Tennis Analytics

Júlia Stancioli Paiva

29 de julho de 2024

Tipo de pesquisa: científica

Orientador: Pedro Olmo S. Vaz de Melo

1 Resumo

Este trabalho aborda a aplicação da ciência de dados ao tênis, com foco na realização de análises estatísticas e no desenvolvimento de algoritmos para ranqueamento dos jogadores. Inicialmente, realizou-se uma investigação abrangente das bases de dados de tênis disponíveis publicamente. Utilizando técnicas de análise estatística, foram testadas diversas crenças amplamente difundidas na comunidade do tênis, com ênfase especial no fenômeno dos "upsets", em que jogadores favoritos, conforme os rankings oficiais, são derrotados por seus oponentes. Posteriormente, desenvolveu-se uma nova metodologia para a quantificação da habilidade dos jogadores, proposta como uma alternativa ao ranking oficial da ATP. Esta metodologia baseia-se no algoritmo PageRank, adotado como uma aproximação assintótica do modelo Bradley-Terry. A eficácia do ranking alternativo foi avaliada por meio da comparação da frequência de upsets por temporada, o que mostrou elevada correlação entre os resultados gerados pelo PageRank e o ranking oficial da ATP, sugerindo que a metodologia proposta pode servir como um complemento ou substituto válido para os sistemas de ranqueamento existentes.

2 Introdução

A área de Sports Analytics refere-se à aplicação de métodos estatísticos, modelagem matemática e tecnologias de informação para analisar dados relacionados a eventos esportivos, com o objetivo de extrair insights significativos para a tomada de decisões estratégicas. Essa disciplina emergente tem revolucionado a maneira como equipes, treinadores e gestores abordam o desenvolvimento tático, o recrutamento de jogadores e a otimização do desempenho atlético. Um exemplo notável de Sports Analytics pode ser encontrado no basquete, onde equipes da NBA utilizam análises avançadas para avaliar a eficácia de jogadas, identificar padrões de comportamento dos jogadores e aprimorar estratégias defensivas e ofensivas. Também no futebol americano, equipes da NFL empregam análises detalhadas para examinar o desempenho de jogadores em diversas situações de jogo, determinar a eficácia de jogadas específicas e até mesmo prever padrões de lesões. Esses exemplos destacam como a área de Sports Analytics não apenas enriquece a compreensão dos esportes, mas também oferece vantagens competitivas relevantes por meio da tomada de decisões baseadas em dados.

Nesse contexto, o tênis se torna um campo propício para a aplicação da ciência de dados, uma vez que os jogos são caracterizados por uma combinação única de estratégia, habilidade física e mental, e decisões táticas em tempo real. Apesar da tendência crescente da área de Sports Analytics, o tênis não apresenta o mesmo avanço nesse âmbito em relação a outros esportes, o que se deve principalmente à descentralização da coleta de dados, que é feita por diferentes organizações, além da reduzida disponibilização dos dados para o público e para os próprios atletas. Assim, a motivação para o desenvolvimento deste trabalho é expandir a aplicação da ciência de dados ao tênis, buscando aproximar-se da robustez analítica já atingida em outros esportes.

Sob essa ótica, a etapa inicial deste trabalho foi a exploração das bases de dados disponíveis publicamente, de modo a mapear a viabilidade para a aplicação de diferentes métodos matemáticos e estatísticos. Além disso, os dados obtidos foram analisados para a identificação de padrões no esporte, como:

- **Padrões associados à ocorrência de upsets**: investigamos a relação entre o ranking dos jogadores e a probabilidade de upsets (resultados inesperados) em diferentes rodadas de Grand Slams. Para o circuito masculino, foi confirmada a máxima de que, conforme o avanço do torneio, maior a chance de vitória dos favoritos (pela ordem do ranking).

- **Evolução da pontuação dos jogadores no topo do ranking**: analisamos a evolução da pontuação dos jogadores líderes do ranking ao longo dos anos, buscando entender o grau de competitividade entre os jogadores. Além disso, observamos um decaimento exponencial da pontuação conforme o avanço do ranking, o que mostra que a diferença entre as posições no ranking não traduz a discrepância de habilidades entre os jogadores, para diferentes faixas do ranking.

- **Duração da partida em diferentes superfícies**: comparamos a duração média das partidas em diferentes tipos de superfície (saibro, grama, dura), observando que os jogos têm se tornado mais longos nos últimos 20 anos.

- **Probabilidade de quebra de serviço**: observamos que a probabilidade de quebra de serviço aumentou ao longo dos últimos anos no circuito masculino, constatando que esse é um dos fatores que contribuem para o aumento da duração dos jogos.

Com base nos resultados anteriores, o foco da segunda etapa deste trabalho foi a busca por quantificar a habilidade de cada jogador: existe alguma maneira mais precisa para ordenar os jogadores, levando em consideração suas habilidades de forma matemática? O objetivo principal é entender a qualidade do atual sistema de rankings oficiais, principalmente quando observamos uma alta frequência de *upsets* nos circuitos (masculino e feminino). Este trabalho propõe um ranking alternativo baseado no algoritmo de PageRank, usado como uma aproximação assintoticamente eficiente para o modelo de Bradley-Terry, sendo aplicado de forma específica para cada superfície de quadra (saibro, grama e quadra dura).

Este trabalho está organizado da seguinte forma: no Capítulo 3, revisaremos a literatura relevante, mostrando a equivalência entre o PageRank e o modelo Bradley-Terry. O Capítulo 4 detalhará a metodologia utilizada para obtenção do novo score de habilidade usado para o ranking. No Capítulo 5, apresentaremos os resultados obtidos, enquanto o Capítulo **??** discutirá as conclusões, limitações do estudo e sugestões para pesquisas futuras. Este documento culmina na contribuição para a compreensão mais aprofundada do tênis por meio da aplicação da ciência de dados.

3 Referencial Teórico

3.1 Ranking

O ranking no tênis profissional é um sistema que classifica os jogadores de acordo com seu desempenho em torneios oficiais. A Associação de Tênis Profissional (ATP) e a Associação Feminina de Tênis (WTA) são responsáveis por manter os rankings masculino e feminino, respectivamente. Além de permitir que os jogadores sejam comparados entre si, o ranking também é usado para determinar quais jogadores podem participar dos principais torneios, garantindo que eles sejam disputados pelos melhores jogadores do mundo. O sistema de ranking atribui pontos aos jogadores conforme o tipo do torneio e a rodada alcançada pelo jogador.

3.1.1 Tipos de Torneios

Os torneios de tênis são divididos em categorias, que determinam a quantidade de pontos que os jogadores podem ganhar ao participar e vencer. Além disso, a quantidade de rodadas do torneio e de jogadores participantes também muda entre as categorias. As principais categorias são:

- **Grand Slams**: são os quatro torneios mais prestigiados do mundo - Australian Open, Roland-Garros, Wimbledon e US Open.

- **ATP Finals/WTA Finals**: reúnem os oito melhores jogadores/jogadoras da temporada.

EVENT	Champ	R-up	SFs	QFs	R16	R32	R64	R128
Grand Slams	2000 (=)	1300 (+100)	800 (+80)	400 (+40)	200 (+20)	100 (+10)	50 (+5)	10 (=)
Masters 1000 (96)	1000 (=)	650 (+50)	400 (+40)	200 (+20)	100 (+10)	50 (+5)	25 (+5)	10 (=)
Masters 1000 (56)	1000 (=)	650 (+50)	400 (+40)	200 (+20)	100 (+10)	50 (+5)	10 (=)	-
ATP Finals	1500 (5-0)	-	-	-	-	-	-	-
ATP 500 (48)	500 (=)	330 (+30)	200 (+20)	100 (+10)	50 (+5)	25 (+5)	-	-
ATP 500 (32)	500 (=)	330 (+30)	200 (+20)	100 (+10)	50 (+5)	-	-	-
ATP 250 (48)	250 (=)	165 (+15)	100 (+10)	50 (+5)	25 (+5)	13 (+3)	-	-
ATP 250 (32)	250 (=)	165 (+15)	100 (+10)	50 (+5)	25 (+5)	-	-	-

Figura 1: Regras de pontuação do ranking ATP 2024 [poi]

- **Masters 1000/WTA 1000**: são os nove torneios mais importantes da temporada, depois dos Grand Slams.
- **ATP 500/WTA 500**: torneios de nível intermediário.
- **ATP 250/WTA 250**: torneios de nível menor.

3.1.2 Regras de pontuação

A quantidade de pontos que um jogador ganha em um torneio depende da categoria do torneio e da fase em que ele é eliminado. O vencedor de um Grand Slam, por exemplo, recebe 2000 pontos, enquanto o vice-campeão recebe 1300 pontos. Já o vencedor de um ATP 250 recebe 250 pontos, enquanto o vice-campeão recebe 165 pontos. De maneira geral, a pontuação aumenta conforme o jogador avança no torneio, mas não de forma proporcional. As pontuações associadas a cada rodada podem ser atualizadas pela ATP e pela WTA ao longo do tempo.

3.1.3 Cálculo do Ranking

O ranking de um jogador é calculado somando os pontos que ele ganhou nas últimas 52 semanas. Os pontos mais antigos vão expirando gradualmente, o que significa que o desempenho recente de um jogador tem mais peso no ranking do que seu desempenho anterior. Além disso, o ranking é atualizado semanalmente, após a conclusão de todos os torneios da semana. Para ver mais detalhes sobre o ranking, visite [ATPd].

3.2 Modelo Bradley-Terry

O modelo Bradley-Terry é um modelo probabilístico usado para prever o resultado de comparações entre pares de elementos que se enfrentam, utilizando a estimativa de um *score* individual para cada elemento. Generalizações deste modelo foram criadas para contemplar a possibilidade de empate nas disputas. Todavia, como não ocorrem empates no tênis, estamos considerando a premissa base de que,

em toda disputa, um dos jogadores vence e o outro perde. No contexto do tênis, o modelo Bradley-Terry pode ser utilizado para estimar a probabilidade de um jogador i vencer o jogador j, em que os *scores* de cada jogador podem ser interpretados como suas respectivas habilidades. Com base nos *scores*, podemos criar um ranking contemplando todos os jogadores.

Para estimar a habilidade de cada jogador, o modelo Bradley-Terry recebe como entrada o histórico de resultados das partidas entre n jogadores. É importante ressaltar que pode não haver dados para todos os pares de jogadores (i, j) possíveis. No entanto, após a obtenção dos *scores* de cada jogador, é possível calcular o resultado esperado para uma partida que nunca ocorreu. Os dados de entrada podem ser representados por meio de uma matriz C, em que cada elemento C_{ij} corresponde à quantidade de vezes que o jogador i venceu o jogador j. Essa mesma representação também será utilizada no algoritmo PageRank ([Sel24]).

3.2.1 Formulação teórica

Existem diversas formalizações matemáticas para o modelo de Bradley-Terry. Em [JMS20], é usada a premissa de $P(i > j) = \frac{\alpha_i}{\alpha_i + \alpha_j}$, em que $i > j$ significa que o jogador i venceu o jogador j, e os valores α_i e α_j correspondem às respectivas habilidades dos jogadores. Os valores de $P(i > j)$ podem ser obtidos diretamente da matriz C de histórico de partidas, calculando $p_{ij} = \frac{C_{ij}}{C_{ij} + C_{ji}}$ (quantidade de vezes em que i venceu j, dentre todas as partidas envolvendo i e j). Para obter os parâmetros α, é utilizado o método de *maximum likelihood estimation* ([Tow]).

No caso do paper [Sel24], que foi utilizado como base para a associação entre o algoritmo PageRank e o modelo Bradley-Terry, é empregada uma formalização alternativa, equivalente à anterior. Nesta formulação, é definido $logP(i > j) = \mu_i - \mu_j$, em que μ_i e μ_j são os *scores* de habilidade dos respectivos jogadores. De forma análoga, com base nos valores de $P(i > j)$, é possível estimar os *scores* μ associados. É válido ressaltar que este modelo considera que os resultados de cada partida são independentes. Além disso, por ser um modelo estatístico, é possível obter medidas de incerteza a partir dos resultados produzidos, o que não acontece no caso do PageRank, a ser detalhado a seguir.

3.3 Page Rank

O algoritmo PageRank foi criado na década de 90 para o desenvolvimento da máquina de busca do Google, buscando ranquear as páginas da internet com base em sua relevância. A ideia que baseia o algoritmo é uma caminhada aleatória no grafo, em que os nós representam páginas web e as arestas representam *hiperlinks* que conectam tais páginas na internet. Um usuário escolhe uma página web aleatoriamente e começa a navegar na internet através dos *hiperlinks*. O tempo que o usuário gasta em cada página é interpretado como a importância daquela página na rede.

A inovação proporcionada pelo Google a este algoritmo foi a inserção do *damping factor* α, que é a probabilidade de o usuário escolher uma nova página aleatória, não conectada à página vigente. Esse fator permite que nós desconectados em um grafo esparso sejam acessados, garantindo que o processo seja *ergódico* (todos os nós da rede sejam associados a qualquer outro nó).

3.3.1 Formulação matemática

Analogamente à notação usada no algoritmo de Bradley-Terry, usaremos a matriz C para representar os resultados históricos de um torneio, em que cada entrada C_{ij} representa a quantidade de vezes que o jogador i venceu o jogador j. Em teoria de grafos, podemos interpretar esses dados como um grafo direcionado, em que cada jogador é representado por um nó e o peso da aresta direcionada e_{ji} é igual ao valor C_{ij}. Além disso, considere a matriz A como uma matriz diagonal correspondente à soma das colunas de C. A seguir, mostramos um exemplo simples das matrizes C e A para um torneio com apenas 3 jogadores.

$$C = \begin{pmatrix} 0 & 1 & 3 \\ 2 & 0 & 1 \\ 1 & 2 & 0 \end{pmatrix} \tag{1}$$

$$A = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{pmatrix} \tag{2}$$

O PageRank será a distribuição estacionária da cadeia de Markov com matriz de transição

$$P_\alpha = \alpha C A^{-1} + \frac{1-\alpha}{n} e e^T \tag{3}$$

em que n é o total de jogadores no torneio e e é um vetor n-dimensional de 1's. O PageRank pode ser computado a partir do autovetor $\pi = P\pi$.

Esta é a formalização matemática usada para comprovar a equivalência metodológica entre o algoritmo PageRank e o modelo Bradley-Terry ([Sel24]). Porém, na implementação do trabalho, o uso do PageRank ocorre de forma direta, a partir de funções previamente implementadas na biblioteca python *networkx*, usada para a manipulação de grafos.

3.4 Equivalência entre Bradley-Terry e PageRank

Conforme mostrado anteriormente, o vetor PageRank pode ser obtido computando-se o autovetor principal π da matriz P de transição (para $\alpha = 1$, $P = CA^{-1}$). O artigo [Sel24] demonstra que o vetor $d = (\mu_1, \mu_2, ..., \mu_n)$ de scores de habilidades dos jogadores, obtido no modelo Bradley-Terry, é também o autovetor principal de uma matriz derivada de C, porém após a aplicação de algumas transformações lineares. Nesse contexto, o artigo prova que os vetores π e d são autovetores correspondentes, porém em diferentes escalas. A seguir, serão mostradas as principais propriedades utilizadas na demonstração, consulte o artigo original para verificar as etapas rigorosamente.

3.4.1 Quasi-simetria

Uma matriz quadrada $Q = (q_{ij})_{nxn}$ é quasi-simétrica quando é válida a decomposição

$$Q = AXB$$

onde A e B são matrizes diagonais, e X é uma matriz simétrica ($X = X^T$). Consequentemente, a matriz Q é quasi-simétrica se, somente se, pode ser escrita na forma

$$Q = DS$$

onde D é uma matriz diagonal e S é uma matriz simétrica.

O modelo de quasi-simetria é um tipo de modelo log-linear aplicado a tabelas de contingência quadradas (como a matriz C de resultados), onde se assume que as contagens esperadas das comparações (*scores* de habilidade) são quasi-simétricas. O modelo Bradley-Terry pode ser visto como uma formulação logística específica desse modelo de quasi-simetria. Portanto, assumindo que a matriz de resultados C de um torneio é quasi-simétrica, o modelo Bradley-Terry se ajustaria perfeitamente a esses dados. Nesse caso, as habilidades dos jogadores, representadas pelos *scores*, seriam calculadas como os logaritmos dos elementos diagonais da matriz D obtida na decomposição quasi-simétrica.

3.4.2 Similaridade de Matrizes

Teorema 1 *Seja $Q = DS$ uma matriz quadrada quasi-simétrica, com D diagonal e S simétrica. Seja e um vetor de 1's. Seja $A = diag(e^T Q)$ uma matriz diagonal correspondente à soma dos elementos das colunas de Q. Seja d o vetor correspondente aos elementos diagonais da matriz D. Então d é o autovalor principal de $A^{-1}Q$.*

O teorema acima implica que, sob a premissa da quasi-simetria, o modelo de Bradley-Terry é uma versão dimensionada do PageRank. Para provar este teorema, precisamos de alguns resultados intermediários.

Lema 1 *Considere as matrizes C e A já definidas anteriormente. A matriz $A^{-1}C$ tem maior autovalor igual a 1.*

Matrizes estocásticas, como CA^{-1} têm maior autovalor igual a 1, com multiplicidade 1. Apesar da matriz $A^{-1}C$ não ser estocástica, ela é similar à matriz CA^{-1}.

Lema 2 *Duas matrizes M e M' são ditas similares quando existe uma matriz invertível X tal que $M' = X^{-1}MX$. Matrizes similares apresentam o **mesmo conjunto de autovalores**.*

Considere as matrizes C e A já definidas anteriormente. CA^{-1} e $A^{-1}C$ são claramente similares, visto que $A^{-1}(CA^{-1})A = A^{-1}C$. Logo, CA^{-1} e $A^{-1}C$ possuem os mesmos autovalores. Portanto, o autovalor principal de $A^{-1}C$ é igual a 1.

Com base nos resultados acima, conseguimos provar o **teorema 1**:

$$\begin{aligned} A^{-1}Cd &= A^{-1}(DS)(De) \\ &= A^{-1}D(e^TDS)^T \\ &= DA^{-1}Ae \\ &= De \\ &= d. \end{aligned}$$

Portanto, d é um autovetor de $A^{-1}C$, com autovalor igual a 1. Logo, d é o principal autovetor de $A^{-1}C$. Assim, conseguimos mostrar que o PageRank sem amortecimento (α) é o autovetor principal $\pi = CA^{-1}\pi$, enquanto o vetor de habilidades d de Bradley-Terry é o autovetor da matriz similar $A^{-1}C$ (não estocástica). Dessa forma, temos que $d = A^{-1}\pi$, ou seja, podemos obter duas métricas equivalentes a partir de uma única computação de autovetores.

4 Contribuição

4.1 Pré-processamento dos dados

As bases de dados usadas neste trabalho ([Sac23a] e [Sac23b]) contêm informações sobre jogadores de tênis da ATP e da WTA - respectivamente, incluindo arquivos de rankings históricos, resultados e estatísticas de partidas. Os rankings da ATP estão disponíveis principalmente de 1985 até o presente, com exceção de 1982 e com rankings intermitentes de 1973 a 1984. Os resultados estão divididos em até três arquivos por temporada, abrangendo partidas do nível principal, classificatórias e *challengers*, além de partidas do nível *future*. A descrição inclui redundância nos arquivos de resultados para facilitar o uso, contendo informações biográficas e de ranking para ambos os jogadores (vencedor e perdedor). As estatísticas de partidas estão disponíveis a partir de 1991 para partidas de nível principal, 2008 para *challengers* e 2011 para classificatórias de nível principal. As estatísticas por partida são apresentadas em totais inteiros, permitindo o cálculo de percentagens tradicionais. Alguns jogos de nível principal podem ter estatísticas ausentes devido à falta de dados da ATP ou a verificações de integridade. Além disso, as partidas da Copa Davis estão incluídas, mas as estatísticas para essas partidas estão disponíveis apenas nas últimas temporadas.

Os arquivos usados neste trabalho consistem no resultado das partidas, sendo destacados os atributos mais relevantes para a metodologia aplicada:

- Dados do torneio
  - $surface$: superfície da quadra;
  - $tourney\_date$: segunda-feira da semana de início do torneio;
- Dados da partida
  - $score$: placar final da partida;
  - $round$: rodada do torneio (semifinal, final, quartas, oitavas, etc.);
  - $minutes$: duração da partida, em minutos;
- Dados do jogador

  Como cada linha do dataset corresponde a uma partida, temos informações sobre o vencedor e o perdedor do jogo. Assim, as colunas tem o mesmo nome, modificando-se apenas o prefixo (*winner* ou *loser*). Veja a descrição abaixo:
  - $winner\_id$: identificador único do ganhador da partida (analogamente, $loser\_id$);
  - $winner\_name$: nome do ganhador da partida (analogamente, $loser\_name$);
  - $winner\_rank$: posição no ranking do ganhador (analogamente $loser\_rank$);

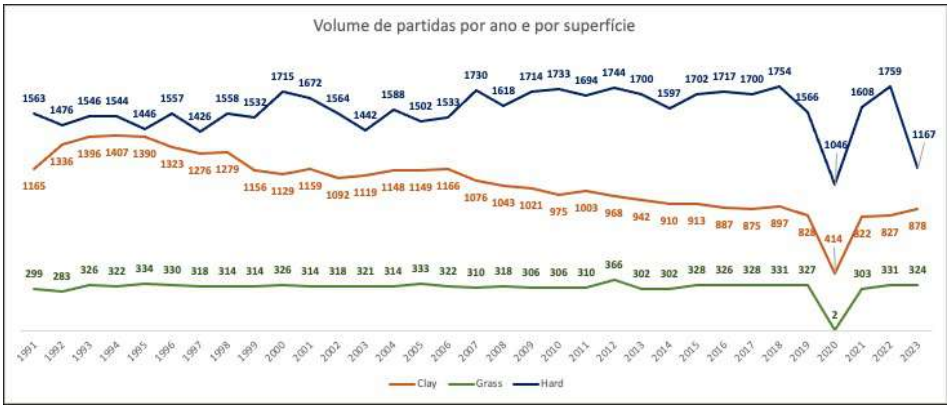


Figura 2: Quantidade de partidas ATP por ano em cada superfície

Neste trabalho, a análise do desempenho dos jogadores foi categorizada de acordo com a superfície da quadra, visto que diferentes jogadores têm estilos mais adaptados a superfícies específicas, e isso não é visualizado no atual sistema de ranking. Há alguns anos, existia outra superfície de quadra, o carpete, que foi gradualmente sendo substituído por quadras duras ou quadras de saibro. Por isso, ao observarmos o volume de partidas em cada superfície, existe uma pequena variação na quantidade total de jogos por ano. Além disso, o volume de jogos em 2020 foi muito inferior aos demais anos por conta da pandemia. Assim, nessa pesquisa, apenas os 3 tipos de superfície utilizados atualmente - quadra dura, saibro e grama - serão considerados. Ao analisar os resultados obtidos, é importante lembrar que um maior volume de partidas contribui para a obtenção de métricas mais precisas.

4.2 Aplicação do algoritmo PageRank

O objetivo de pesquisa foi, para cada temporada entre 1991 e 2023, estabelecer um ranking dos jogadores ativos para cada superfície, com base nos *scores* de habilidade obtidos via PageRank. Antes da aplicação do algoritmo, foi realizada uma etapa de análise exploratória para entendimento do panorama geral dos dados, sob a perspectiva do enfrentamento entre pares de jogadores. Para isso, foram construídos alguns grafos direcionados em que:

- Cada nó representa um jogador;
- Cada aresta direcionada (i, j) indica que o jogador j venceu o jogador i;
- O gradiente de coloração dos nós indica a taxa de vitórias do jogador naquela temporada;
- O tamanho do nó indica a quantidade de partidas jogadas na temporada;

A figura 3 representa uma ampliação desta estrutura de grafo para os jogos de quadra dura da temporada de 2022. Nessa figura, podemos visualizar rapidamente alguns jogadores com posições elevadas no ranking da época: Carlos Alcaraz, Novak Djokovic, Rafael Nadal, Daniil Medvedev, entre outros, cujos nós apresentam colorações mais escuras, associadas a uma taxa elevada de vitórias. Essa representação nos dá uma ideia sobre fatores que serão considerados no algoritmo do PageRank: não basta vencer frequentemente, vencer bons jogadores tem um peso grande no cálculo da habilidade de um jogador.

Para cada temporada, foram executadas as mesmas etapas de processamento dos dados até o cálculo dos *scores* de habilidade:

- Seleção dos jogos associados à superfície e à temporada analisada;
- Contagem da quantidade de vitórias para cada par de jogadores no dataset (construção da matriz C de resultados);
- Criação de um grafo direcionado G, em que cada aresta (i, j) representa que j venceu i, e o peso da aresta é a frequência desse resultado;

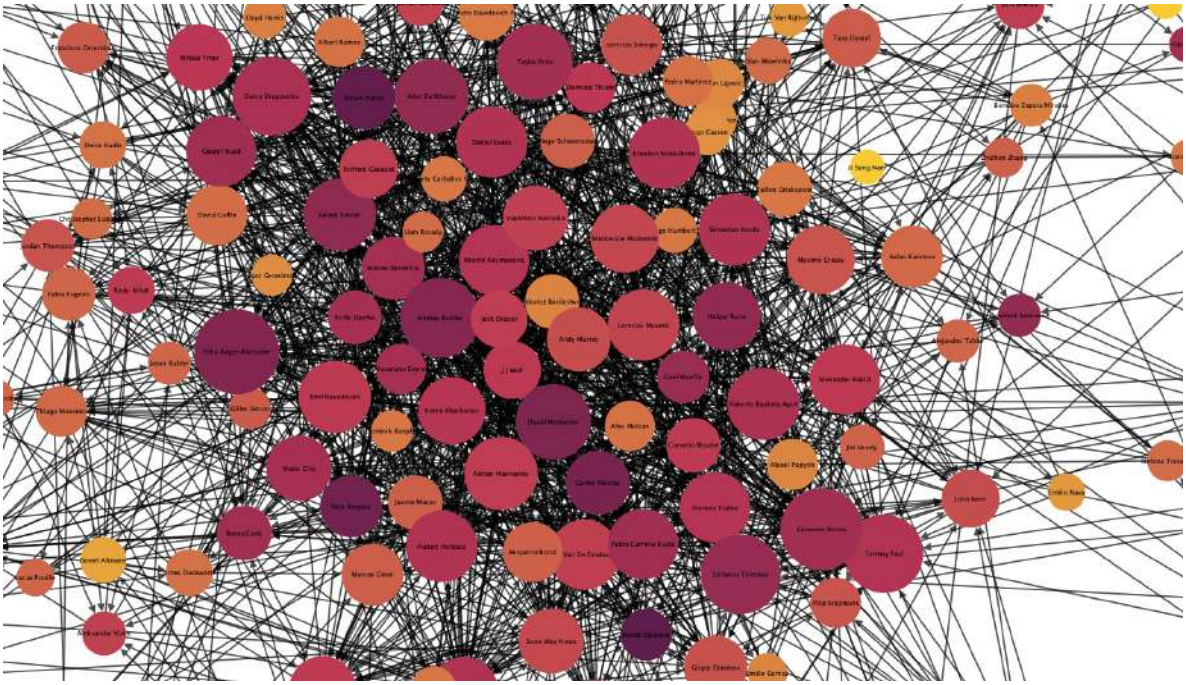


Figura 3: Grafo direcionado ampliado - temporada ATP 2022 em quadra dura

- Execução do algoritmo PageRank pré-implementado na biblioteca python *networkx*, com $\alpha = 0.85$, para obtenção do *score* de habilidade de cada jogador;

Apesar de a equivalência entre o modelo Bradley-Terry e o algoritmo PageRank ter sido demonstrada somente para $\alpha = 1$, é preciso levar em consideração que os grafos direcionados G de cada temporada são esparsos, visto que apenas alguns subconjuntos de jogadores se enfrentam. O grafo apresentado na figura 4 ilustra a esparsidade inerente aos dados de enfrentamento dos jogadores no circuito. Sob essa ótica, para construirmos um ranking comparável entre todos os jogadores, consideramos adequado o uso do fator de amortecimento $\alpha = 0.85$ (valor padrão definido no algoritmo).

4.3 Validações preliminares

Após a geração da nova métrica de ranking usando o PageRank, específico para cada superfície da quadra, foram realizados testes iniciais para verificar o sentido lógico dos resultados obtidos. Para isso, analisamos o ranking proposto e o *score* de habilidade obtido para os jogadores do Big-3 (Novak Djokovic, Rafael Nadal e Roger Federer) ao longo das temporadas entre 2004 e 2022, dados que estão representados nas figuras 5, 7 e 6. Abaixo, apresentamos como o ranking proposto retrata os momentos de melhor performance dos 3 jogadores, em cada superfície, ao longo de suas ilustres carreiras.

4.3.1 Rafael Nadal - 'Rei do saibro'

Rafael Nadal é mundialmente conhecido como o 'rei do saibro', não somente por ter sido 14 vezes campeão de Roland-Garros - o único Grand Slam disputado no saibro -, mas também por seu estilo de jogo muito adaptado à superfície, sempre buscando longas trocas de bola com os adversários. Seu sucesso não se concretizou apenas em Roland-Garros, mas também nos demais torneios importantes da temporada de saibro. A seguir, destacamos os torneios vencidos no saibro pelo jogador durante suas duas melhores temporadas ([ATPa]):

- Temporada de 2005
  - Masters 1000 Monte Carlo
  - ATP 500 Barcelona
  - Masters 1000 Roma

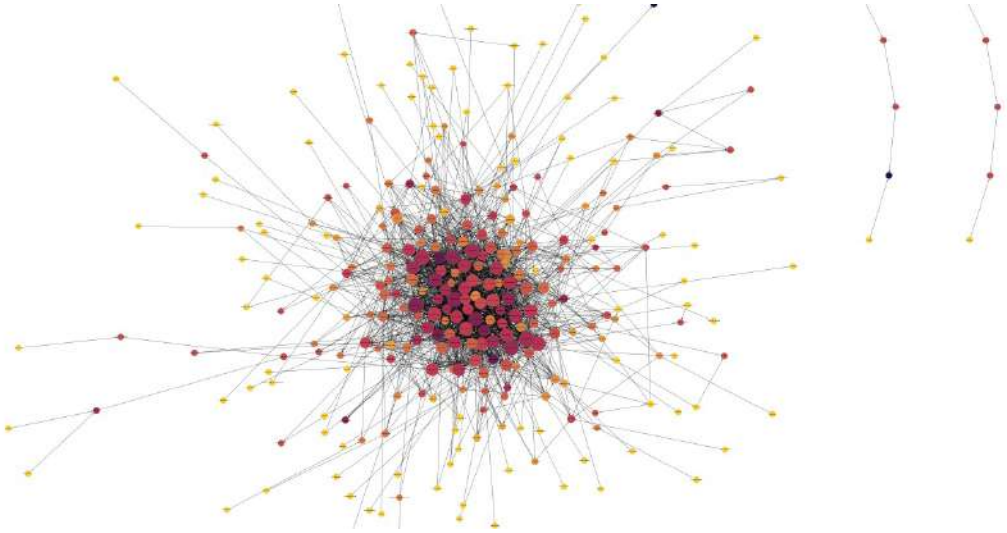


Figura 4: Grafo direcionado - temporada ATP 2022 em quadra dura

  - Roland-Garros

- Temporada de 2010
  - Masters 1000 Monte Carlo
  - Masters 1000 Madrid
  - Masters 1000 Roma
  - Roland-Garros

Podemos observar que o gráfico 5 ilustra bem o cenário de Rafael Nadal. Desde 2005 até 2021, o jogador apresentou *score* de habilidade no saibro dentro do top-3, sendo o melhor jogador da superfície em 8 temporadas. Apesar de não ser uma temporada de destaque para o jogador considerando todos os torneios disputados, o ano de 2017 mostra um elevado pico no *score* de habilidade. Esse resultado está condizente com as temporadas de 2005 e de 2010, visto que em 2017 o jogador também venceu 4 dos 5 principais torneios da temporada de saibro (Monte Carlo, Barcelona, Madrid e Roland-Garros). Sob essa ótica, analisando o desempenho de Rafael Nadal, o ranking proposto mostra-se coerente.

4.3.2 Novak Djokovic e Roger Federer - quadra rápida e grama

Em relação aos demais dois integrantes do Big-3, Novak Djokovic e Roger Federer apresentam estilos de jogo mais agressivo e rápido, que são mais eficientes nas quadras rápidas e de grama. Ambos os jogadores obtiveram excelentes resultados em tais superfícies, conquistando diversos Grand Slams e outros torneios muito importantes da ATP.

Considerando o *score* de habilidade para a grama, é válido considerar que existem poucos torneios disputados nessa superfície, cuja temporada é a mais curta do circuito. Assim, o vencedor de Wimbledon deve naturalmente estar associado aos *scores* mais altos, visto que precisa vencer 7 partidas consecutivas e jogadores no topo do ranking para conquistar o título. Na figura 6, podemos observar que Roger Federer apresenta o maior *score* de habilidade na grama em 6 dos 8 anos em que foi o vencedor do torneio (2003, 2004, 2005, 2006, 2007, 2009, 2012, 2017), sendo que o ano de 2003 não foi representado. Analogamente, em 5 das 7 vezes em que foi campeão de Wimbledon (2011, 2014, 2015, 2018, 2019, 2021, 2022), Novak Djokovic também apresenta o maior *score*.

No caso das quadras rápidas, que contêm o maior volume de torneios ao longo do ano, temos uma aproximação melhor do contexto global da temporada. Entre 2004 e 2007, Roger Federer conquistou 11 dos 16 Grand Slams disputados [ATPb], apresentando o maior *score* de habilidade em 3 temporadas, e a segunda posição em 2005. Por sua vez, Novak Djokovic, que teve o melhor desempenho de sua carreira em 2015 [ATPc], apresentou nesse ano o maior *score* de habilidade em quadras rápidas dentre todas as temporadas, considerando todos os jogadores do circuito.

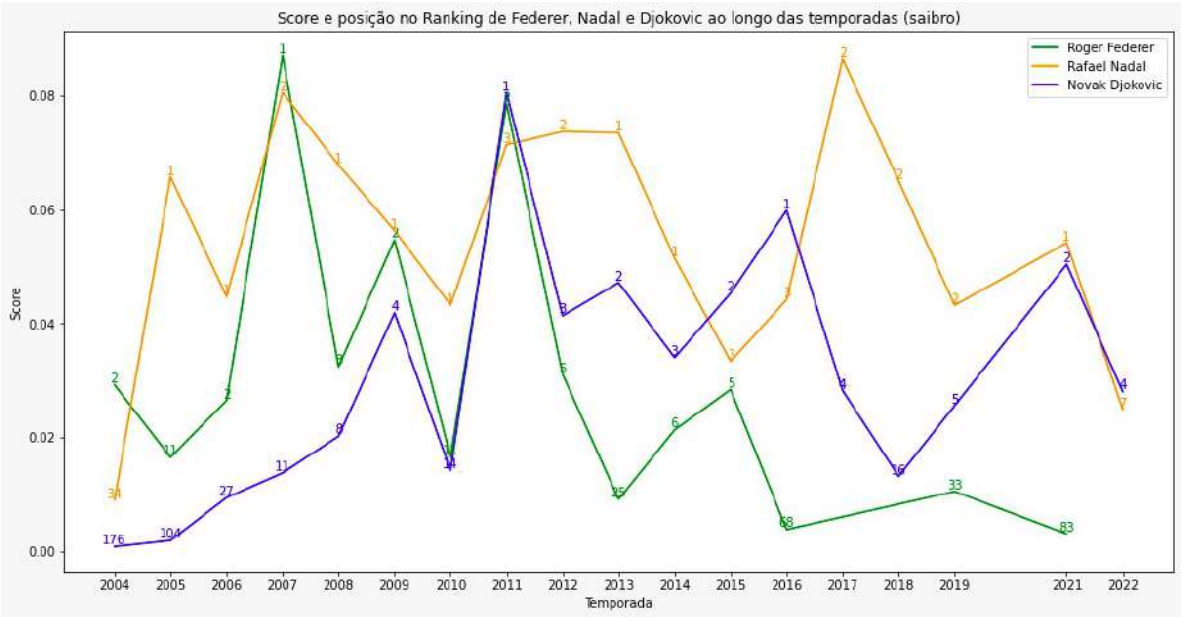


Figura 5: Performance do Big-3 em quadras de saibro, entre 2004 e 2022

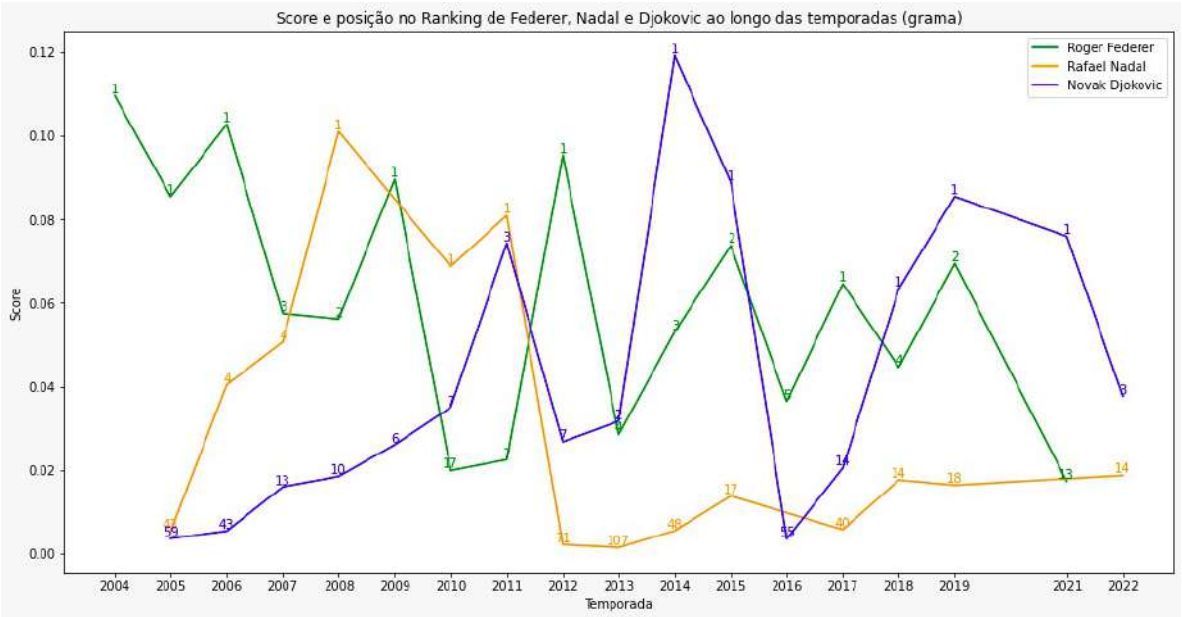


Figura 6: Performance do Big-3 em quadras de grama, entre 2004 e 2022

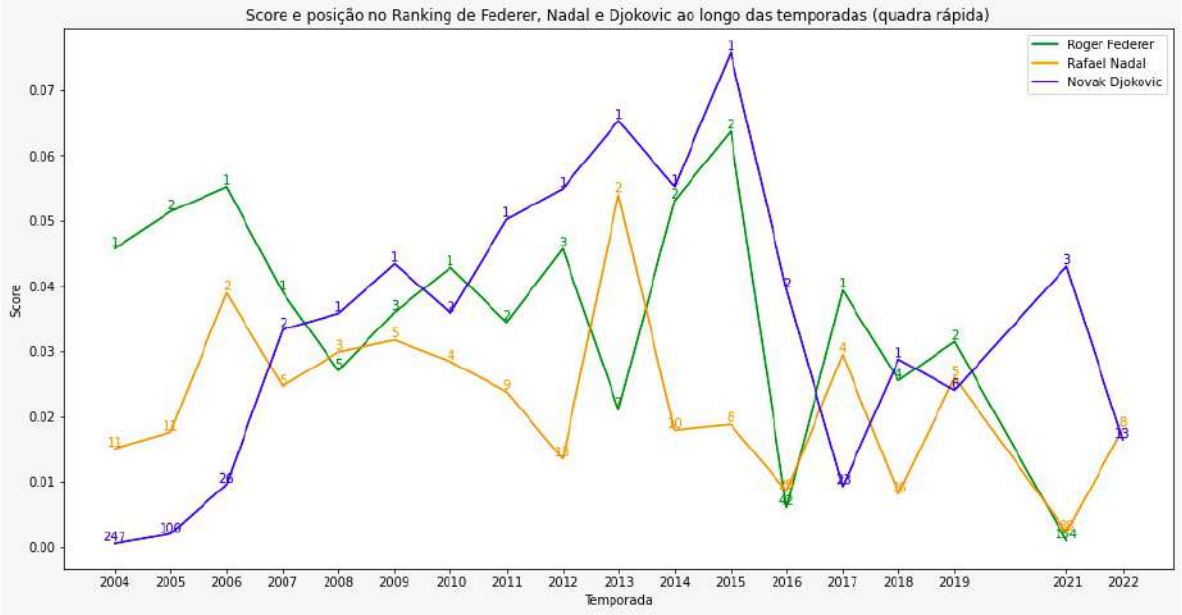


Figura 7: Performance do Big-3 em quadras rápidas, entre 2004 e 2022

Considerando os 3 gráficos anteriores (5, 6 e 7), é possível entender claramente o impacto do Big-3 para o tênis. Durante aproximadamente 20 anos consecutivos, foram líderes absolutos do ranking, vencendo mais de 80% dos Grand-Slams disputados. Apesar de terem maneiras de jogar distintas, todos conseguiram desenvolver um nível de tênis muito elevado nas 3 diferentes superfícies - o que é retratado pelos altos valores no *score* de habilidades, evidenciando tamanho domínio sobre o circuito masculino profissional de tênis.

4.4 Análise dos resultados

Para a avaliação dos *scores* de forma extensiva, buscamos comparar o ranking proposto com o ranking oficial da ATP sob a ótica de previsão dos *upsets*, já estudados na primeira etapa do trabalho. Para isso, após a obtenção do *score* de habilidade (μ) de cada jogador para as três superfícies, foi calculada a probabilidade de *upset* para cada partida da temporada correspondente, usando a formulação matemática do modelo de Bradley-Terry:

$$P(upset) = \frac{\mu_i}{\mu_i + \mu_j}; \mu_i < \mu_j \tag{4}$$

Além disso, foi calculada a razão de habilidade entre os jogadores de cada partida, como uma maneira de metrificar o nível de competitividade de cada jogo.

$$skill\_ratio = \frac{\mu_i}{\mu_j}; \mu_i < \mu_j \tag{5}$$

Com base nessas informações, agrupamos os jogos de cada temporada, em uma dada superfície, de acordo com o *skill_ratio* dos jogadores, dividido em faixas (*skill_ratio_bin*) de 0.1. Para cada faixa de *skill_ratio*, comparamos a média da probabilidade de *upsets* prevista pelo *score* com a taxa real de *upsets*, obtida por meio do ranking original da ATP para cada jogador. Veja na tabela 1 alguns exemplos de combinações de jogadores em cada *skill_ratio_bin*.

Os gráficos 8, 9, 10, 11, 12,13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24 e 25 mostram os resultados obtidos. No eixo x, temos o *skill_ratio_bin* (de 1 a 10), enquanto as linhas do eixo y representam os percentuais de *upset* e as colunas mostram o volume de partidas em cada bin. De maneira geral, podemos observar que a curva de *upsets* estimados a partir dos *scores* de habilidade apresenta variações mínimas entre temporadas, para uma mesma superfície. Além disso, a estimativa de *upsets* feita pelo cálculo dos *scores* se mostra uma boa aproximação para a taxa real de *upsets*, visto que as séries são

Tabela 1: *Exemplos de combinações de jogadores por skill ratio bin*

Temporada	Superfície	Jogador 1	Score 1	Jogador 2	Score 2	Score Ratio	Bin
2019	Grama	Felix A. Aliassime	0.0182	Grigor Dimitrov	0.0014	0.079	1
2022	Saibro	Carlos Alcaraz	0.0533	Alex De Minaur	0.0074	0.139	2
2021	Saibro	Novak Djokovic	0.0504	Jannik Sinner	0.0122	0.241	3
2017	Grama	Roger Federer	0.0646	Alexander Zverev	0.0216	0.335	4
2012	Quadra rápida	Andy Murray	0.0486	Milos Raonic	0.0199	0.411	5
2017	Saibro	Alexander Zverev	0.0230	Fernando Verdasco	0.0133	0.578	6
2015	Quadra rápida	Stan Wawrinka	0.0233	Andy Murray	0.0377	0.617	7
2022	Quadra rápida	Rafael Nadal	0.0180	Daniil Medvedev	0.0239	0.753	8
2019	Grama	Novak Djokovic	0.0855	Roger Federer	0.0694	0.812	9
2022	Quadra rápida	Carlos Alcaraz	0.0169	Casper Ruud	0.0166	0.978	10

altamente correlacionadas (tabela 2). Esse comportamento pode ser explicado pela sutil semelhança entre o PageRank e a metodologia atual de ranqueamento da ATP.

No caso do ranking da ATP, conforme mencionado na seção 3.1, existem dois fatores principais que influenciam na pontuação atribuída ao jogador:

- Categoria do torneio (ATP 250, ATP 500, Masters 1000, Grand Slams);
- Rodada atingida pelo jogador (quartas de final, semifinal, final, ...);

É importante destacar que os torneios maiores contam com a participação de uma maior quantidade de jogadores bem colocados no ranking, inclusive por acontecerem de maneira isolada, enquanto podem ocorrer diferentes torneios ATP 250 e 500 simultaneamente, em países distintos. Essa concorrência de torneios menores acaba dividindo os melhores jogadores em algumas semanas da temporada, enquanto a maioria absoluta sempre se reúne nos Grand Slams. Note que, para vencer um Grand Slam, naturalmente será necessário vencer vários jogadores bem colocados no ranking (no top-20, por exemplo), fato que também tem uma contribuição positiva no processamento do PageRank. Alternativamente, em torneios menores, se enfrentam jogadores em patamares menores do ranking, que terão um peso menor no PageRank.

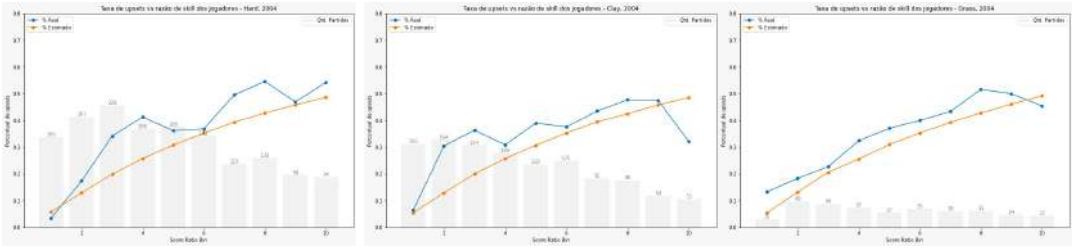
Sob essa ótica, podemos perceber que o PageRank e a metodologia de ranking da ATP são semelhantes devido à dinâmica de participação dos jogadores em torneios de diferentes categorias. A principal diferença está na quantificação: enquanto a ATP atribui os pontos de forma discreta, o PageRank é contínuo. Além disso, a metodologia da ATP não faz distinção entre resultados esperados e inesperados. Por exemplo, vencer a semifinal de um Grand Slam tem sempre a mesma importância, independente de o adversário ser top-3 ou top-100. No caso do PageRank, vencer um top-3 terá uma contribuição maior para o *score* de habilidade do que vencer um top-100. Esse fato explica, em parte, por que a taxa real de *upsets* é geralmente maior do que o valor estimado pelo algoritmo. Ao considerar a qualidade do adversário como um fator relevante para julgar a performance de um jogador, o *score* de habilidades baseado no PageRank se mostra uma possível melhoria para a metodologia de ranking vigente.

Tabela 2: *Correlação de pearson entre upsets reais e estimados*

Superfície	Correlação
Grama	0.815
Saibro	0.868
Quadra rápida	0.925

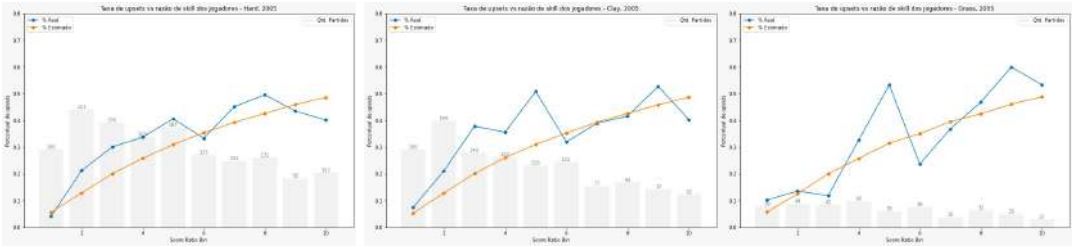
5 Fechamento

A motivação para a pesquisa desenvolvida foi o estudo dos *upsets* feito na primeira etapa do trabalho, que levou ao questionamento sobre a relação entre o ranking e a real habilidade dos jogadores. Com esse objetivo, foi desenvolvida uma metodologia para quantitficar a habilidade dos jogadores no circuito profissional de tênis, visando entender a qualidade do sistema de ranking e explicar a ocorrência de *upsets*. Utilizando dados das temporadas de 2004 a 2022, implementamos o algoritmo PageRank para computar o *score* de habilidade dos jogadores em cada superfície. Considerando a equivalência



(a) Quadra rápida (b) Saibro (c) Grama

Figura 8: Temporada de 2004 em diferentes superfícies



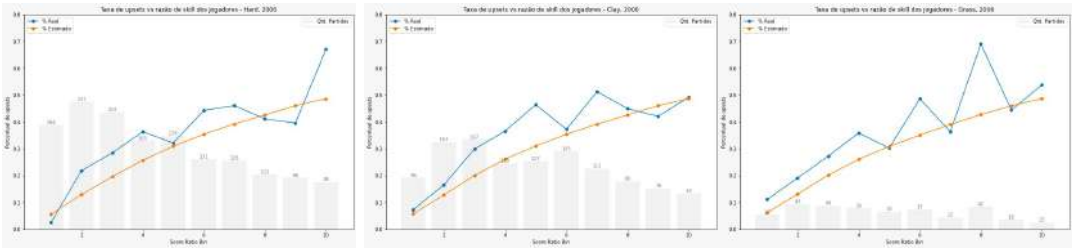
(a) Quadra rápida (b) Saibro (c) Grama

Figura 9: Temporada de 2005 em diferentes superfícies

do PageRank com o modelo de Bradley-Terry, empregamos este modelo para prever o resultado das partidas de cada temporada. Além do estudo da tendência de *upsets* em cada superfície, fizemos uma análise detalhada dos *scores* e rankings propostos para o Big-3 ao longo dos anos, comparando os resultados do modelo com a performance dos jogadores. Nossos resultados demonstraram que a tendência de *upsets* com base no ranking original é altamente correlacionada com as estimativas feitas pelo modelo proposto.

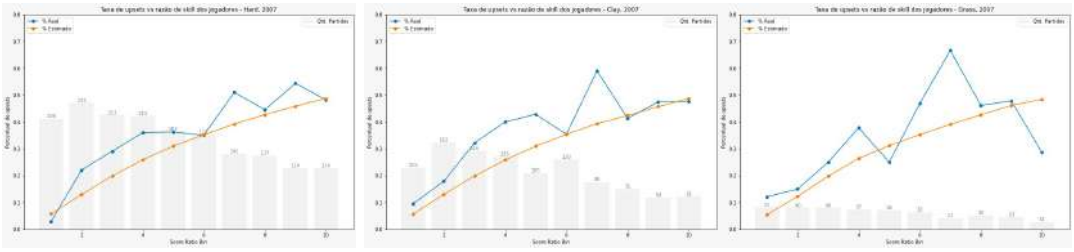
A semelhança entre o PageRank e o sistema de ranking pode ser explicada pela estrutura de organização dos torneios no circuito profissional de tênis. No entanto, identificamos que o ranking atual apresenta mais *upsets* porque não considera a qualidade do adversário vencido ao atribuir pontos ao jogador, mas apenas o nível do torneio e a rodada são relevantes. Portanto, o modelo proposto representa uma melhoria significativa para a medida de ranking atual, permitindo também a análise segmentada da habilidade dos jogadores por superfície.

Para futuros trabalhos, sugerimos inicialmente a aplicação desta metodologia ao circuito feminino, além da utilização do modelo para a previsão de resultados de partidas. Adicionalmente, há oportunidades para o desenvolvimento de projetos de pesquisa que utilizem outras bases de dados encontradas no desenvolvimento deste trabalho, as quais contêm informações detalhadas de cada batida dos jogadores ao longo dos jogos. Essa abordagem pode fornecer insights ainda mais profundos sobre o desempenho e a habilidade dos jogadores de tênis, contribuindo para o avanço da análise estatística e preditiva no esporte.



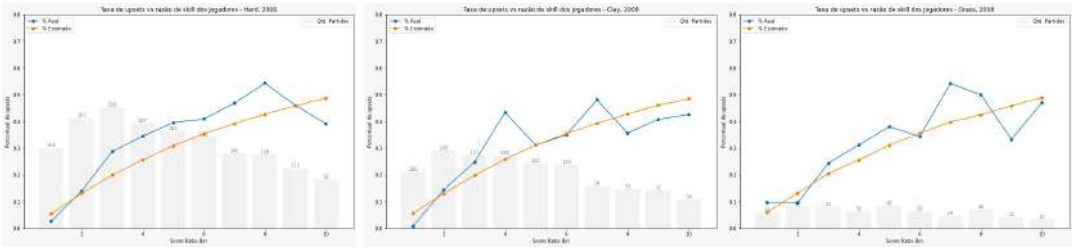
(a) Quadra rápida (b) Saibro (c) Grama

Figura 10: Temporada de 2006 em diferentes superfícies



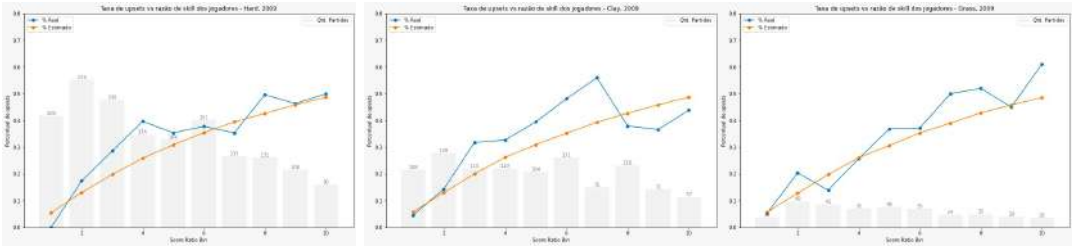
(a) Quadra rápida (b) Saibro (c) Grama

Figura 11: Temporada de 2007 em diferentes superfícies



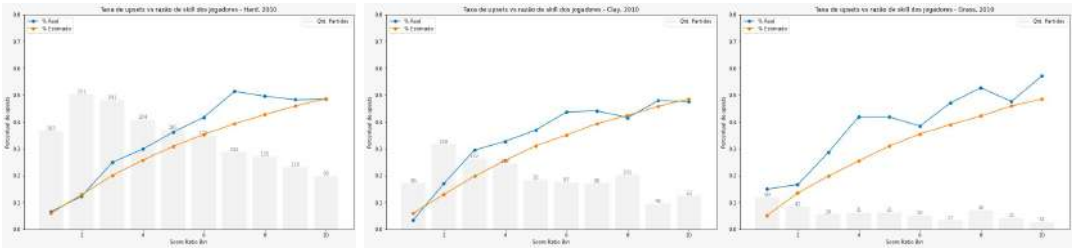
(a) Quadra rápida (b) Saibro (c) Grama

Figura 12: Temporada de 2008 em diferentes superfícies



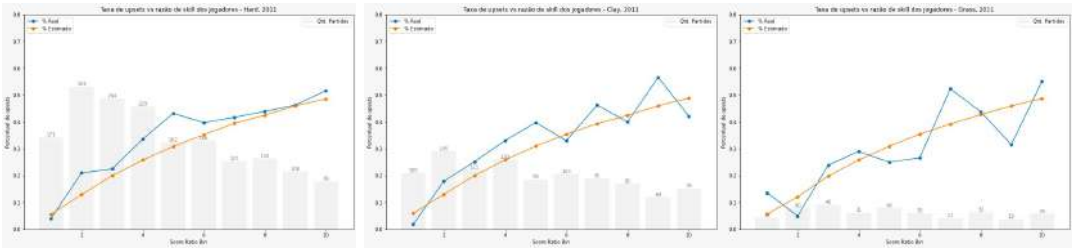
(a) Quadra rápida (b) Saibro (c) Grama

Figura 13: Temporada de 2009 em diferentes superfícies



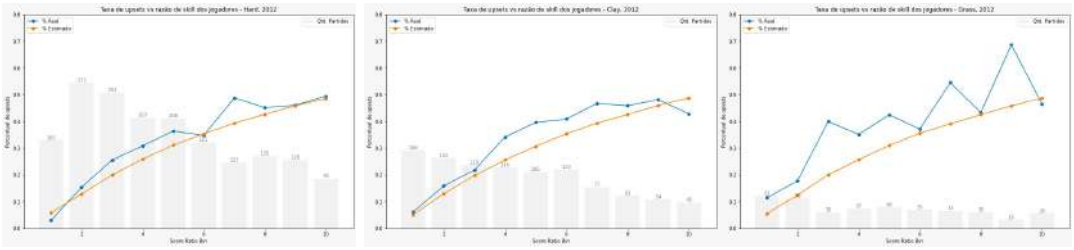
(a) Quadra rápida (b) Saibro (c) Grama

Figura 14: Temporada de 2010 em diferentes superfícies



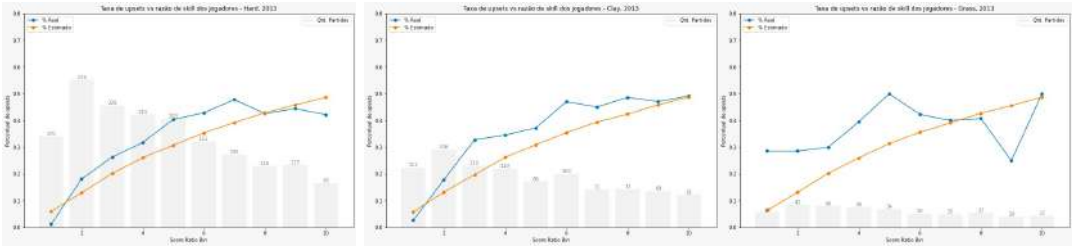
(a) Quadra rápida (b) Saibro (c) Grama

Figura 15: Temporada de 2011 em diferentes superfícies



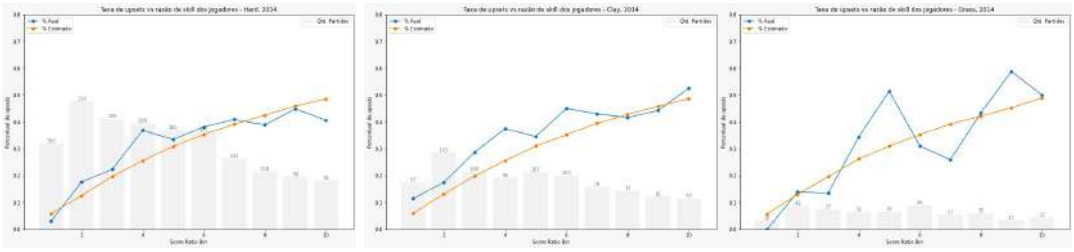
(a) Quadra rápida (b) Saibro (c) Grama

Figura 16: Temporada de 2012 em diferentes superfícies



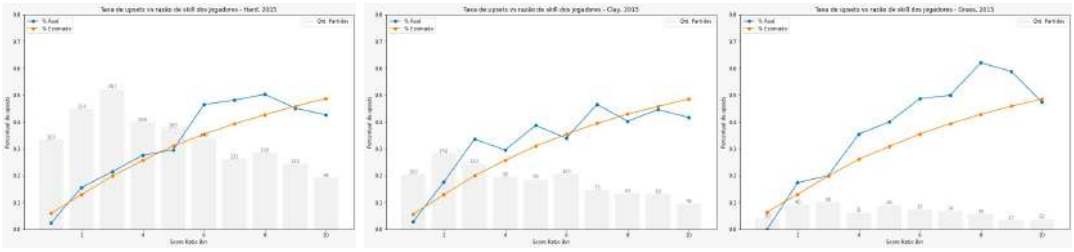
(a) Quadra rápida (b) Saibro (c) Grama

Figura 17: Temporada de 2013 em diferentes superfícies



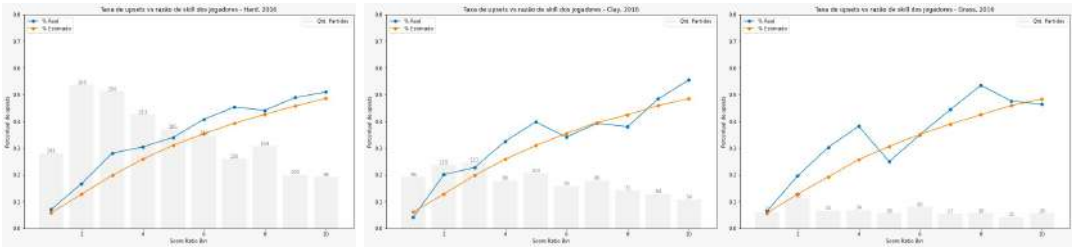
(a) Quadra rápida (b) Saibro (c) Grama

Figura 18: Temporada de 2014 em diferentes superfícies



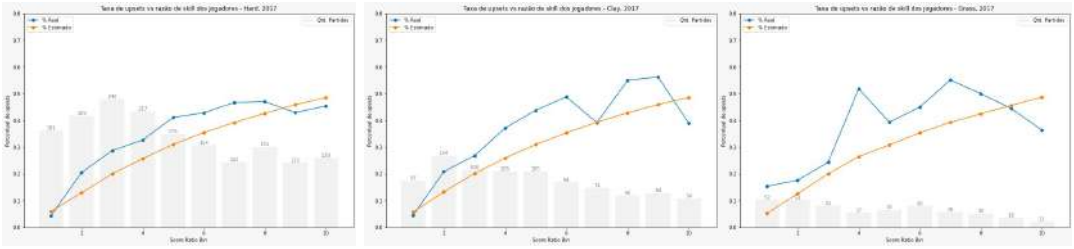
(a) Quadra rápida (b) Saibro (c) Grama

Figura 19: Temporada de 2015 em diferentes superfícies



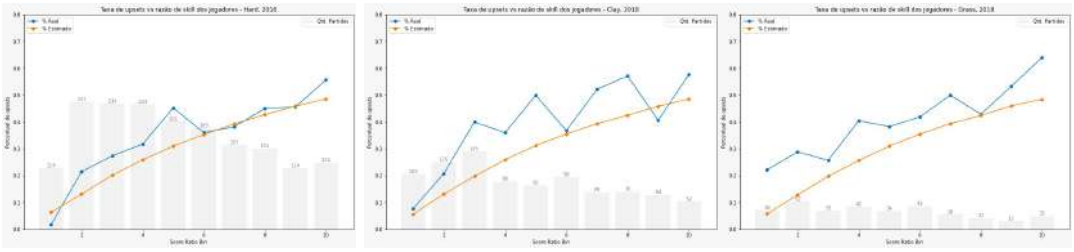
(a) Quadra rápida (b) Saibro (c) Grama

Figura 20: Temporada de 2016 em diferentes superfícies



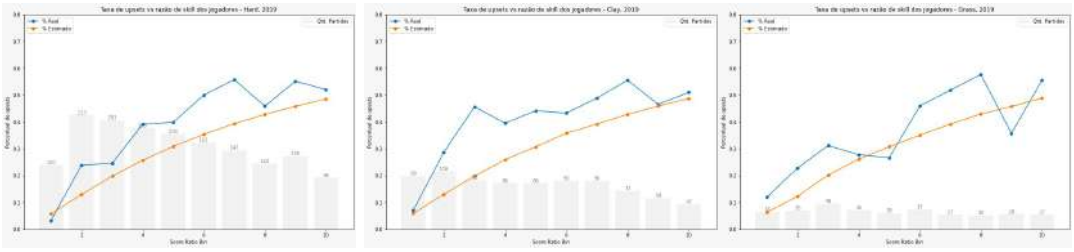
(a) Quadra rápida (b) Saibro (c) Grama

Figura 21: Temporada de 2017 em diferentes superfícies



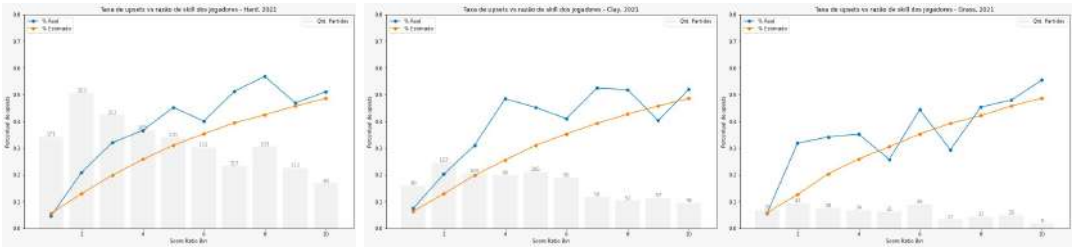
(a) Quadra rápida (b) Saibro (c) Grama

Figura 22: Temporada de 2018 em diferentes superfícies



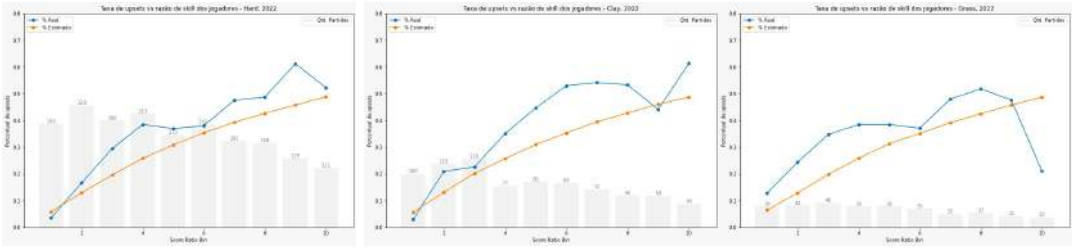
(a) Quadra rápida (b) Saibro (c) Grama

Figura 23: Temporada de 2019 em diferentes superfícies



(a) Quadra rápida (b) Saibro (c) Grama

Figura 24: Temporada de 2021 em diferentes superfícies



(a) Quadra rápida (b) Saibro (c) Grama

Figura 25: Temporada de 2022 em diferentes superfícies

Referências

[ATPa] https://www.atptour.com/en/news/nadal-infosys-btn-september-2023. Acesso em 24/07/2024.

[ATPb] https://www.atptour.com/en/news/roger-federer-best-seasons-career-feature. Acesso em 24/07/2024.

[ATPc] https://www.eurosport.com/tennis/after-bitter-end-how-does-2023-compare-to-novak-djokovic-s-
sto9898892/story.shtml. Acesso em 24/07/2024.

[ATPd] Regras gerais do ranking atp. https://www.atptour.com/en/rankings/rankings-faq. Acesso em 02/04/2024.

[JMS20] Ali Jadbabaie, Anuran Makur, and Devavrat Shah. Estimation of skill distribution from a tournament. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8418–8429. Curran Associates, Inc., 2020.

[poi] Pontuações para o ranking atp 2024. https://www.reddit.com/r/tennis/comments/18rxbpg/jose_morgado_new_atp_ranking_points_breakdown/?rdt=59510. Acesso em 02/04/2024.

[Sac23a] Jeff Sackmann. Repositório tennis-atp no github. https://github.com/JeffSackmann/tennis_atp, 2023. Acesso em: 20/09/2023.

[Sac23b] Jeff Sackmann. Repositório tennis-wta no github. https://github.com/JeffSackmann/tennis_wta, 2023. Acesso em: 20/09/2023.

[Sel24] David Antony Selby. Pagerank and the bradley-terry model. *arXiv*, 2024.

[Tow] https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-
Acesso em 16/07/2024.