-0	1-0	0-2	1-0	1-0	2-2	2-0	0-1	1-0	0-0	1-1
Coritiba	1-0	0-1	2-0	0-1	1-0	1-0	1-0	R-37	—	1-0	R-36	3-2	2-1	3-0	1-1	2-2	0-2	2-1	1-2	1-1
Cuiabá	2-1	0-1	1-1	1-1	R-34	2-0	0-0	1-0	R-38	—	1-2	0-1	0-1	1-2	1-1	1-0	R-36	1-1	0-0	1-1
Flamengo	3-0	5-0	4-1	1-0	R-38	0-1	1-1	R-35	2-0	2-0	—	1-2	1-2	1-0	0-0	4-0	0-0	4-1	3-2	3-1
Fluminense	0-2	2-1	0-2	5-3	2-0	2-2	2-1	4-0	5-2	1-0	1-2	—	2-1	R-37	0-1	4-0	1-1	2-1	0-0	R-36
Fortaleza	1-0	0-0	R-36	0-0	2-0	1-3	0-1	1-0	R-34	0-1	3-2	0-1	—	1-1	3-0	1-1	0-0	R-37	0-0	1-1
Goiás	2-2	2-1	2-1	2-2	1-1	0-1	1-1	R-32	1-0	1-0	1-1	2-3	0-1	—	1-2	R-36	1-1	1-1	1-0	R-38
Internacional	1-0	R-36	1-1	3-0	0-0	2-3	2-1	2-2	3-0	1-0	3-1	3-0	2-1	4-2	—	4-0	R-38	0-0	1-0	3-3
Juventude	0-1	1-3	1-1	1-2	1-1	2-2	1-0	2-2	R-35	0-1	R-37	1-0	1-1	0-0	1-1	—	0-3	2-2	1-2	1-2
Palmeiras	R-37	0-2	4-2	0-0	3-0	4-0	2-3	3-0	4-0	1-0	1-1	1-1	R-35	3-0	2-1	2-1	—	2-0	1-0	0-0
RB Bragantino	R-36	4-2	4-0	1-1	4-0	0-1	1-1	0-1	4-2	2-1	1-0	R-38	2-1	1-1	0-2	1-0	2-2	—	0-2	1-1
Santos	3-0	2-0	1-0	1-2	R-36	2-0	0-0	0-1	2-1	4-1	1-2	2-2	R-38	1-2	1-1	4-1	0-1	2-2	—	1-0
São Paulo	1-0	4-0	R-34	R-35	4-0	0-1	2-2	1-1	3-1	2-1	0-2	2-2	0-1	3-3	R-37	0-0	1-2	3-0	2-1	—

Fazendo o processo do método direto premiando vitórias com 1 e empates com $\frac{1}{2}$, temos o seguinte ranking, apresentado em comparação com o ranking oficial:

Equipes	Método Direto	Oficial
Palmeiras	1	1
Internacional	2	2
Flamengo	3	3
Fluminense	4	4
Corinthians	5	5
Atlético Mineiro	6	7
Athletico Paranaense	7	6
Fortaleza	8	11
São Paulo	9	9
Botafogo	10	8
América Mineiro	11	10
Goiás	12	13
Santos	13	12
Ceará	14	16
Red Bull Bragantino	15	14
Coritiba	16	15
Atlético Goianiense	17	17
Cuiabá	18	18
Avaí	19	19
Juventude	20	20

É visível que as mudanças foram pontuais, já que o campeonato está quase completo. Entretanto, podemos estender esse ranqueamento usando dados mais antigos, como todos os jogos da primeira divisão do Campeonato Brasileiro desde 2003:

Ranking (Método Direto)	Pos.	Equipe	P	J	V	E	D	GP	GC	SG	Aproveitamento (%)
1912	1.	São Paulo	1264	775	349	217	209	1114	834	280	54.36
1861	2.	Palmeiras	1100	692	307	179	206	1006	802	204	52.98
1841	3.	Corinthians	1149	737	312	213	212	947	780	167	51.96
1837	4.	Flamengo	1216	776	335	211	230	1109	913	196	52.23
1835	5.	Internacional	1155	738	321	192	225	989	806	183	52.16
1806	6.	Santos	1190	776	328	206	242	1135	910	225	51.11
1795	7.	Cruzeiro	1030	666	293	151	222	978	818	160	51.55
1790	8.	Grêmio	1065	700	295	180	225	956	778	178	50.71
1760	9.	Atlético-MG	1104	736	305	189	242	1072	942	130	50.0
1702	10.	Fluminense	1111	776	302	205	269	1039	984	55	47.72
1683	11.	Bragantino	150	110	37	39	34	151	133	18	45.45
1676	12.	Athletico-PR	1053	738	293	174	271	974	917	57	47.56
1613	13.	São Caetano	236	172	65	41	66	209	199	10	45.73
1584	14.	Goiás	727	547	199	130	218	750	776	-26	44.3
1582	15.	Barueri	49	38	12	13	13	59	52	7	42.98
1578	16.	Botafogo-RJ	861	654	227	180	247	813	839	-26	43.88
1566	17.	Vasco	765	590	196	177	217	747	832	-85	43.22
1523	18.	Figueirense	550	438	142	124	172	530	622	-92	41.85
1503	19.	Coritiba	681	547	179	144	224	646	707	-61	41.49
1492	20.	Fortaleza	339	273	90	69	114	317	371	-54	41.39
1468	21.	Ceará	305	262	71	92	99	273	312	-39	38.8
1466	22.	Paraná	304	248	83	55	110	312	369	-57	40.86
1452	23.	Portuguesa	131	114	31	38	45	137	157	-20	38.3
1445	24.	Sport	500	418	131	107	180	463	568	-105	39.87
1440	25.	Bahia	458	388	115	113	160	430	512	-82	39.34
1439	26.	Ponte Preta	433	362	114	91	157	414	534	-120	39.87
1427	27.	Juventude	330	281	85	75	121	330	431	-101	39.14
1416	28.	Vitória	468	396	123	99	174	492	581	-89	39.39
1389	29.	Atlético-GO	292	261	72	76	113	282	355	-73	37.29
1388	30.	Guarani	147	130	36	39	55	140	180	-40	37.69
1380	31.	Cuiabá	78	71	17	27	27	59	74	-15	36.61
1374	32.	Paysandu	154	134	41	31	62	193	245	-52	38.3
1356	33.	Chapecoense	287	265	70	77	118	261	359	-98	36.1
1354	34.	América-MG	204	186	51	51	84	178	247	-69	36.55
1338	35.	Criciúma	188	168	50	38	80	194	264	-70	37.3
1283	36.	Brasiliense	42	42	10	12	20	49	68	-19	33.33
1283	37.	Santo André	41	38	11	8	19	46	61	-15	35.96
1264	38.	Náutico	200	190	54	38	98	224	318	-94	35.08
1261	39.	Avai	264	261	64	72	125	268	410	-142	33.71
1111	40.	Ipatinga	35	38	9	8	21	37	67	-30	30.7
1055	41.	CSA	32	38	8	8	22	24	58	-34	28.07
1044	42.	Grêmio Prudente	31	38	7	10	21	39	64	-25	27.19
1019	43.	Joinville	31	38	7	10	21	26	48	-22	27.19
964	44.	Santa Cruz	59	76	15	14	47	86	145	-59	25.87
555	45.	América-RN	17	38	4	5	29	24	80	-56	14.91

Onde é mais claro como o ranqueamento opera: não há favorecimento para times com mais jogos, já que normalizamos essa condição, e o aproveitamento de pontos dos times (de fórmula $(3V + E)/(3J)$) não se traduz na ordenação dos times, muito em razão pela matriz de preferência premiar uma vitória com o equivalente de dois empates, enquanto os pontos seguem a lógica de que uma vitória equivale a três empates (uma estratégia do campeonato de incentivar a competitividade).

Entretanto, há como melhorar nossa matriz de preferência.

3.4 Especificando a Preferência

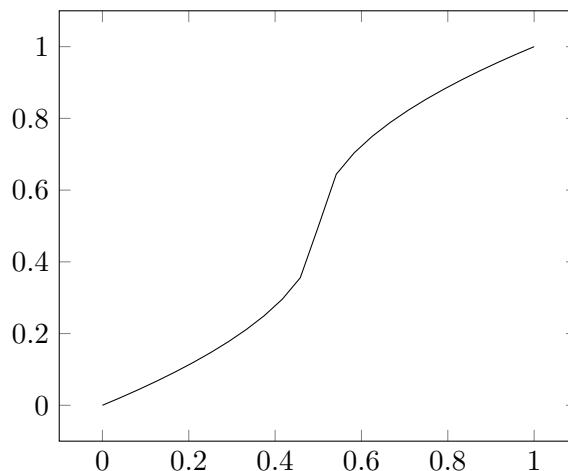
A forma como definimos nossa matriz de preferência tem alguns problemas: quando nos restringimos a uma temporada (ou uma parte dela), contamos com poucos jogos entre pares de times, o que acarreta em poucas informações sobre seus confrontos. Além disso, as vitórias passam a ser “genéricas”, isto é, uma vitória por 1 a 0 vale o mesmo que uma vitória por 7 a 1, por exemplo. Um outro problema técnico enfrentado é a possibilidade de haver um time sem vitórias ou empates. Nesse caso, a matriz de preferência fica irredutível, e nossos cálculos podem se tornar mais complicados.

Dessa forma, imaginemos que numa partida entre os times i e j , i marca G_{ij} gols e j marca G_{ji} gols. Uma boa estratégia de preferência é premiar o time i com $\frac{G_{ij}+1}{G_{ij}+G_{ji}+2}$, e o time j , com $\frac{G_{ji}+1}{G_{ij}+G_{ji}+2}$. Desta forma, cada partida premia um total de $\frac{G_{ij}+1+G_{ji}+1}{G_{ij}+G_{ji}+2} = 1$, e cada time ganha uma quantidade não-nula de pontos. Note, porém, que a pontuação que o time i ganha cresce conforme mais gols o time marca. Assim, para tornar os goleadas “menos decisivas”, impedindo que os times se aproveitem indevidamente de times mais fracos, podemos utilizar uma função não-linear como

$$h(x) = \frac{1}{2} + \frac{1}{2} \operatorname{sgn}\left(x - \frac{1}{2}\right) \sqrt{|2x - 1|}$$

onde $\operatorname{sgn}(x)$ é a função sinal (1 para números positivos, 0 para 0, e -1 para números negativos).

Essa função é interessante para nosso caso pois $h(x) > x$ para $x > \frac{1}{2}$ e $h(x) < x$ para $x < \frac{1}{2}$, $h\left(\frac{1}{2}\right) = \frac{1}{2}$, $h\left(\frac{1}{2} + x\right) + h\left(\frac{1}{2} - x\right) = 1$, e $h(x)$ tem uma taxa de variação baixa para x próximo de 0 ou 1, e alta pra x perto de $\frac{1}{2}$. Abaixo, um gráfico de $y = h(x)$.



E, assim, fazamos um ranking melhorado dos times da Série A desde 2003, usando

$$a_{ij} = h \left(\frac{G_{ij} + 1}{G_{ij} + G_{ji} + 2} \right)$$

como entradas da matriz de preferência:

Ranking (Método Direto Melhorado)	Pos.	Equipe	P	J	V	E	D	GP	GC	SG	Aproveitamento (%)
1743	1.	São Paulo	1264	775	349	217	209	1114	834	280	54.36
1711	2.	Palmeiras	1100	692	307	179	206	1006	802	204	52.98
1702	3.	Internacional	1155	738	321	192	225	989	806	183	52.16
1702	4.	Flamengo	1216	776	335	211	230	1109	913	196	52.23
1696	5.	Corinthians	1149	737	312	213	212	947	780	167	51.96
1695	6.	Santos	1190	776	328	206	242	1135	910	225	51.11
1688	7.	Grêmio	1065	700	295	180	225	956	778	178	50.71
1674	8.	Cruzeiro	1030	666	293	151	222	978	818	160	51.55
1657	9.	Atlético-MG	1104	736	305	189	242	1072	942	130	50.0
1625	10.	Bragantino	150	110	37	39	34	151	133	18	45.45
1616	11.	Fluminense	1111	776	302	205	269	1039	984	55	47.72
1607	12.	Athletico-PR	1053	738	293	174	271	974	917	57	47.56
1606	13.	Barueri	49	38	12	13	13	59	52	7	42.98
1579	14.	São Caetano	236	172	65	41	66	209	199	10	45.73
1552	15.	Goiás	727	547	199	130	218	750	776	-26	44.3
1546	16.	Botafogo-RJ	861	654	227	180	247	813	839	-26	43.88
1541	17.	Vasco	765	590	196	177	217	747	832	-85	43.22
1515	18.	Coritiba	681	547	179	144	224	646	707	-61	41.49
1502	19.	Figueirense	550	438	142	124	172	530	622	-92	41.85
1497	20.	Ceará	305	262	71	92	99	273	312	-39	38.8
1491	21.	Fortaleza	339	273	90	69	114	317	371	-54	41.39
1491	22.	Portuguesa	131	114	31	38	45	137	157	-20	38.3
1480	23.	Bahia	458	388	115	113	160	430	512	-82	39.34
1476	24.	Paraná	304	248	83	55	110	312	369	-57	40.86
1465	25.	Sport	500	418	131	107	180	463	568	-105	39.87
1454	26.	Juventude	330	281	85	75	121	330	431	-101	39.14
1452	27.	Ponte Preta	433	362	114	91	157	414	534	-120	39.87
1450	28.	Vitória	468	396	123	99	174	492	581	-89	39.39
1447	29.	Cuiabá	78	71	17	27	27	59	74	-15	36.61
1435	30.	Guarani	147	130	36	39	55	140	180	-40	37.69
1433	31.	Atlético-GO	292	261	72	76	113	282	355	-73	37.29
1418	32.	Paysandu	154	134	41	31	62	193	245	-52	38.3
1407	33.	Chapecoense	287	265	70	77	118	261	359	-98	36.1
1406	34.	Criciúma	188	168	50	38	80	194	264	-70	37.3
1406	35.	América-MG	204	186	51	51	84	178	247	-69	36.55
1396	36.	Santo André	41	38	11	8	19	46	61	-15	35.96
1371	37.	Brasiliense	42	42	10	12	20	49	68	-19	33.33
1356	38.	Náutico	200	190	54	38	98	224	318	-94	35.08
1340	39.	Avai	264	261	64	72	125	268	410	-142	33.71
1246	40.	Ipatinga	35	38	9	8	21	37	67	-30	30.7
1238	41.	Grêmio Prudente	31	38	7	10	21	39	64	-25	27.19
1226	42.	Joinville	31	38	7	10	21	26	48	-22	27.19
1209	43.	CSA	32	38	8	8	22	24	58	-34	28.07
1198	44.	Santa Cruz	59	76	15	14	47	86	145	-59	25.87
884	45.	América-RN	17	38	4	5	29	24	80	-56	14.91

4 Método de Probabilidades

Uma outra pergunta que surge frequentemente em círculos de debate esportivo é sobre quem tem mais probabilidade de ganhar um confronto direto. Note que estamos descartando a possibilidade de empate. Essa suposição é especialmente útil para competições como a Copa do Mundo, em que dada fase da competição é mata-mata. Para tal fim, podemos usar uma visão probabilística em nossa matriz de preferência. Assim, supondo que \mathbf{p} seja o

vetor pontuação dos times, e π_{ij} seja a probabilidade do time i ganhar do time j , digamos que

$$\pi_{ij} = \frac{p_i}{p_i + p_j}.$$

E, como $\pi_{ij} + \pi_{ji} = 1$, temos que

$$\pi_{ji}p_i - \pi_{ij}p_j = 0.$$

É notável que, sabendo π_{ij} , poderíamos determinar p_i . Como não temos uma definição exata das probabilidades, podemos estimá-las com P_{ij} definidos como

$$P_{ij} = \frac{S_{ij}}{S_{ij} + S_{ji}},$$

sendo S_{ij} alguma variável que represente o histórico do time i sobre o time j . Assim,

$$S_{ji}p_i - S_{ij}p_j = 0.$$

Tomando um histórico extenso o bastante, vemos que teremos mais equações do que variáveis. Assim, como não é garantia que teremos uma solução única que fará todas as equações serem satisfeitas, podemos ao menos achar a solução com menor “distância” às equações.

4.1 Achando uma Boa Solução

A ideia de achar uma solução que esteja “próxima” de equações lembra a ideia dos mínimos quadrados, isto é, considerar os quadrados das equações, e minimizar a soma, já que assim, estamos tornando cada equações próxima de zero.

Entretanto, como o vetor nulo $\mathbf{p} = (0, \dots, 0)$ satisfaz todas as equações, devemos alguma restrição em nosso cálculo, como, por exemplo, que o vetor \mathbf{p} tenha norma euclidiana 1. Assim, expressando essa restrição através de um multiplicador de Lagrange μ , temos que nossa função a ser minimizada será:

$$\left(\sum_{ij} (S_{ji}p_i - S_{ij}p_j)^2 \right) - \mu \left(\left(\sum_i p_i^2 \right) - 1 \right).$$

Para achar o mínimo, diferenciamos a função anterior em relação a p_i , e a igualamos a zero, donde temos:

$$\begin{aligned} & \left(\sum_{\substack{1 \leq j \leq N \\ j \neq i}} 2(S_{ji}p_i - S_{ij}p_j)S_{ji} \right) - 2\mu p_i = 0 \\ & \iff \sum_{\substack{1 \leq j \leq N \\ j \neq i}} (S_{ji}^2 p_i - S_{ij}S_{ji}p_j) = \mu p_i \\ & \iff B\mathbf{p} = \mu\mathbf{p}, \end{aligned}$$

onde B é uma matriz $N \times N$ definida como $b_{ij} = \begin{cases} \sum_{k \neq i} S_{ki}^2, & \text{se } i = j \\ -S_{ij}S_{ji}, & \text{se } i \neq j \end{cases}$. Ou seja, B é uma matriz tal que seus elementos na diagonal são positivos, e os elementos fora da diagonal são não-positivos. Provaremos que \mathbf{p} é um autovetor positivo único em uma matriz relacionada a B .

Primeiro, note que, devido ao grande número de jogos esperados, é bem provável que as colunas de B sejam linearmente independentes, ou seja, B é invertível. Além disso, existe um valor real positivo λ tal que a matriz $B' = B + \lambda I$ é diagonalmente dominante (isto é, para todo $i = 1, \dots, N$, temos que $|b'_{ii}| > \sum_{j \neq i} |b'_{ij}|$). Assim, para o vetor $\mathbf{v}_0 = (1, \dots, 1)$, $B'\mathbf{v}_0 > 0$. Agora, analisemos $B'\mathbf{x}$ para $\mathbf{x} \geq 0$, com $\mathbf{x} \not\geq 0$. Sendo \mathbf{v}_j um vetor de entradas positivas, com exceção da j -ésima entrada, nula, então a j -ésima entrada de $B'\mathbf{v}_j$ será negativa ou nula. Supondo que tenhamos tido jogos o suficiente para que $(B'\mathbf{v}_j)_j$ seja estritamente negativo, então $B'\mathbf{x} \not\geq 0$ para \mathbf{x} tal que $\mathbf{x} \geq 0$ e $\mathbf{x} \not\geq 0$. Como existe um vetor, \mathbf{v}_0 , que é mapeado por B' do ortante positivo para o ortante positivo, então B' é uma transformação linear que leva o ortante positivo a um conjunto que contém o ortante positivo. Assim, B'^{-1} é uma transformação linear que leva o ortante positivo ao ortante positivo, o que mostra que $B'^{-1} \geq 0$.

Assim, pelo Teorema de Perron-Frobenius, B'^{-1} possui um autovetor positivo \mathbf{p} correspondente a um autovalor μ , maior ou igual em módulo a todos os outros autovalores de B'^{-1} . Assim, \mathbf{p} é autovetor de B' , correspondente ao autovalor de menor módulo dentre os autovalores de B' .

4.2 Prevendo a Copa do Mundo

Agora, botemos em prática nossa teoria.

Usando $S_{ij} = \sum h\left(\frac{G_{ij}+1}{G_{ij}+G_{ji}+2}\right)$, isto é, somatório das pontuações estabelecidas no método anterior, temos o seguinte ranking de seleções para a Copa do Mundo 2022.

Ranking (Método Probabilístico)	Equipe
1.	Brasil
2.	Argentina
3.	Senegal
4.	Marrocos
5.	Equador
6.	Uruguai
7.	Japão
8.	Tunísia
9.	Estados Unidos
10.	México
11.	Canadá
12.	Coreia do Sul
13.	Arábia Saudita
14.	Irã
15.	Gana
16.	Camarões
17.	Costa Rica
18.	Holanda
19.	Austrália
20.	Catar
21.	Espanha
22.	Inglaterra
23.	Alemanha
24.	Sérvia
25.	França
26.	Portugal
27.	Bélgica
28.	Suíça
29.	Croácia
30.	Dinamarca
31.	Polônia
32.	Gales

5 Conclusão

Podemos concluir que o Teorema de Perron-Frobenius é de grande valia, podendo inclusive ser peça-chave dos mais diversos e eficientes métodos de ranqueamento esportivo. Apesar de ser uma técnica da Álgebra Linear, pudemos usá-lo de forma mais refinada para resolver um problema de mínimos quadrados, inclusive. Podemos também perceber que métodos alternativos de ranqueamento são úteis para verificar pontos fortes de times que não são valorizados nos ranqueamentos comuns.

No mais, resta-nos torcer para que o ranking gerado por Perron-Frobenius esteja correto e nos leve ao hexacampeonato.

Referências

[Keener 1993] KEENER, J. The Perron-Frobenius Theorem and the Ranking of Football Teams. **SIAM Review**. Society for Industrial and Applied Mathematics, v. 35, n.1, p. 80–93, Mar. 1993.

[Lancaster 1969] LANCASTER, P. Theory of matrices. **Academic Press**. New York.