UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE FÍSICA

André Seiji Sunahara

Padrões de Impacto e Produtividade em Carreiras Científicas

Maringá, 29 de novembro de 2023.

UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE FÍSICA

André Seiji Sunahara

Padrões de Impacto e Produtividade em Carreiras Científicas

Tese de doutorado apresentada ao Programa de Pós-Graduação em Física da Universidade Estadual de Maringá.

Orientador: Prof. Dr. Haroldo Valentin Ribeiro

Maringá, 29 de novembro de 2023.

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

S957p

Sunahara, André Seiji

Padrões de impacto e produtividade em carreiras científicas / André Seiji Sunahara. -- Maringá, PR, 2023.

139 f.: il. color., figs., tabs.

Orientador: Prof. Dr. Haroldo Valentin Ribeiro.

Tese (Doutorado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Física, Programa de Pós-Graduação em Física, 2023.

1. Sistemas complexos. 2. Física estatística. 3. Análise de dados. I. Ribeiro, Haroldo Valentin, orient. II. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Física. Programa de Pós-Graduação em Física. III. Título.

CDD 23.ed. 530.072

Rosana de Souza Costa de Oliveira - 9/1366

ANDRE SEIJI SUNAHARA

PADRÕES DE IMPACTO E PRODUTIVIDADE EM CARREIRAS CIENTÍFICAS

Tese apresentada à Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de doutor.

Aprovado em: Maringá, 29 de novembro de 2023.

BANCA EXAMINADORA

_____

Prof. Dr. Haroldo Valentin Ribeiro
Universidade Estadual de Maringá – UEM

_____

Prof. Dr. José Soares de Andrade Júnior
Universidade Federal do Ceará – UFC

_____

Prof. Dr. Hygor Piaget Monteiro Melo
Sony Computer Science Laroratories Rome - Itália

_____

Prof. Dr. Luiz Roberto Evangelista
Universidade Estadual de Maringá – UEM

_____

Prof. Dr. Luis Carlos Malacarne
Universidade Estadual de Maringá – UEM

Resumo

Nesta tese, exploramos diversos tópicos da disciplina de ciência da ciência (*science of science*) usando a Plataforma Lattes como fonte primária de dados. Na primeira parte, estudamos a associação entre indicadores de produtividade e impacto de jornal ao longo de carreiras científicas. Encontramos que essa relação é específica para cada disciplina, dependente do estágio da carreira e similar entre pesquisadores com performances *outlier* ou não *outlier*. Pesquisadores *outliers* têm performances muito acima da média em produtividade ou em impacto de jornal, mas raramente conseguem atingir esses mesmos níveis de performance em ambas as categorias simultaneamente. Pesquisadores não *outliers* mostram uma associação negativa entre produtividade e impacto de jornal com intensidades específicas para cada disciplina. Pesquisadores em nossa base de dados tendem a manter os níveis de produtividade e impacto de jornal similares em anos consecutivos. Eles também apresentam padrões médios de carreira de produtividade e impacto de jornal específicos para cada disciplina, mas que frequentemente são compostos por padrões com um máximo no começo da carreira para impacto de jornal e padrões crescentes para produtividade com a progressão da carreira. Na segunda parte, restringimos nossa atenção ao indicador produtividade. Usando dados de carreiras científicas individuais, empregamos uma abordagem orientada por dados composta por métodos de análise de séries temporais, redução de dimensionalidade e ciência de redes para revelar padrões universais de produtividade em carreiras científicas. Encontramos seis padrões gerais de produtividade: constante, em forma de U, decrescente, periódico, crescente e com aspecto canônico, sendo esse último padrão referente a trajetórias nas quais a produtividade aumenta no início da carreira, passa por um pico e depois diminui com o avanço da idade do pesquisador. Carreiras crescentes e com aspecto canônico representam quase três quartos dos pesquisadores, sendo a categoria de aspecto canônico a mais prevalente. Trajetórias curtas mais frequentemente são representadas por carreiras crescentes, enquanto trajetórias longas mais frequentemente são representadas por carreiras com aspecto canônico. De maneira contrária às expectativas da literatura, encontramos que carreiras com aspecto canônico apresentam mais frequentemente um pico em produtividade na região intermediária da carreira em vez de no início da carreira.

Palavras-chave: Sistemas Complexos. Ciência da Ciência. Análise de Dados. Física Estatística.

Abstract

In this thesis, we explore a diverse set of topics on the discipline of science of science using the Lattes Platform as our primary data source. In the first part, we study the association between indicators of productivity and journal impact across scientific careers. We find this relationship is discipline-specific, career-age dependent, and similar among researchers with outlier and nonoutlier performance. Outlier researchers outperform either in productivity or journal impact, but rarely outperform simultaneously in both categories. Nonoutlier researchers display a negative association between productivity and journal impact with discipline-specific intensity. Researchers in our dataset are prone to maintain their productivity and journal impact levels in consecutive years. They also display average career patterns of productivity and journal impact that are discipline-specific, but that often present an early-career high in journal impact and an increasing pattern in productivity with career progression. In the second part, we focus solely on the productivity indicator. Using individual data on scientific careers, we employ a data-driven approach based on methods of time series analysis, dimensionality reduction, and network science to uncover the universal productivity patterns in research careers. We find six career patterns of productivity: constant, u-shaped, decreasing, periodic-like, increasing, and canonical, with the latter pattern being that of trajectories in which productivity increases in early maturity, peaks, and then decreases with career progression. Increasing and canonical careers comprise almost three-fourths of researchers, with the canonical category being the most prevalent. Shorter trajectories are more often represented by increasing careers, while longer trajectories are more often represented by canonical careers. Contrary to expectations, the canonical pattern displays a productivity peak more frequently around mid-career rather than in the beginning.

Keywords: Complex Systems. Science of Science. Data Science. Statistical Physics.

Sumário

Introdução

A crescente disponibilidade de informação tem possibilitado esforços interdisciplinares em direção de um melhor entendimento quantitativo da empreitada científica: uma ciência da ciência [1–3]. Para além da questão acadêmica de encontrar os mecanismos que impulsionam a ciência, essas iniciativas visam melhorar a eficiência científica por meio da identificação de boas práticas e políticas, que incluem, num aspecto mais amplo, a escolha de prioridades científicas nacionais até, num aspecto mais local, a seleção de projetos de pesquisa e a contratação de professores. Atualmente, o progresso científico é fortemente dependente dos processos de avaliação, pois eles regulam o fluxo de ideias viabilizando projetos de pesquisa por meio da alocação de recursos financeiros [2, 4–6]. Nesse contexto, a revisão por pares é considerada a abordagem padrão para avaliar performance acadêmica [7]. Entretanto, esse processo é laborioso e apresenta várias desvantagens como viés, falta de consistência e até mesmo fraude [7–10]. O número crescente de publicações científicas [11] e a expansão da classe trabalhadora científica [12] acarretam limitações adicionais para o método de revisão por pares [6]. Como consequência direta dessas dificuldades, houve um crescimento no uso de índices bibliométricos (ou bibliometrias) para a classificação da performance acadêmica [6, 13], especialmente depois dos anos 2000 [14].

É fato que avaliações por meio de bibliometrias apresentam caráter mais objetivo. Porém, não existe um consenso sobre quais índices são os mais adequados para mensurar performance acadêmica. Pesquisas recentes corroboram essa indefinição sugerindo que a natureza intrínseca dos processos científicos só pode ser precisamente quantificada por abordagens multidimensionais [15, 16]. Para além da questão sobre a viabilidade da avaliação, o uso de bibliometrias impõe uma enorme pressão aos cientistas, particularmente aos mais jovens [17], para publicar em grandes quantidades, em jornais de prestígio[1] e para desenvolver pesquisas altamente citadas [6, 18, 19]. Por isso, o uso de bibliometrias tem sido alvo de muitas críti-

[1] A partir daqui, utilizaremos os termos *prestígio* e *impacto* de modo intercambiável.

cas [7, 20–22]. É nesse contexto controverso que a produtividade e as medidas de impacto são frequente e amplamente utilizadas para quantificar performance acadêmica. Se por um lado a produtividade é simplesmente definida como o número de documentos acadêmicos produzidos num dado período, o impacto tem um caráter mais subjetivo e usualmente é medido pelo número de citações, pela fração de documentos entre os mais citados ou pelo prestígio do meio de publicação. A presente tese consiste de investigações sobre diversos aspectos da vida acadêmica de pesquisadores brasileiros relacionados aos indicadores bibliométricos produtividade e impacto de jornal, utilizando a Plataforma Lattes como base de dados primária.

No Capítulo 2, estudamos a inter-relação entre os indicadores produtividade e impacto de jornal. A avaliação de pesquisas científicas por meio de bibliometrias tem levantado um debate sobre "qualidade *versus* quantidade" desde sua concepção [23–38] e ainda não existe consenso sobre a natureza da associação entre essas duas variáveis. Por exemplo, enquanto Larivière e Costas [33] encontraram uma associação positiva entre a produtividade e o número de artigos altamente citados, Bornmann e Tekles [37] mostraram que os autores mais produtivos têm geralmente uma fração menor de publicações entre os artigos mais citados, isto é, uma associação negativa entre produtividade e impacto em níveis muito elevados de produtividade. Essas discrepâncias refletem determinadas características da associação entre produtividade e impacto, pois a associação depende da disciplina, do estágio da carreira, da escala e da presença de indivíduos *outliers*[2]. Todavia, ainda existe uma escassez de trabalhos que levam em consideração todos esses fatores simultaneamente para revelar a complexidade geral da relação "quantidade *versus* qualidade".

Investigamos aspectos multifacetados dessa associação ao analisar a carreira científica de mais de 6 mil cientistas brasileiros de 14 disciplinas. Determinamos seus números de publicações anuais e os respectivos valores médios do impacto de jornal. É importante pontuar que o uso de métricas em nível de jornal para avaliar a performance individual é bastante controverso [20, 39]. No entanto, essa abordagem ainda permanece difundida e amplamente utilizada [40], especialmente no Brasil, em que várias universidades e agências de fomento usam o prestígio de jornal ou indicadores derivados para diversas tarefas [41], desde a contratação de professores [42] até a concessão de recursos financeiros [43]. Além disso, trabalhos recentes demostram que métricas em nível de jornal carregam informação sobre a performance acadêmica [44–48] e são correlacionadas com citações, indicando que essas duas métricas podem ser parcialmente consideradas como substitutas. Apesar de não termos uma resposta definitiva se métricas em nível de jornal, ou até mesmo as citações, são apropriadas para avaliação científica, fato é que essas métricas são importantes para a comunidade acadêmica. Dessa forma, novas investigações podem trazer à luz aspectos

[2]Indivíduos que apresentam produtividade ou impacto de jornal com valores atípicos e muito maiores do que os valores médios calculados a partir de pesquisadores da mesma disciplina em um determinado ano.

relevantes para melhorar o processo de avaliação.

Nossa pesquisa examina padrões na associação entre produtividade e métricas de jornal em carreiras de pesquisadores de diferentes disciplinas. Em contraste com trabalhos anteriores, usamos medidas padronizadas para levar em consideração efeitos de inflação temporal e de especificidade das disciplinas. A medida padronizada referente ao prestígio de jornal também corrigiu vieses relacionados à escala da produtividade associada. Nossos resultados permitiram identificar indivíduos *outliers* em produtividade e/ou em impacto de jornal. Mostramos que esses acadêmicos performam muito acima da média em produtividade ou em impacto de jornal, mas raramente em ambas as categorias. Também descobrimos que os acadêmicos são aversos a mudanças simultâneas nos níveis de produtividade e impacto de jornal, preferindo manter esses níveis aproximadamente constantes em anos consecutivos de suas carreiras. Para indivíduos não *outliers*, nossos resultados indicam uma correlação negativa entre produtividade e prestígio de jornal para a maioria dos pesquisadores e para a maioria das disciplinas. Porém, mostramos que os padrões médios de carreira de produtividade e prestígio de jornal são específicos para cada disciplina, com fato comum de que o impacto de jornal é maior no início das carreiras e a produtividade média cresce com o tempo atingindo um pico ou um platô.

No Capítulo 3, estudamos as carreiras científicas de pesquisadores brasileiros pela perspectiva do indicador produtividade. Os padrões de produtividade em carreiras acadêmicas desde muito tempo têm sido objeto de estudo de investigações científicas. A monografia de Lehman é considerada como o trabalho seminal nessa linha de pesquisa [49]. Em 1953, Lehman observou que a contribuição agregada média de cientistas, compositores musicais, artistas e escritores exibia um padrão de produtividade crescente no início da carreira seguido por um declínio gradual com o tempo. Esse padrão foi observado em vários outros contextos e conjuntos de dados, sendo assim reconhecido na literatura como a "narrativa canônica de produtividade" [49–60].

Todavia, pesquisas recentes contestam a existência de um padrão universal de produtividade em carreiras científicas. As evidências sugerem a prevalência de uma variedade de padrões, que inclui trajetórias constantes [53, 58], decrescentes [54, 61, 62], crescentes [58] e periódicas [52, 63, 64]. Entretanto, muitas dessas investigações usam dados agregados, que podem introduzir vieses derivados da "falácia composicional" [56]. Esse problema é definido como o comportamento médio estimado a partir das séries temporais de vários indivíduos que não corresponde aos padrões observados para cada série individualmente. Alguns desses estudos restringiram as análises a conjuntos de dados reduzidos, com limitações na quantidade de anos de carreira analisados [61, 62] ou utilizando disciplinas específicas [61, 62, 65]. Além disso, frequentemente, investigações basearam suas estimativas em regressões lineares [55, 58, 61, 62, 64, 65], que podem não capturar toda a complexidade dos padrões de produtividade. Alguns autores também propuseram modelos generativos para curvas de

produtividade [55, 56, 66], mas não foram capazes de validar esses padrões com evidências empíricas.

Estudos de larga escala que investigam individualmente as formas das trajetórias de produtividade são escassos, com o trabalho de Way *et al.* [65] sendo uma das poucas exceções. Com dados de mais de dois mil professores de ciência da computação dos Estados Unidos e do Canadá, eles aplicaram um modelo linear segmentado, composto por duas retas em sequência, em carreiras científicas para avaliar a universalidade da narrativa de produtividade canônica. A pesquisa deles encontrou que quase metade das carreiras nesse conjunto de dados é consistente com padrões estritamente constantes, crescentes ou decrescentes. Por outro lado, apenas 20% das trajetórias apresentaram um crescimento rápido seguido por um declínio gradual em produtividade, o que sugere que a narrativa canônica pode não ser tão prevalente quanto se imaginava anteriormente. Entretanto, o uso de regressões segmentadas limita a emergência de padrões não lineares, como trajetórias periódicas, e o foco na disciplina de ciência da computação pode limitar a generalização dessas conclusões para outras disciplinas acadêmicas. Além disso, as pesquisas anteriores não levaram em consideração que mudanças estruturais na empreitada científica – como o aumento na colaboração científica [67, 68] e a pressão para produzir em grandes quantidades [19, 69, 70] – podem impactar a cultura de pesquisa de diferentes gerações e suas trajetórias de produtividade.

Para esclarecer essas questões, investigamos as trajetórias acadêmicas de produtividade de mais de oito mil pesquisadores da comunidade científica brasileira, englobando mais de cinquenta disciplinas acadêmicas. Empregamos uma abordagem orientada por dados que combina métodos de análise de séries temporais, redução de dimensionalidade e ciência de redes para agrupar as trajetórias de produtividade de acordo com as dissimilaridades calculadas para cada par de curvas. Diferentemente de trabalhos anteriores, nossa abordagem considera as trajetórias individualmente, a natureza ruidosa das trajetórias, a taxa de inflação temporal em produtividade específica para cada disciplina [59, 71, 72] e possíveis efeitos geracionais. Notadamente, não assumimos nenhuma forma predeterminada para as curvas de produtividade, o que possibilita a emergência natural dos padrões universais de produtividade. Nossa pesquisa identifica padrões de produtividade que foram apenas conjecturados [64] ou encontrados em estudos utilizando dados agregados [53, 54, 58, 61, 62, 64]. Identificamos seis categorias de trajetórias de produtividade: constante, em forma de U, decrescente, periódica, crescente e canônica, sendo que as últimas duas categorias representam quase três quartos dos pesquisadores. As trajetórias crescentes são muito mais frequentes entre pesquisadores em início de carreira do que entre pesquisadores seniores, enquanto trajetórias canônicas são muito mais frequentes entre pesquisadores seniores do que entre pesquisadores em início de carreira. No entanto, os anos iniciais das carreiras de pesquisadores seniores são categorizados como padrões crescentes com prevalência levemente menor do que a prevalência para pesquisadores mais novos. Apenas uma pequena fração de pes-

quisadores seniores com tendências inicialmente crescentes é capaz de manter esse padrão, enquanto grande parte do restante progride para carreiras canônicas.

A tese está dividida em quatro capítulos. O Capítulo 1 fundamenta os métodos utilizados nos dois trabalhos. O Capítulo 2 apresenta o trabalho sobre a associação entre produtividade e impacto de jornal considerando efeitos de disciplina, escala e carreira [59]. O Capítulo 3 apresenta o trabalho sobre padrões universais de produtividade em carreiras científicas [73]. Por fim, o Capítulo 4 apresenta a nossa discussão dos resultados e as perspectivas futuras. Além dos trabalhos apresentados nesta tese que se enquadram na disciplina de ciência da ciência, o autor também produziu outros dois artigos sobre a pandemia de COVID-19 [74,75] durante o doutorado.

CAPÍTULO 1

Métodos para análise de dados

Neste capítulo, fundamentamos os métodos empregados nas duas investigações que compõem esta tese. Sugerimos a leitura a partir do capítulo seguinte para o leitor com maior familiaridade com os tópicos a seguir.

1.1 Regressão logística

A regressão logística é empregada no estudo de variáveis binárias, isto é, uma variável dependente y_i que representa o sucesso ($y_i = 1$) ou o fracasso ($y_i = 0$) de um evento i [76]. A Figura 1.1A mostra um gráfico de dispersão de uma variável dependente binária y em função de uma variável independente contínua x. Podemos entender cada y_i como um ensaio de Bernoulli [77], com probabilidade $P(y)$ definida como

$$P(y) = \pi(x)^y [1 - \pi(x)]^{1-y} , \tag{1.1}$$

em que $\pi(x)$ é a probabilidade de sucesso e $[1-\pi(x)]$ é a probabilidade de fracasso do ensaio. No modelo logístico, modelamos a probabilidade $\pi(x)$ como função de uma variável arbitrária x. Para cada valor de x, existe uma proporção de sucessos que corresponde à probabilidade $\pi(x)$. Parametrizamos o valor de y por meio de uma função sigmoide para mapear a variável x para o intervalo $[0, 1]$ [77]

$$S(y) = \frac{\exp y}{1 + \exp y} .$$

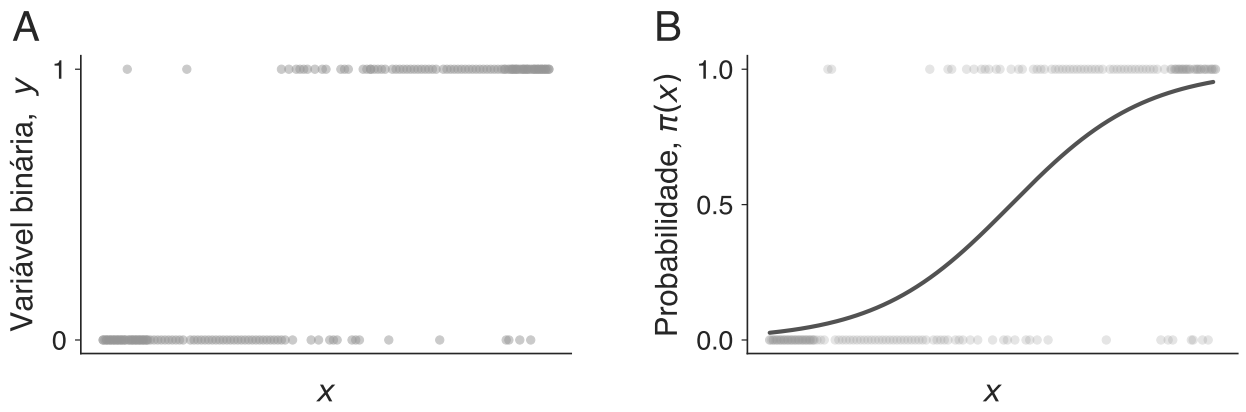

Figura 1.1: Exemplo de regressão logística. (A) Diagrama de dispersão de uma variável binária y em relação a uma variável contínua x. (B) Exemplo de ajuste de uma curva logística.

A variável parametrizada torna-se, então,

$$\hat{\pi}(x) = S(\beta_0 + \beta_1 x) = \frac{\exp\left[\beta_0 + \beta_1 x\right]}{1 + \exp\left[\beta_0 + \beta_1 x\right]}\,,$$

em que definimos $y = \beta_0 + \beta_1 x$. Essa equação pode ser reescrita como

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x\,, \tag{1.2}$$

em que definimos a função *logit* (logística) [77] de $\pi(x)$ numa forma similar à expressão de um modelo linear simples [78]. A Figura 1.1B ilustra a curva logística ajustada a um conjunto de dados binários gerado artificialmente em função de uma variável contínua x. Na expressão com *logit*, a relação linear pode ser interpretada como o logaritmo da chance, pois a chance é definida como a razão entre as probabilidades de dois eventos,

$$\log(\text{chance}) \coloneqq \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x\,.$$

Ainda, podemos escrever a relação para cada valor de x, ou seja,

$$\sum_{i=1}^{k} y_i = \mathbb{E}[y|x] + \varepsilon_i = k\pi(x) + \varepsilon_i\,, \tag{1.3}$$

em que k é o número total de eventos para um valor específico de x e ε_i é o erro correspondente. Como estamos tratando de eventos de Bernoulli, o erro ε_i pode ser caracterizado

como uma distribuição binomial com valor esperado dado por [79]

$$\begin{aligned}\mathbb{E}[\varepsilon_i] &= \mathbb{E}\left[\sum_{i=1}^{k} y_i\right] - \mathbb{E}[k\pi(x)] \\ &= k\pi(x) - k\pi(x) \\ &= 0\end{aligned} \tag{1.4}$$

e variância dada por

$$\begin{aligned}Var[\varepsilon_i] &= Var\left[\sum_{i=1}^{k} y_i\right] - Var[k\pi(x)] \\ &= k\pi(x)[1-\pi(x)] - 0 \\ &= k\pi(x)[1-\pi(x)]\,.\end{aligned} \tag{1.5}$$

Estimamos os parâmetros do modelo por meio do método da máxima verossimilhança [80]. Considerando um conjunto de dados de uma variável dependente binária y_i e uma variável independente x_i com $i = 1, \ldots, N$, podemos estimar as contribuições de cada par (x_i, y_i) para a verossimilhança por meio de

$$\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}\,. \tag{1.6}$$

Supondo independência entre as observações, a verossimilhança assume a forma do produto das contribuições individuais, ou seja,

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}\,. \tag{1.7}$$

Aplicando o logaritmo à verossimilhança para linearizar a equação, temos

$$\log \mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \{y_i \log \pi(x_i) + (1-y_i)\log[1-\pi(x_i)]\}\,. \tag{1.8}$$

Para encontrar as estimativas dos parâmetros β_0 e β_1, diferenciamos a Equação 1.8 em relação a cada um deles e igualamos as expressões a zero, obtendo

$$\begin{aligned}&\sum_{i=1}^{N}[y_i - \pi(x_i)] = 0\,, \\ &\sum_{i=1}^{N} x_i[y_i - \pi(x_i)] = 0\,.\end{aligned} \tag{1.9}$$

No caso da regressão logística, não é possível encontrar um resultado analítico para essas equações. Dessa forma, recorremos à utilização de rotinas numéricas para estimar β_0 e β_1 conforme implementadas no pacote *statsmodels* [81] do *Python*.

1.2 Regressão linear mista

Dada sua simplicidade, a regressão linear simples é amplamente utilizada para estimar a relação linear entre uma variável dependente y e uma variável independente x, seguindo uma relação do tipo [78]

$$y = \beta_0 + \beta_1 x + \varepsilon\,, \tag{1.10}$$

em que β_0 representa o intercepto constante, β_1 representa o coeficiente linear da influência de x em y e ε representa o erro do modelo. Entretanto, essa formulação apresenta determinadas limitações em função de suas suposições: relação linear, normalidade do erro, homoscedasticidade (variância constante), independência estatística e distribuição idêntica dos dados [78, 82]. Na prática, várias dessas suposições são violadas em dados do mundo real.

Para ilustrar essas limitações, considere um estudo longitudinal em que investigamos o efeito da taxa de forrageamento no peso de ninhos de aves passeriformes [83]. Considerando vários ninhos, os dados são registrados semanalmente e correspondem às quantidades de idas das mães aos alimentadores e aos pesos de seus respectivos ninhos. Notamos que cada ninho pode apresentar um efeito particular entre essas duas variáveis. Esse fato decorre de inúmeras causas – genéticas e ambientais – e tem como consequência o não colapso das curvas de taxa de forrageamento *versus* peso de diferentes ninhos. Nesse caso, há violação da hipótese de independência estatística, pois existe correlação entre os dados de cada ninho, e também da hipótese de distribuição idêntica das variáveis, uma vez que os dados provêm de distribuições com médias distintas. Além disso, determinados ninhos podem apresentar mais observações, sendo o conjunto de dados, nessa situação, denominado desbalanceado. Como efeito, os ninhos com maior quantidade de observações serão mais influentes no resultado da regressão, ocultando a verdadeira relação entre as variáveis estudadas. Para solucionar esses problemas, recorremos à regressão linear mista [84], que considera a estrutura hierárquica dos dados como pressuposto do modelo.

Na regressão linear mista, supomos que os parâmetros também são variáveis aleatórias, cada qual com sua distribuição de probabilidade. Nesse contexto, esses parâmetros são comumente denominados "efeitos aleatórios" [83]. Para entender melhor a estrutura do modelo, vamos considerar outro simples exemplo. Considere que gostaríamos de modelar a progressão dos salários y de funcionários de uma empresa após t anos de sua admissão. Uma possível suposição é que diferentes cargos j possuem diferentes salários iniciais, mas as taxas de crescimento são mais ou menos equivalentes e lineares. Para um grande número de funcionários

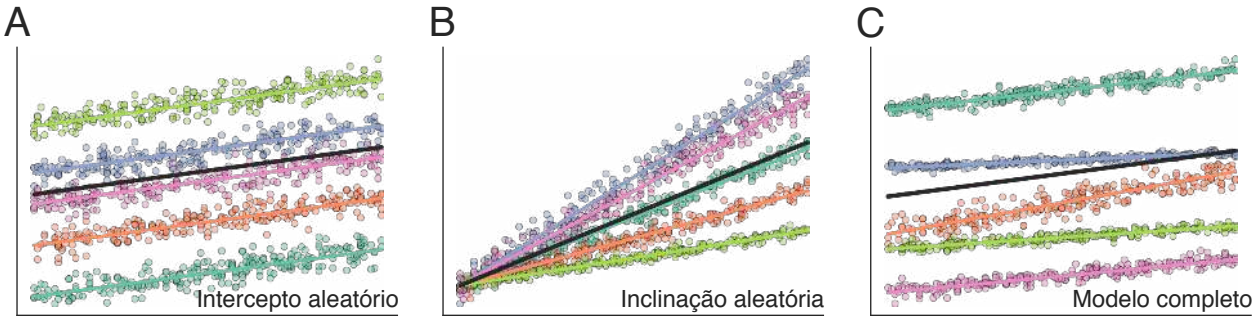

Figura 1.2: Ilustração de modelos lineares mistos. (A) Intercepto aleatório. (B) Inclinação aleatória. (C) Modelo completo com inclinação e intercepto aleatórios. As linhas contínuas em preto representam os valores médios do modelo.

e suas séries temporais, podemos considerar os salários iniciais (interceptos) como sendo normalmente distribuídos com média μ_0 e variância σ_0, isto é,

$$\beta_0 \sim \mathcal{N}(\mu_0, \sigma_0)\,.$$

Podemos escrever a equação do modelo como

$$y_i = \mu_0 + b_{0j} + \beta_1 t_i + \varepsilon_i\,,$$

em que o índice i refere-se à i-ésima observação, o índice j refere-se ao j-ésimo cargo e b_{0j} representa a variação no intercepto do cargo j. De forma mais resumida, temos

$$\begin{aligned} y_i &= \beta_{0j} + \beta_1 x_i + \varepsilon_i\,, \\ \beta_{0j} &= \mu_0 + b_{0j}\,. \end{aligned}$$

A Figura 1.2A ilustra o modelo com interceptos aleatórios. A linha contínua em preto representa o valor médio do modelo, isto é, o comportamento médio global da progressão salarial considerando o salário inicial médio μ_0.

De outra forma, podemos supor que a empresa estipula um salário inicial fixo para todos os funcionários, entretanto, a depender do cargo j, as taxas de crescimento salarial variam. Dessa forma, as inclinações estariam distribuídas normalmente com média μ_1 e variância σ_1, isto é,

$$\beta_1 \sim \mathcal{N}(\mu_1, \sigma_1)\,,$$

sendo as equações que descrevem o modelo dadas por

$$\begin{aligned} y_i &= \beta_0 + \beta_{1j} x_i + \varepsilon_i\,, \\ \beta_{1j} &= \mu_1 + b_{1j}\,, \end{aligned}$$

em que b_{1j} refere-se à variação na inclinação do cargo j. A Figura 1.2B ilustra o modelo com

inclinações aleatórias. A linha contínua em preto representa o valor médio do modelo, isto é, o comportamento médio global da progressão salarial considerando a taxa de crescimento média μ_1.

Finalmente, podemos supor que tanto o salário inicial quanto as taxas de progressão variam entre cargos numa empresa com política salarial distinta das anteriores. O modelo, que aqui chamamos de completo, pode ser descrito pela seguinte equação

$$y_i = \beta_{0j} + \beta_{1j} x_i + \varepsilon_i \,.$$

A Figura 1.2C mostra o modelo com inclinações e interceptos aleatórios. A linha contínua em preto representa o valor médio do modelo, isto é, o comportamento médio global da progressão salarial considerando μ_0 e μ_1. Após inspecionar a Figura 1.2, fica evidente que a regressão linear simples não é capaz de capturar a estrutura hierárquica desse tipo de dados por conta das limitações explicitadas anteriormente. Em contraposição, a regressão linear mista é uma escolha mais adequada nesses casos, pois incorpora a correlação dentro dos grupos e o desbalanceamento dos dados como suposições do modelo.

Estrutura matemática do modelo

Após a introdução qualitativa, vamos detalhar a estrutura matemática da regressão linear mista. Nesta subseção, adotamos a notação utilizada no manual do pacote *lme4* [85]. Definimos o modelo como a distribuição condicional da variável dependente aleatória Y dado que $\mathcal{B} = \boldsymbol{b}$, sendo $\mathcal{B}$ o vetor de efeitos aleatórios, isto é,

$$(Y|\mathcal{B} = \mathbf{b}) \sim \mathcal{N}(\boldsymbol{\mu}_{Y|\mathcal{B}=\mathbf{b}}, \sigma^2 \mathbf{W}^{-1}) \,,$$

com

$$\boldsymbol{\mu}_{Y|\mathcal{B}=\mathbf{b}} = \boldsymbol{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{o} + \boldsymbol{\varepsilon} \,,$$

em que $\boldsymbol{\mu}_{Y|\mathcal{U}=\mathbf{u}}$ é o vetor de preditores lineares, $\mathbf{Z}\mathbf{b}$ é a parcela dos efeitos aleatórios, $\boldsymbol{Z}$ é a matriz *design* para os efeitos aleatórios $\boldsymbol{b}$, $\mathbf{o}$ é o *offset* definido caso haja informações previamente conhecidas sobre o sistema e $\boldsymbol{W}$ é a matriz diagonal de pesos preestabelecidos da variância para modelagem de sua estrutura na regressão.

Supomos que a distribuição dos efeitos aleatórios $\mathcal{B}$ é multivariada e normalmente distribuída com matriz de covariância positiva e semi-definida $\boldsymbol{\Sigma}$,

$$\mathcal{B} \sim \mathcal{N}(0, \boldsymbol{\Sigma}) \,.$$

Com intuito de permitir a singularidade em $\boldsymbol{\Sigma}$ [85], definimos $\boldsymbol{\Sigma}$ em termos de um fator relativo de covariância $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$, cujos parâmetros $\boldsymbol{\theta}$ correspondem aos pesos dos elementos da

matriz covariância dos efeitos aleatórios, isto é,

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \sigma^2 \boldsymbol{\Lambda}_{\boldsymbol{\theta}} \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{\top} .$$

A seguir, supomos que $\mathcal{U}$ é uma variável esférica dos efeitos aleatórios, ou seja,

$$\mathcal{U} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}) ,$$

e realizamos a transformação $\mathcal{B} \to \mathcal{U}$ por meio de

$$\mathcal{B} = \boldsymbol{\Lambda}_{\boldsymbol{\theta}} \mathcal{U} .$$

Assim, a distribuição de $\mathcal{B}$ pode ser descrita como uma função de $\mathcal{U}$. Essa transformação é essencial, pois, quando $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$ é singular e estabelecemos $\mathcal{U}$ em função de $\mathcal{B}$, a distribuição esférica não pode ser estimada [85]. Com essas definições, podemos escrever o modelo como

$$\begin{aligned} (Y|\mathcal{U} = \mathbf{u}) &\sim N(\boldsymbol{\mu}_{Y|\mathcal{U}=\mathbf{u}}, \sigma^2 \mathbf{W}^{-1}) \\ \boldsymbol{\mu}_{Y|\mathcal{U}=\mathbf{u}} &= \boldsymbol{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\Lambda}_{\boldsymbol{\theta}}\boldsymbol{u} + \mathbf{o} + \boldsymbol{\varepsilon} \end{aligned} ,$$

em que $\boldsymbol{\mu}_{Y|\mathcal{U}=\mathbf{u}}$ é a média condicional da variável aleatória esférica $\mathcal{U}$ dado o conjunto de observações do vetor de variáveis dependentes. Na prática, as matrizes do nosso modelo têm a seguinte estrutura

$$\underbrace{\boldsymbol{Y}}_{N\times 1} = \underbrace{\overbrace{\boldsymbol{X}}^{N\times p}\ \overbrace{\boldsymbol{\beta}}^{p\times 1}}_{N\times 1} + \underbrace{\overbrace{\boldsymbol{Z}}^{N\times q}\ \overbrace{\boldsymbol{\Lambda}_{\boldsymbol{\theta}}}^{q\times q}\ \overbrace{\boldsymbol{u}}^{q\times 1}}_{N\times 1} + \underbrace{\boldsymbol{\varepsilon}}_{N\times 1} , \tag{1.11}$$

em que N é o número de observações, p é o número de parâmetros e $q = lp'$ é o número de grupos l vezes o número de parâmetros p' modelados como efeitos aleatórios.

Para analisar a forma de $\boldsymbol{Z}$ e $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$, retomemos o exemplo da progressão salarial em um empresa. Suponha que tenhamos um conjunto de dados estruturados como mostra a Figura 1.3 e que queremos modelar a taxa de crescimento salarial como um efeito aleatório. A

$$\boldsymbol{x} = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\} \qquad N = 10$$
$$\boldsymbol{Y} = \{y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9, y_{10}\} \qquad l = 4$$

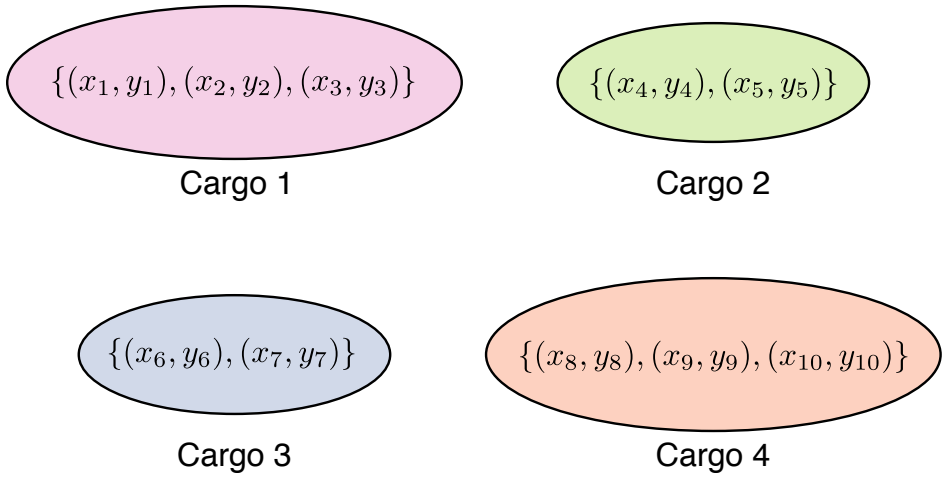

Figura 1.3: Disposição hipotética dos dados de salário por cargo em uma empresa. Nesse caso, o conjunto de dados consiste de $N = 10$ observações (funcionários) que estão distribuídas entre $l = 4$ grupos (cargos).

matriz *design* de efeitos aleatórios $\boldsymbol{Z}$ é escrita como

$$\boldsymbol{Z} = \begin{pmatrix} x_1 & 0 & 0 & 0 \\ x_2 & 0 & 0 & 0 \\ x_3 & 0 & 0 & 0 \\ 0 & x_4 & 0 & 0 \\ 0 & x_5 & 0 & 0 \\ 0 & 0 & x_6 & 0 \\ 0 & 0 & x_7 & 0 \\ 0 & 0 & 0 & x_8 \\ 0 & 0 & 0 & x_9 \\ 0 & 0 & 0 & x_{10} \\ \underbrace{}_{\text{Cargo 1}} & \underbrace{}_{\text{Cargo 2}} & \underbrace{}_{\text{Cargo 3}} & \underbrace{}_{\text{Cargo 4}} \end{pmatrix}.$$

Se adicionarmos a hipótese de que o intercepto (salário inicial) também é um efeito aleatório,

a matriz *design* $\boldsymbol{Z}$ torna-se

$$\boldsymbol{Z} = \begin{pmatrix} x_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ x_2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ x_3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & x_4 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & x_5 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & x_6 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & x_7 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & x_8 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & x_9 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & x_{10} & 1 \\ \underbrace{}_{\text{Cargo 1}} & & \underbrace{}_{\text{Cargo 2}} & & \underbrace{}_{\text{Cargo 3}} & & \underbrace{}_{\text{Cargo 4}} & \end{pmatrix}. \tag{1.12}$$

A matriz de covariância relativa correspondente é dada por

$$\boldsymbol{\Lambda_\theta} = \begin{pmatrix} a & . & . & . & . & . & . & . \\ c & b & . & . & . & . & . & . \\ . & . & a & . & . & . & . & . \\ . & . & c & b & . & . & . & . \\ . & . & . & . & a & . & . & . \\ . & . & . & . & c & b & . & . \\ . & . & . & . & . & . & a & . \\ . & . & . & . & . & . & c & b \\ \underbrace{}_{\text{Cargo 1}} & & \underbrace{}_{\text{Cargo 2}} & & \underbrace{}_{\text{Cargo 3}} & & \underbrace{}_{\text{Cargo 4}} & \end{pmatrix},$$

em que os elementos da diagonal são os parâmetros de variância e os elementos não diagonais são parâmetros de covariância. Assim, para esse modelo hipotético específico, os parâmetros $\boldsymbol{\theta}$ da matriz $\boldsymbol{\Lambda_\theta}$ são

$$\boldsymbol{\theta} = (a, b, c)\,.$$

De forma geral, o número de parâmetros m do vetor $\boldsymbol{\theta}$ pode ser obtido por

$$m = \binom{p+1}{2} = \frac{(p+1)!}{2!(p-1)!}\,.$$

1.3 Modelos hierárquicos bayesianos

Diferentemente do tratamento dado à regressão logística, utilizamos uma abordagem bayesiana para estimar os parâmetros da regressão linear mista. Naquele contexto, empre-

gamos o método frequentista de estimação dos parâmetros por máxima verossimilhança. Neste contexto, empregamos o Teorema de Bayes para estimar uma distribuição de probabilidade dos parâmetros. Considerando um conjunto de dados D e parâmetros $\boldsymbol{\theta}$, o Teorema de Bayes pode ser escrito como [86]

$$P(\boldsymbol{\theta}|D) = \frac{P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(D)}\,, \tag{1.13}$$

em que $P(D|\boldsymbol{\theta})$ é a verossimilhança (distribuição de probabilidade da amostra D supondo que o modelo para os parâmetros $\boldsymbol{\theta}$ é o correto), $P(\boldsymbol{\theta})$ é a distribuição a *priori* (distribuição de probabilidade dos parâmetros do modelo antes de termos qualquer informação via dados), $P(D)$ é a probabilidade de obtermos uma determinada amostra sob qualquer hipótese (também pode ser interpretado como um fator de normalização da distribuição a *posteriori*) e $P(\boldsymbol{\theta}|D)$ é a distribuição a *posteriori* (distribuição de probabilidade dos parâmetros $\boldsymbol{\theta}$ após atualizarmos a distribuição a *priori* por meio das informações do conjunto de dados D).

Inicialmente, precisamos definir a distribuição de probabilidade a *posteriori* do modelo hierárquico. Começamos considerando um sistema de dois níveis que possui uma estrutura hierárquica genérica com l grupos. O número de observações N é dado pela soma dos elementos respectivos de cada grupo l (n_j), isto é, $N = \sum_{j=1}^{l} n_j$. A verossimilhança da i-ésima observação e j-ésimo grupo pode ser escrita como

$$\boldsymbol{Y}_{ij}|\boldsymbol{\theta}_j \sim P(y_{ij}|\boldsymbol{\theta}_j)\,. \tag{1.14}$$

Se as observações de cada grupo são independentes entre si, é possível escrever a verossimilhança do j-ésimo grupo como o produto das verossimilhanças individuais [87], isto é,

$$P(\boldsymbol{y}_j|\boldsymbol{\theta}_j) = \prod_{i=1}^{n_j} P(y_{ij}|\boldsymbol{\theta}_j)\,. \tag{1.15}$$

Além disso, sendo os parâmetros específicos de cada j-ésimo grupo independentes, podemos escrever a distribuição a *priori* como [87]

$$P(\boldsymbol{\theta}|\boldsymbol{\phi}) = \prod_{j=1}^{l} P(\boldsymbol{\theta}_j|\boldsymbol{\phi})\,, \tag{1.16}$$

em que $\boldsymbol{\phi}$ são os hiperparâmetros do modelo. Em termos bayesianos, os hiperparâmetros são os "parâmetros dos parâmetros" e suas distribuições a *priori* correspondentes são chamadas distribuições a *hiperpriori* [88]. No modelo hierárquico, supomos que os parâmetros $\boldsymbol{\beta}$ provêm da distribuição a *hiperpriori* $P(\boldsymbol{\phi})$ dos hiperparâmetros $\boldsymbol{\phi}$. Como os parâmetros de cada grupo estão correlacionados via distribuição dos hiperparâmetros, grupos com pouca quantidade de dados conseguem "emprestar força estatística" dos grupos com maior

quantidade de dados [89].

Escolhemos as distribuições a *priori* e a *hiperpriori* adequadas e, a partir do Teorema de Bayes, conseguimos definir a distribuição hierárquica para o modelo hierárquico de dois níveis como [87]

$$\begin{aligned} P(\boldsymbol{\theta}, \boldsymbol{\phi}|D) &= \frac{P(D|\boldsymbol{\theta})P(\boldsymbol{\theta}|\boldsymbol{\phi})P(\boldsymbol{\phi})}{P(D)} \\ &= \frac{P(\boldsymbol{\phi})\prod_{j=1}^{l} P(\boldsymbol{\theta}_j|\boldsymbol{\phi})P(\boldsymbol{y}_j|\boldsymbol{\theta}_j)}{P(D)} \\ &\propto P(\boldsymbol{\phi})\prod_{j=1}^{l} P(\boldsymbol{\theta}_j|\boldsymbol{\phi})P(\boldsymbol{y}_j|\boldsymbol{\theta}_j)\,. \end{aligned} \tag{1.17}$$

1.4 Amostrador No-U-Turn

Em problemas reais, o cálculo da distribuição a *posteriori* dificilmente é realizado de forma analítica por meio da Equação 1.13. A razão da dificuldade está no cálculo do denominador da expressão, isto é, no cálculo da integral

$$P(D) = \int P(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}\,.$$

Sua solução analítica é trabalhosa e frequentemente inviável, sendo possível apenas nos casos mais simples. A solução numérica, por sua vez, funciona somente para um número reduzido de dimensões em tempo computacional razoável. Esse não é o caso para maior parte dos problemas reais. A título de exemplo, considere um modelo linear misto com intercepto β_0 e inclinação β_1 aleatórios considerando a estrutura da Figura 1.3. Os parâmetros do modelo são $\boldsymbol{\theta} = \{\beta_{0j}, \beta_{1j}\}$ com $j = 1, \ldots, 4$ e os hiperparâmetros são $\boldsymbol{\Phi} = \{\mu_0, \sigma_0, \mu_1, \sigma_1\}$. A integral pode ser escrita como

$$P(D) = \int_{\boldsymbol{\theta},\boldsymbol{\Phi}} P(\boldsymbol{\phi})P(\boldsymbol{\theta}|\boldsymbol{\phi})P(D|\boldsymbol{\theta})d\boldsymbol{\theta}d\boldsymbol{\Phi}\,.$$

Nesse simples caso, a integral já apresenta doze dimensões (8 parâmetros e 4 hiperparâmetros). Além disso, o cálculo de certas estatísticas como a média,

$$\mathbb{E}[\boldsymbol{\theta}|D] = \int_{\text{todo } \boldsymbol{\theta}} \boldsymbol{\theta}P(\boldsymbol{\theta}|D)d\boldsymbol{\theta} \tag{1.18}$$

e a variância

$$\begin{aligned} Var[\boldsymbol{\theta}|D] &= \mathbb{E}[\boldsymbol{\theta}^2|D] - (\mathbb{E}[\boldsymbol{\theta}|D])^2 \\ &= \int \boldsymbol{\theta}^2 P(\boldsymbol{\theta}|D)d\boldsymbol{\theta} - \left[\int \boldsymbol{\theta}P(\boldsymbol{\theta}|D)d\boldsymbol{\theta}\right]^2, \end{aligned} \tag{1.19}$$

também envolve o cálculo de integrais do mesmo tipo. Na prática, porém, precisamos apenas estimar o numerador do Teorema de Bayes, pois o denominador atua apenas como um fator de normalização. No caso mais simples, precisamos estimar

$$\begin{aligned} P(\boldsymbol{\theta}|D) &= \frac{P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(D)} \\ &\propto P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})\,, \end{aligned} \tag{1.20}$$

ou, num modelo hierárquico, por meio da relação expressa na Equação 1.17. Portanto, uma possível solução é replicar a distribuição a *posteriori* por meio de métodos de amostragem que não utilizam informações sobre a estrutura global da distribuição (que é muito complexa), mas que focam em passos locais correlacionados. Essas técnicas pertencem à classe de métodos de amostragem estocástica via *Markov Chain Monte Carlo* (MCMC) [90]. A vantagem dos métodos MCMC é que não precisamos saber antecipadamente a forma da distribuição a *posteriori*. A amostragem da distribuição a *posteriori* é realizada construindo cadeias de Markov que resultam na distribuição de equilíbrio alvo, a distribuição a *posteriori*, por meio de caminhantes aleatórios que exploram o espaço de parâmetros. A dinâmica do processo estocástico é regida pela escolha de algoritmos que privilegiam a exploração de regiões que contribuem mais para a distribuição alvo. Podemos destacar como algoritmos MCMC mais simples o algoritmo de Metropolis [91] e o algoritmo de Gibbs [92]. Entretanto, esses algoritmos apresentam dificuldades ao tratar de modelos multidimensionais. A principal dificuldade decorre do fenômeno conhecido como "concentração de medida" [93]. Esse fenômeno consiste na discrepância entre o hipervolume da distribuição alvo, que se torna cada vez mais singular quanto maior o número de dimensões, e o hipervolume de seus arredores como mostra a Figura 1.4. Dessa forma, o caminhante aleatório não consegue explorar todo o espaço dos parâmetros. Por esse motivo, neste trabalho, optamos por utilizar

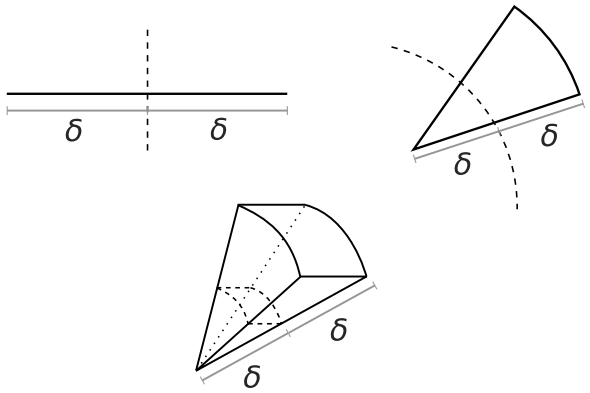

Figura 1.4: Efeito da concentração de medida. O aumento da dimensionalidade da distribuição estudada torna o volume externo à região de interesse drasticamente maior do que o volume da própria região.

o amostrador de Monte Carlo Hamiltoniano (HMC), que apresentaremos a seguir.

O amostrador HMC consiste em uma analogia física para propor os passos dos caminhantes aleatórios em que os parâmetros da distribuição a *posteriori* são considerados como a posição de um sistema físico de mecânica clássica [94, 95]. Por meio das equações de Hamilton, calculamos a trajetória do sistema de modo determinístico para explorar o espaço da *posteriori* segundo sua geometria. Primeiramente, consideramos a distribuição de probabilidade conjunta

$$\pi(\boldsymbol{q}, \boldsymbol{p}) = \pi(\boldsymbol{p}|\boldsymbol{q})\pi(\boldsymbol{q})\,, \tag{1.21}$$

em que $\boldsymbol{q}$ são os parâmetros da *posteriori* (a posição do sistema físico), $\boldsymbol{p}$ são as variáveis auxiliares representando as coordenadas de momento com mesma dimensão da posição, $\pi(\boldsymbol{p}|\boldsymbol{q})$ é a distribuição do momento condicionada à posição e $\pi(\boldsymbol{q})$ é a distribuição a *posteriori*. A partir disso, podemos definir o hamiltoniano como

$$\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}) = -\log \pi(\boldsymbol{q}, \boldsymbol{p})\,, \tag{1.22}$$

cuja interpretação é de um *ensemble* canônico com probabilidade definida como

$$\begin{aligned} P &\propto e^{-\mathcal{H}(\boldsymbol{q},\boldsymbol{p})} \\ P &\propto \pi(\boldsymbol{q}, \boldsymbol{p})\,. \end{aligned} \tag{1.23}$$

Na equação acima, cada termo do hamiltoniano corresponde a certa parcela da energia total do sistema, isto é,

$$\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}) = K(\boldsymbol{p}, \boldsymbol{q}) + V(\boldsymbol{q})\,, \tag{1.24}$$

em que $K(\boldsymbol{p}, \boldsymbol{q}) = -\log \pi(\boldsymbol{p}|\boldsymbol{q})$ é a energia cinética e $V(\boldsymbol{q}) = -\log \pi(\boldsymbol{q})$ é a energia potencial. A energia potencial é definida como o logaritmo da distribuição a *posteriori*. A energia cinética pode ser escolhida de maneira mais conveniente para cada modelo estudado [93]. Se consideramos um sistema na ausência de atrito, a energia total é constante e as trajetórias ficam confinadas a um nível energético determinado, ou seja,

$$\mathcal{H}^{-1}(E) = \{\boldsymbol{q}, \boldsymbol{p}|\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}) = E\}\,, \tag{1.25}$$

hipersuperfícies em $(2D-1)$ dimensões do espaço de parâmetros com dimensão D. Podemos também decompor a distribuição canônica em termos microcanônicos, isto é,

$$\pi(\boldsymbol{q}, \boldsymbol{p}) = \pi(q_E|E)\pi(E)\,, \tag{1.26}$$

em que $\pi(q_E|E)$ é a distribuição microcanônica e $\pi(E)$ é a distribuição marginal de energias.

De modo geral, podemos dizer que o algoritmo HMC consiste da repetição de duas etapas: *i)* o cálculo determinístico das trajetórias no espaço de parâmetros mantendo a energia

fixa (considerando o sistema sem atrito) e *ii)* a exploração estocástica dos níveis de energia, representados pela distribuição marginal de energias na Equação 1.26, pelo sorteio das coordenadas de momento. Podemos visualizar essas duas etapas na Figura 1.5A. As elipses concêntricas representam os níveis energéticos $\mathcal{H}$ e as respectivas coordenadas permitidas. As curvas coloridas em cinza mostram a trajetória no espaço de fase (a primeira etapa). As linhas verticais e o marcador indicam a posição final que é armazenada para construção da distribuição $\pi(\boldsymbol{q})$. Após o sorteio estocástico do momento, as trajetórias recomeçam em outro nível energético (a segunda etapa).

Nesse contexto, a escolha da energia cinética é realizada de modo a favorecer a exploração dos níveis energéticos com eficiência. Em outras palavras, a escolha da energia cinética deve fazer com que a distribuição de transições energéticas convirja para a distribuição marginal de níveis de energia $\pi(E)$ da Equação 1.26, sendo esta representativa de todas as energias relevantes ao sistema. Neste caso ideal, a amostragem é realizada de maneira independente [93]. Comumente, a escolha da forma da energia cinética restringe-se a dois tipos: Gaussiana-Euclidiana e Gaussiana-Riemanniana [93].

A diferença dessas energias cinéticas reside na métrica empregada em sua construção. Por exemplo, a energia cinética Gaussiana-Euclidiana faz uso da métrica euclidiana para construir energias do tipo

$$K(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{2}\boldsymbol{p}^{\top}\boldsymbol{M}^{-1}\boldsymbol{p} + \log|\boldsymbol{M}| + \text{constante}\,, \tag{1.27}$$

em que $\boldsymbol{M}$ é a matriz de massa responsável por realizar transformações lineares na *posteriori* [96]. Os elementos de variância esticam ou comprimem os parâmetros da *posteriori* para que eles apresentem a mesma escala, enquanto os elementos de covariância rotacionam

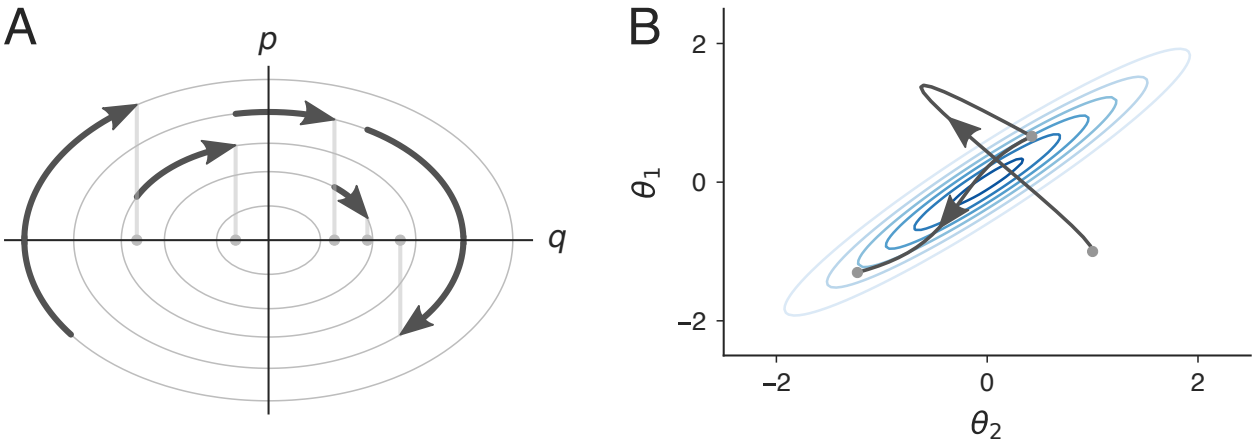

Figura 1.5: Amostrador de Monte Carlo Hamiltoniano. (A) Espaço de fases e transições energéticas do algoritmo HMC. As elipses concêntricas ilustram os níveis energéticos e as coordenadas permitidas. As curvas em cinza representam as trajetórias no espaço de fase. As linhas verticais indicam as posições finais. (B) Trajetória do caminhante aleatório no espaço de parâmetros de acordo com o algoritmo HMC.

a *posteriori* para que os parâmetros sejam considerados independentes entre si. Se a matriz de massa é similar à matriz de covariância da *posteriori*, a amostragem pode ser considerada independente. A dificuldade é que raramente sabemos qual a verdadeira forma da matriz de covariância da *posteriori*.

Outra possível definição de energia cinética é por meio da métrica Gaussiana-Riemanniana na matriz de massa, isto é,

$$K(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{2}\boldsymbol{p}^\top [\boldsymbol{\Sigma}(\boldsymbol{q})]^{-1}\boldsymbol{p} + \log|\boldsymbol{\Sigma}(\boldsymbol{q})| + \text{constante}\,, \tag{1.28}$$

em que $\boldsymbol{M} = \boldsymbol{\Sigma}(\boldsymbol{q})$ é a matriz de massa dependente da posição $\boldsymbol{q}$. Nessa abordagem, consideramos que tanto a matriz de massa quanto a métrica dependem da posição no espaço [93], o que aumenta a eficiência em regiões de alta curvatura espacial.

Vamos exemplificar o funcionamento do amostrador HMC por meio da energia gaussiana mais simples, expressa por

$$K(\boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^\top\boldsymbol{p} + \text{constante}\,, \tag{1.29}$$

com distribuição gaussiana do momento $\boldsymbol{p} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. As equações de Hamilton podem ser escritas como

$$\begin{aligned} \frac{d\boldsymbol{q}}{dt} &= \frac{\partial\mathcal{H}(\boldsymbol{p},\boldsymbol{q})}{\partial\boldsymbol{p}} = \frac{\partial K(\boldsymbol{p})}{\partial\boldsymbol{p}} + \frac{\partial V(\boldsymbol{q})}{\partial\boldsymbol{p}}\,, \\ \frac{d\boldsymbol{p}}{dt} &= -\frac{\partial\mathcal{H}(\boldsymbol{p},\boldsymbol{q})}{\partial\boldsymbol{q}} = -\frac{\partial K(\boldsymbol{p})}{\partial\boldsymbol{q}} - \frac{\partial V(\boldsymbol{q})}{\partial\boldsymbol{q}}\,. \end{aligned} \tag{1.30}$$

Após realizar as simplificações, temos que

$$\begin{aligned} \frac{d\boldsymbol{q}}{dt} &= \boldsymbol{p}\,, \\ \frac{d\boldsymbol{p}}{dt} &= -\frac{\partial V(\boldsymbol{q})}{\partial\boldsymbol{q}}\,. \end{aligned} \tag{1.31}$$

Em posse das equações de Hamilton, podemos agora definir o procedimento de amostragem:

1. Amostragem do momento $\boldsymbol{p} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$;

2. Simulação das trajetórias no espaço de parâmetros, $\boldsymbol{q}(t)$ e $\boldsymbol{p}(t)$, por meio das equações de Hamilton em T passos temporais;

3. Armazenamento da posição final $\boldsymbol{q}(T)$.

A amostragem do momento e o armazenamento da posição final são os passos mais simples do algoritmo. Em contraste, a simulação da trajetória depende da resolução das equações de Hamilton, que não é possível de maneira analítica a não ser nos exemplos mais

simples [93]. Recorremos a métodos numéricos para discretizar as equações de Hamilton. Consideramos o tamanho do passo ϵ e o número total de passos L (a duração da trajetória) como parâmetros. O método numérico muito empregado para resolução das equações de Hamilton é o algoritmo *leapfrog* [97] descrito pelo conjunto de equações

$$\begin{aligned} \boldsymbol{p}_{t+\epsilon/2} &= \boldsymbol{p}_t + (\epsilon/2)\nabla_{\boldsymbol{q}} V(\boldsymbol{q}_t)\,, \\ \boldsymbol{q}_{t+\epsilon} &= \boldsymbol{q}_t + \epsilon \boldsymbol{p}_{t+\epsilon/2}\,, \\ \boldsymbol{p}_{t+\epsilon} &= \boldsymbol{p}_{t+\epsilon/2} + (\epsilon/2)\nabla_{\boldsymbol{q}} V(\boldsymbol{q}_{t+\epsilon})\,, \end{aligned} \tag{1.32}$$

em que o índice indica o número de iterações do algoritmo e ∇_q é o diferencial espacial, ou seja,

$$\nabla_{\boldsymbol{q}} V(\boldsymbol{q}_t) \to \frac{\partial V(q_{t,i})}{\partial q_i}\,, \tag{1.33}$$

em que o índice i indica a i-ésima coordenada.

O algoritmo *leapfrog* atualiza a posição no espaço de fase utilizando as próprias coordenadas. Essa característica é importante pois garante que passos sucessivos preservam o hipervolume e respeitam o princípio de balanço detalhado na cadeia de Markov [98]. Após realizar a trajetória por L passos, a trajetória é realizada com probabilidade

$$\alpha = \min\left(1, \frac{\exp V(\tilde{\boldsymbol{q}}) - \frac{1}{2}\tilde{\boldsymbol{p}}\cdot\tilde{\boldsymbol{p}}}{\exp V(\boldsymbol{q}_0) - \frac{1}{2}\boldsymbol{p}_0\cdot\boldsymbol{p}_0}\right)\,, \tag{1.34}$$

em que $\tilde{\boldsymbol{p}}$ é o momento da última iteração, $\tilde{\boldsymbol{q}}$ é a posição da última iteração, $\boldsymbol{p}_0$ é o momento sorteado no início e $\boldsymbol{q}_0$ é a posição inicial. Qualitativamente, a razão de probabilidades na Equação 1.34 indica a energia perdida de 0 a $T = \epsilon L$. A desvantagem desse tipo de algoritmo é que existe uma diferença entre a trajetória real e a trajetória calculada por se tratar de uma aproximação discreta [93]. Se pudéssemos calcular exatamente a trajetória, sempre obteríamos $\alpha = 1$ e as proposições seriam sempre aceitas. O Algoritmo 1 descreve o código de uma implementação do HMC [98]. Para que haja reversibilidade temporal e respeito do balanço detalhado, trajetórias aceitas têm coordenadas de momento armazenadas com sinal trocado [93].

Em amostradores mais simples, como o amostrador de Metropolis e Gibbs, as proposições dos passos localizam-se aleatoriamente ao redor da posição e, em decorrência disso, passos subsequentes são altamente correlacionados. Esse fato pode gerar dificuldades para o caminhante alcançar regiões distantes de sua posição inicial dependendo da geometria da distribuição a *posteriori* [96]. Para o amostrador HMC, a correlação entre os passos ainda existe, mas num grau muito menor se comparado aos amostradores mencionados anteriormente, pois o método utiliza informações sobre a geometria da distribuição a *posteriori* para

Algoritmo 1 Amostrador de Monte Carlo Hamiltoniano

Inicialização das variáveis: $\boldsymbol{q}_0$, ϵ, L
for $i = 1, 2, \ldots$ **do**
    Amostrar o momento: $\boldsymbol{p}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
    Definir: $\tilde{\boldsymbol{q}} \leftarrow \boldsymbol{q}_{i-1}, \tilde{\boldsymbol{p}} \leftarrow \boldsymbol{p}_0$
    for $j = 1$ to L **do**
        Definir: $\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}} \leftarrow \text{Leapfrog}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}}, \epsilon)$
    Definir a probabilidade de aceitação: $\alpha = \min\left(1, \dfrac{\exp V(\tilde{\boldsymbol{q}}) - \frac{1}{2}\tilde{\boldsymbol{p}} \cdot \tilde{\boldsymbol{p}}}{\exp V(\boldsymbol{q}_{i-1}) - \frac{1}{2}\boldsymbol{p}_0 \cdot \boldsymbol{p}_0}\right)$
    $u \sim \mathcal{U}(0, 1)$
    if $u < \alpha$ **then**
        Aceitar a proposição: $\boldsymbol{q}_i \leftarrow \tilde{\boldsymbol{q}}, \boldsymbol{p}_i \leftarrow -\tilde{\boldsymbol{p}}$
    else
        Rejeitar a proposição: $\boldsymbol{q}_i \leftarrow \boldsymbol{q}_{i-1}, \boldsymbol{p}_i \leftarrow \boldsymbol{p}_{i-1}$
function $\text{Leapfrog}(\boldsymbol{q}, \boldsymbol{p}, \epsilon)$
Definir: $\tilde{\boldsymbol{p}} \leftarrow \boldsymbol{p} + (\epsilon/2)\nabla_{\boldsymbol{q}} V(\boldsymbol{q})$
Definir: $\tilde{\boldsymbol{q}} \leftarrow \boldsymbol{q} + \epsilon\tilde{\boldsymbol{p}}$
Definir: $\tilde{\boldsymbol{p}} \leftarrow \tilde{\boldsymbol{p}} + (\epsilon/2)\nabla_{\boldsymbol{q}} V(\tilde{\boldsymbol{q}})$
return $\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}}$

propor os passos. Como exemplo disso, a Figura 1.5B mostra duas iterações do Algoritmo 1 em que o caminhante se desloca para regiões distintas do espaço da *posteriori*.

É importante ressaltar que a escolha de valores adequados para os parâmetros ϵ (tamanho do passo) e L (número total de passos) é essencial a fim de que o algoritmo HMC atinja uma performance satisfatória, pois o algoritmo é muito sensível à variação desses parâmetros [98]. No caso de trajetórias curtas, o comportamento é o mesmo de um caminhante aleatório, que avança apenas para regiões circundantes à sua posição. No caso de trajetórias longas, por outro lado, o caminhante, num comportamento cíclico, pode acabar visitando as mesmas regiões desnecessariamente. Por isso, incorporamos ao amostrador HMC um procedimento que cessa a progressão das trajetórias na ocorrência de uma meia volta (no inglês, *U-Turn*) para escolher o tamanho ideal da trajetória. Esse procedimento dá um novo nome ao amostrador HMC de *NUTS* (*No U-Turn Sampler*) [98].

No amostrador NUTS, construímos a trajetória a partir de seu prolongamento em direções aleatoriamente determinadas. No primeiro passo, a trajetória é calculada com dois passos para frente. Em passos subsequentes, uma direção aleatória é escolhida por meio do sorteio $u \sim \mathcal{U}(\{-1, 1\})$. O tamanho do deslocamento é sempre o dobro do tamanho na iteração anterior[1] para construirmos trajetórias de tamanhos variados. O critério de parada é a ocorrência de uma meia volta. Para definir esse acontecimento matematicamente, considere que $\boldsymbol{q}_-(t)$ e $\boldsymbol{q}_+(t)$ são as posições das extremidades da trajetória no tempo t e $\boldsymbol{p}_\pm(t)$ são suas

[1] O número de dobras também é chamado de "comprimento da árvore".

respectivas coordenadas de momento. O critério de parada do algoritmo, para uma métrica euclidiana, pode ser definido como [98]

$$\begin{aligned} & \boldsymbol{p}_+(t)^\top \cdot [\boldsymbol{q}_+(t) - \boldsymbol{q}_-(t)] < 0 \\ \text{e} \quad & \boldsymbol{p}_-(t)^\top \cdot [\boldsymbol{q}_-(t) - \boldsymbol{q}_+(t)] < 0\,, \end{aligned} \tag{1.35}$$

em que se define que os momentos das extremidades estão alinhados de maneira contrária à linha que une suas posições [93]. Definimos também um critério de parada adicional em que os valores de energia total $\mathcal{H} \to \infty$, ou seja, quando ocorre a divergência do erro de aproximação. Essa situação denomina-se *transição divergente* e acarreta baixas taxas de aceitação [96]. Após a parada, a amostra da posição é sorteada a partir de todas as posições por que o caminhante passou até a última iteração. Em outras palavras, sorteamos o tamanho do passo L utilizando apenas posições da trajetória em que o caminhante ainda não realizou uma meia-volta. Como não há probabilidade de aceitação no NUTS, empregamos a taxa de aceitação média do algoritmo hamiltoniano tradicional na última dobra [98].

O tamanho do passo ϵ é outro parâmetro que pode afetar o desempenho do algoritmo NUTS. Passos de tamanho grande contribuem com o aumento do erro de aproximação, o que pode causar uma divergência no valor da energia total ($\mathcal{H} \to \infty$). Por outro lado, passos de tamanho pequeno fazem com que capacidade de processamento computacional seja desperdiçada no cálculo de trajetórias demasiadamente detalhadas. Dessa forma, precisamos encontrar um tamanho de passo ϵ ideal. Para isso, utilizamos um método adaptativo. Seja a estatística G_t definida como

$$G_t = \delta - \alpha_t\,, \tag{1.36}$$

em que δ é a probabilidade de aceitação desejada e α_t é a probabilidade de aceitação no tempo t. O valor esperado de G_t é dado por

$$\mathbb{E}_t[G_t|\epsilon] = g(\epsilon) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[G_t|\epsilon]\,. \tag{1.37}$$

Ainda, consideramos uma função $g(\epsilon)$ não decrescente e atualizações do tipo

$$\begin{aligned} \epsilon_{t+1} &\leftarrow \mu - \frac{\sqrt{t}}{\gamma} \frac{1}{t+t_0} \sum_{i=0}^{t} G_i\,, \\ \bar{\epsilon}_{t+1} &\leftarrow \nu_t \epsilon_{t+1} + \bar{\epsilon}_t - \nu_t \bar{\epsilon}_t \end{aligned} \tag{1.38}$$

em que μ é o valor de convergência escolhido para ϵ_t, t_0 é um valor que estabiliza as primeiras iterações, ν_t é o tamanho do passo em cada iteração, $\gamma > 0$ define a intensidade de concentração para μ, a somatória refere-se ao valor médio de G_t até o tempo t e definimos $\bar{\epsilon}_1 = \epsilon_1$. Desejamos que $g(\epsilon) \to 0$, isto é, obter a taxa de aceitação desejada $\delta \approx \alpha_t$. Da

literatura, temos que esses resultados são alcançados quando $\sum_t \nu_t \to \infty$ e $\sum_t \nu_t^2 < \infty$ [98]. Uma possível escolha da função ν_t é dada por $\nu_t = t^{-\kappa}$ com $\kappa \in (0.5, 1]$. Para mais detalhes sobre esse processo adaptativo, recomendamos as referências [98–100].

Definimos, assim, uma maneira de encontrar os parâmetros L (tamanho da trajetória) e ϵ (tamanho do passo) ótimos para que o amostrador hamiltoniano funcione mais efetiva e automatizadamente. Para fins de comparação, as Figuras 1.6A-C apresentam a performance de três amostradores (respectivamente, Metropolis, Gibbs e HMC) em suas 100 primeiras iterações para uma distribuição alvo bidimensional definida por

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\sigma} = \begin{bmatrix} 1 & 0.95 \\ 0.95 & 1 \end{bmatrix} \right),$$

com alto grau de correlação entre as variáveis θ_1 e θ_2, começando do ponto inicial $(-2, 2)$. A Figura 1.6D simula uma amostragem independente da mesma distribuição. Notamos que o algoritmo de Metropolis amostra menos pontos do que o restante por rejeitar mais pro-

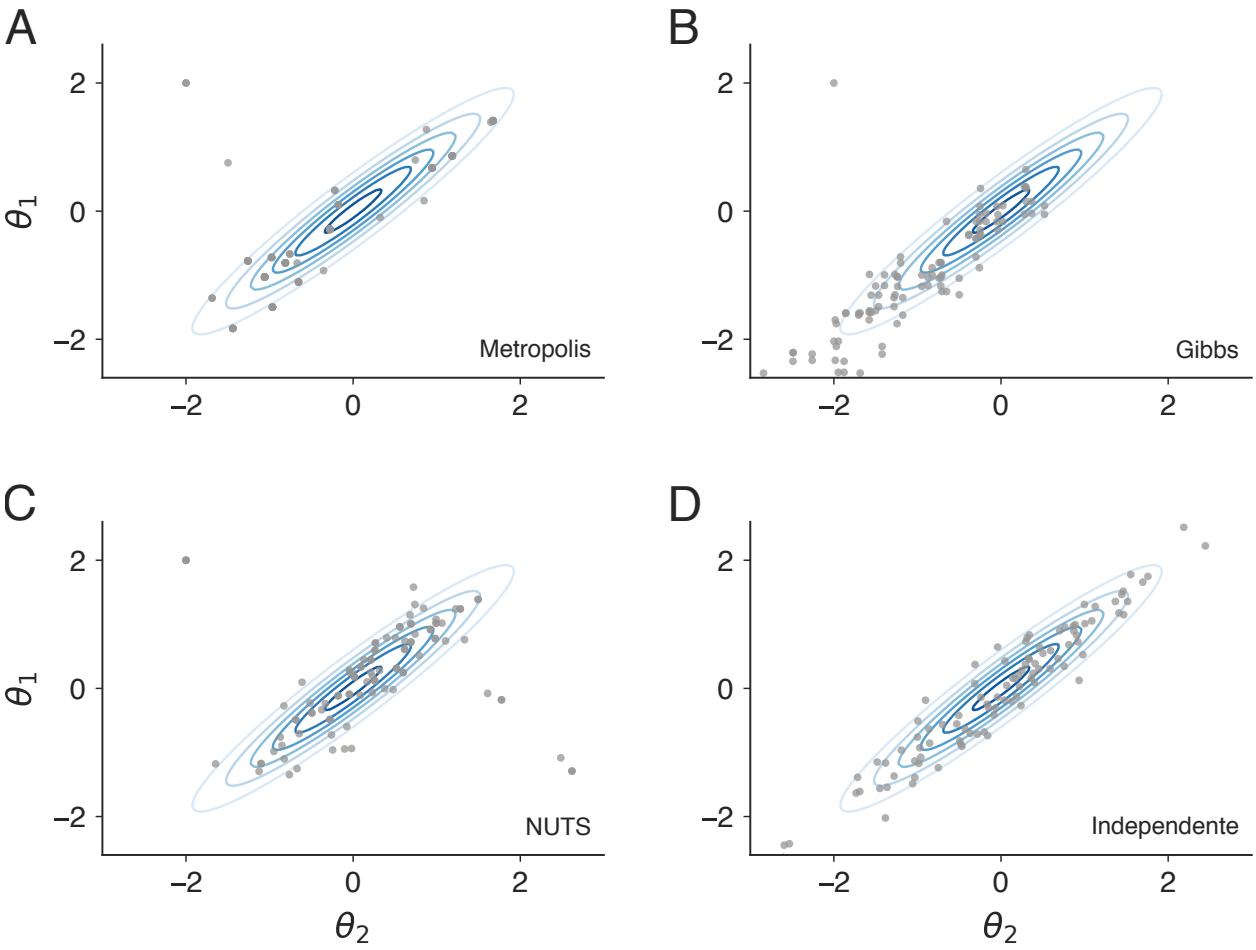

Figura 1.6: Tipos de amostradores e sua performance. (A) Amostragem de Metropolis. (B) Amostragem de Gibbs. (C) Amostragem NUTS. (D) Amostragem independente. Os marcadores em cinza representam os pontos amostrados.

posições (Figura 1.6A). O algoritmo de Gibbs, por sua vez, não apresenta o problema de alta taxa de rejeição, porém, tem dificuldades em explorar as regiões da distribuição alvo uniformemente (Figura 1.6B). O algoritmo HMC (Figura 1.6C) é o que mais se aproxima de uma amostragem independente (Figura 1.6D), pois realiza a amostragem por meio de uma analogia física que evita trajetórias redundantes, conseguindo explorar o espaço dos parâmetros mais efetivamente. É importante ressaltar que, para geometrias simples, todos esses algoritmos têm boa performance para um número razoável de iterações, mas, para geometrias mais complicadas, a variante NUTS do algoritmo HMC é a escolha mais apropriada[2].

Convergência da cadeia de Markov

Como optamos por amostrar a distribuição a *posteriori*, não sabemos qual sua verdadeira forma nem quantas iterações são necessárias para representá-la precisamente. Precisamos de métricas para quantificar a convergência da distribuição. Com esse intuito, rodamos múltiplas cadeias de Markov e verificamos a convergência das distribuições correspondentes. A concordância entre as distribuições é um bom indicativo de que a distribuição amostrada representa a distribuição verdadeira [88]. Considerando um sistema unidimensional do parâmetro θ, a primeira métrica que pode nos auxiliar nessa tarefa é a variância de uma única cadeia dada por

$$W = \frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} (\theta_{ij} - \overline{\theta}_j)^2 \,, \tag{1.39}$$

em que m é o número total de cadeias e n é o número total de iterações para cada cadeia. O índice j refere-se à j-ésima cadeia, o índice i refere-se à i-ésima observação e $\overline{\theta}_j$ é o valor médio do parâmetro na cadeia j. Outra métrica que pode ser definida é a variância entre cadeias definida como

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\overline{\theta}_j - \overline{\theta})^2 \,,$$

em que $\overline{\theta}$ é o valor médio do parâmetro considerando todas as cadeias. Se as variâncias entre cadeias e de cada cadeia tiverem valores próximos, temos indícios de que as cadeiras estão bem "misturadas", isto é, alcançaram o estado de equilíbrio que muito possivelmente reflete uma distribuição a *posteriori* bem estimada. Com esse intuito, Gelman e Rubin propuseram a métrica de variância da *posteriori*, similar ao método empregado na modelagem ANOVA, escrita como [101]

$$\begin{aligned} \mathrm{Var}(\hat{\theta}|D) &= \frac{n-1}{n} W + \frac{1}{n} B \\ &= W + \frac{1}{n}(B - W) \,, \end{aligned} \tag{1.40}$$

[2]Para uma comparação entre as diversas técnicas de amostragem, recomendamos o aplicativo do link https://chi-feng.github.io/mcmc-demo/app.html. [Último acesso em 14 de agosto de 2023]

uma variância superestimada, mas não enviesada no estado estacionário das cadeias [101]. No limite em que $B \to W$ ou $n \to \infty$, a variância da distribuição a *posteriori* converge exatamente para a variância interna $var(\hat{\theta}|D) \to W$, traduzindo a ideia de mistura das cadeias. Outra métrica proposta por Gelman e Rubin é o "R chapéu", definido como [88,101]

$$\hat{R} = \sqrt{\frac{W + \frac{1}{n}(B - W)}{W}}, \tag{1.41}$$

cuja interpretação é a razão entre a variância estimada e a variância desejada W. Nas iterações iniciais, como a variância entre as cadeias é maior do que nas cadeias ($B \gg W$) temos $\hat{R} \gg 1$. Porém, com a progressão do processo de amostragem, a tendência é que as variâncias convirjam para o mesmo valor ($B \to W$). Nessa situação, a estatística converge para um ($\hat{R} \to 1$). Na prática, é comum considerar o valor $\hat{R} \approx 1.1$ como indício de boa mistura das cadeias [88].

Outra ferramenta útil para avaliar a mistura das múltiplas cadeias de Markov é a visualização denominada *trace plot* [88,102]. O *trace plot* ilustra a evolução temporal (de iterações) das séries da estimativa do parâmetro amostrado para todas as cadeias. Cadeias com boa mistura apresentam *trace plot* com curvas flutuando ao redor de um único valor como no exemplo apresentado na Figura 1.7.

A utilização de métodos de amostragem via cadeias de Markov implica autocorrelação entre passos sucessivos. Dessa forma, podemos dizer que existe uma quantidade efetiva de passos que pode ser estimada por [87]

$$n_{ef} = \frac{mT}{1 + 2\sum_{\tau=1}^{\infty} \rho_\tau}, \tag{1.42}$$

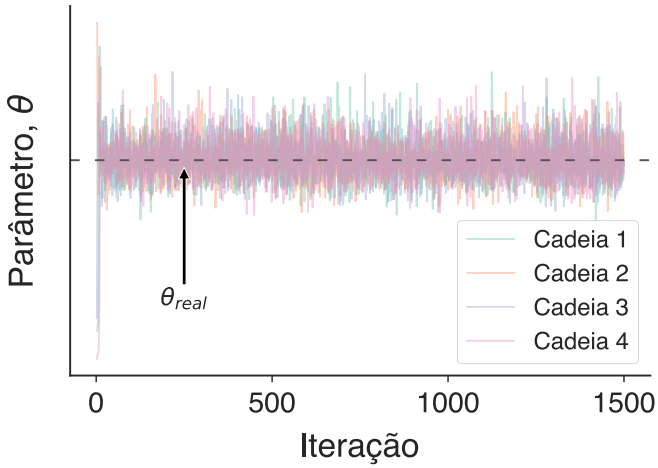

Figura 1.7: Mistura das cadeias de Markov. *Trace plot* de um processo ilustrativo de amostragem usando quatro cadeias de Markov.

em que m é a quantidade de cadeias de Markov, T é a quantidade de passos em cada cadeia e ρ_τ é a autocorrelação com atraso τ. Normalmente, não sabemos o valor exato de ρ_τ e, portanto, utilizamos a estimativa amostral $\hat{\rho}_\tau$. Uma outra interpretação de n_{ef} é a quantidade de passos realizados de maneira independente para um amostrador que utiliza cadeias de Markov.

1.5 Estimadores-M

Em análise de dados, recorremos a estatísticas descritivas para caracterizar conjuntos de dados. Por exemplo, muito usualmente calculamos a média de uma variável unidimensional y para estimar sua tendência média de localização por meio de

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i \,, \tag{1.43}$$

em que y_i é a i-ésima observação e N é o tamanho amostral. Para uma caracterização mais completa, podemos calcular uma medida de dispersão (também denominada de escala) pela variância amostral, dada por

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N} (y_i - \mu)^2 \,. \tag{1.44}$$

Entretanto, essas medidas deixam de ser representativas na presença de *outliers*[3]. Na presença de um único *outlier* divergente ($y_k \to \infty$), as medidas de média e variância divergem.

Para lidar com a presença de *outliers*, precisamos utilizar estatísticas descritivas robustas a esse tipo de comportamento. A mediana é uma medida de localização robusta comum. Ela é definida como o valor central que divide a amostra ordenada em duas metades. A escala, por sua vez, pode ser estimada pelo desvio da mediana (MAD), definido como a mediana dos desvios absolutos da mediana, ou seja,

$$\text{MAD} = k\ \text{mediana}(|y_i - \text{mediana}(y)|) \,, \tag{1.45}$$

em que estabelecemos a constante $k = 1.4826$ para que o estimador MAD seja consistente com o estimador desvio padrão [103]. Entretanto, apesar da simplicidade dessas medidas, qualitativamente elas deixam de apresentar a interpretação de valor médio e variância, representando, respectivamente, o ponto médio e o desvio absoluto do ponto médio.

Vamos definir um conjunto de estatísticas descritivas com propriedades robustas e com interpretação de média e desvio padrão, os chamados estimadores-M. Suponha que a dis-

[3] *Outliers* são observações com valores muito discrepantes quando comparados ao restante da amostra.

tribuição de probabilidade $f(y;\mu,\sigma)$ descreva uma variável aleatória y. Os parâmetros de localização μ e de escala σ não são inicialmente conhecidos. Podemos estimar esses parâmetros considerando a verossimilhança da amostra definida por

$$\mathcal{L}(\mu,\sigma) = \sum_i \sigma^{-1} f\left(\frac{y_i-\mu}{\sigma}\right), \tag{1.46}$$

em que a somatória abrange todo o conjunto de dados e a distribuição de probabilidade é normalizada e centrada na origem. Assim como realizamos na Seção 1.1, aplicamos uma transformação logarítmica que preserva as características do máximo da verossimilhança e tomamos seu negativo, isto é,

$$\rho = -\log \mathcal{L}(\mu,\sigma) = \sum_i \left[\log\sigma - \log f\left(\frac{y_i-\mu}{\sigma}\right)\right]. \tag{1.47}$$

Com essas mudanças, podemos tratar esse problema como a estimativa dos parâmetros por meio do método da maximização da verossimilhança[4]. A origem do nome dessa classe de estatísticas provém do nome do método utilizado para sua obtenção. O "M" dos estimadores-M decorre da "m"aximização da verossimilhança. Dessa forma, um estimador-M pode ser definido como qualquer parâmetro que maximiza a expressão

$$\sum_i \psi(y_i;\theta) = 0\,, \tag{1.48}$$

em que $\psi(y_i;\theta)$ é a derivada de ρ em relação ao parâmetro θ [104]. As relações da equação de maximização para os parâmetros de localização e escala definidos na Equação 1.47 são dadas, respectivamente, por

$$\begin{aligned} &\sum_i \psi\left(\frac{y_i-\mu}{\sigma}\right) = 0\,, \\ &\sum_i \left[\left(\frac{y_i-\mu}{\sigma}\right)\psi\left(\frac{y_i-\mu}{\sigma}\right) - 1\right] = 0\,. \end{aligned} \tag{1.49}$$

O último passo para obtenção dos estimadores-M é a escolha de uma função ψ adequada [105]. A função ψ representa a função geradora da amostra a partir de qual queremos inferir as medidas de localização e escala. Vamos listar algumas possíveis escolhas para a função ψ.

[4]Na realidade, ao tomar o negativo da verossimilhança, o problema em questão é de minimização. No entanto, decidimos seguir essa linha de raciocínio para manter o procedimento e notação originais de Huber [104].

Primeiramente, temos a função

$$\psi(y) = \begin{cases} y & \text{se } |y| < c \\ 0 & \text{caso contrário} \end{cases}, \tag{1.50}$$

que é a média cortada. Designamos um peso nulo para *outliers* definidos como valores superiores, em módulo, a uma constante arbitrária c que define o ponto de corte. Outra possibilidade de escolha para o ψ é a função

$$\psi(y) = \begin{cases} -c & \text{se } y < -c \\ y & \text{se } |y| < c \\ c & \text{se } y > c \end{cases}, \tag{1.51}$$

em que o peso para *outliers* é não nulo de valor c. Essa é a função geradora originalmente proposta por Huber [106]. Ao integrarmos ψ, obtemos a distribuição de probabilidade geradora a menos de uma constante. Para os dois exemplos anteriores, a parte central da distribuição de probabilidade é uma distribuição gaussiana. No caso da média cortada, as caudas da distribuição de probabilidade são nulas, enquanto que, para a proposta de Huber, as caudas são distribuições exponenciais duplas, ou seja,

$$\rho_H(y) = \begin{cases} y^2 & \text{se } |y| < c \\ c(2|y| - c) & \text{caso contrário} \end{cases}, \tag{1.52}$$

em que ρ_H é o logaritmo da verossimilhança para proposta de Huber. Outras escolhas comuns para a função ψ são a função de duplo peso de Tukey

$$\psi(y) = y\left[1 - \left(\frac{y}{R}\right)^2\right]_+^2, \tag{1.53}$$

em que R é uma constante e o símbolo + refere-se à parte positiva da função, e a função de Hampel

$$\psi(y) = sign(x) \begin{cases} |y| & \text{se } 0 < |y| < a \\ a & \text{se } a < |y| < b \\ a(c - |y|)/(c - b) & \text{se } b < |y| < c \\ 0 & \text{se } c < |y| \end{cases}, \tag{1.54}$$

em que a, b e c são constantes. A Figura 1.8 apresenta a representação gráfica das quatro funções ψ descritas anteriormente. Em nosso trabalho, escolhemos a função de Huber para calcular os estimadores-M de localização e escala. Dessa forma, como estamos interessados em ambos os parâmetros, precisamos estimá-los conjuntamente. Nesse cenário, é necessário

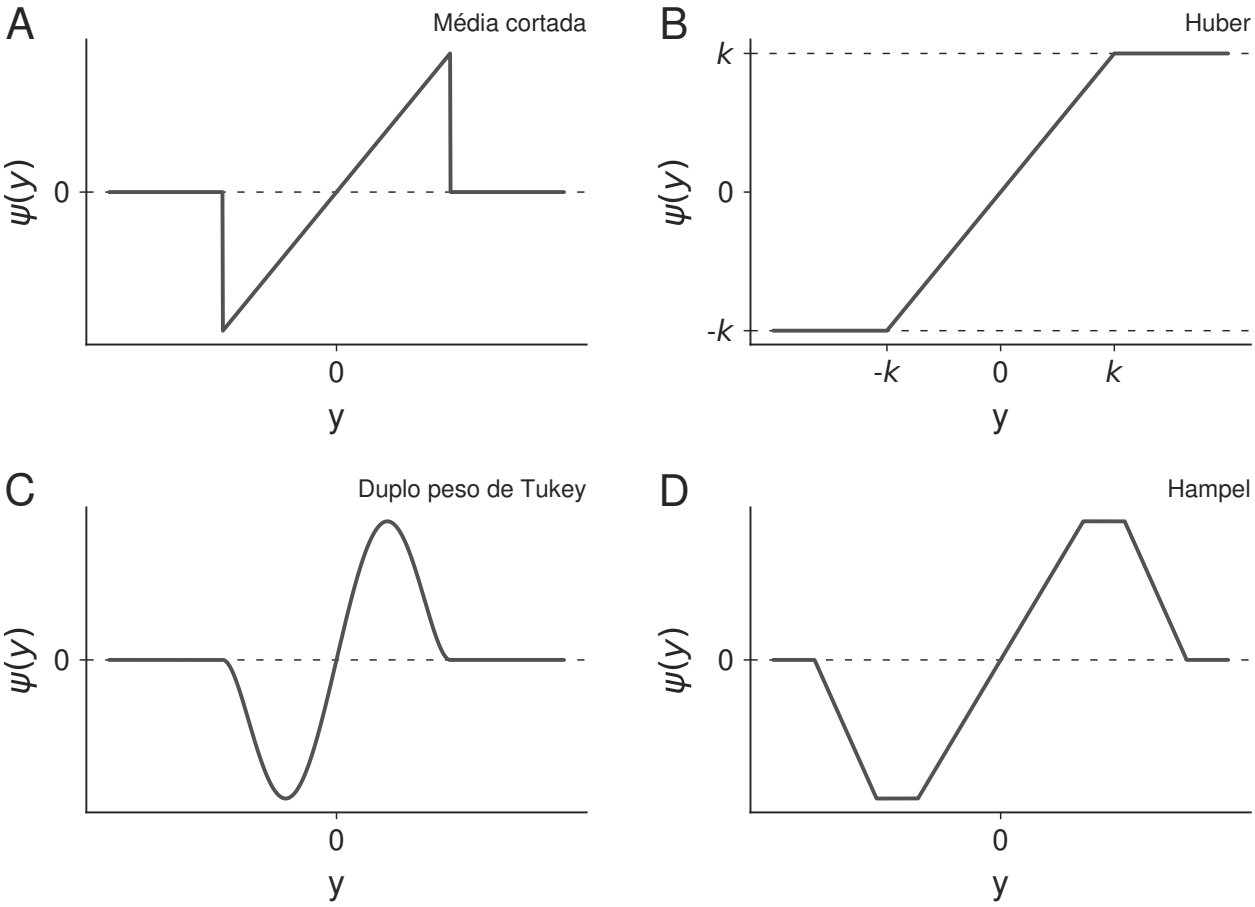

Figura 1.8: Diferentes funções $\psi(y)$ para determinação do estimador-M. (A) Média cortada. (B) Função de Huber. (C) Função duplo peso de Tukey. (D) Função de Hampel.

realizar uma pequena correção na equação de maximização da verossimilhança do parâmetro de escala a fim de que a estimativa seja não enviesada em relação à distribuição normal [105, 107]. As equações tornam-se, então,

$$\begin{aligned} \sum_i \left[\left(\frac{y_i - \mu}{\sigma} \right) \psi \left(\frac{y_i - \mu}{\sigma} \right) \right] &= (n-1)a(c)\,, \\ \sum_i \psi \left(\frac{y_i - \mu}{\sigma} \right) &= 0\,, \end{aligned} \tag{1.55}$$

em que $a(c)$ é a constante adicionada com o intuito de tornar a estimativa não enviesada.

Computacionalmente, a solução das Equações 1.55 e, portanto, a estimativa dos parâmetros de localização e de escala, raízes da equação, pode ser realizada por meio do método numérico de Newton [108]. Partindo de valores próximos ao valor verdadeiro do parâmetro, o método consiste em aproximar a função como a reta tangente à própria função para estimar uma raiz aproximada, repetindo o procedimento até que a variação entre iterações sucessivas seja pequena. Matematicamente, considerando uma função arbitrária $h(x)$ e um valor inicial

x_n, a reta tangente à função é dada por

$$y = h'(x_n)(x - x_n) + h(x_n)\,. \tag{1.56}$$

O valor da raiz aproximada é obtida igualando a Equação 1.56 a zero, isto é,

$$x_{n+1} = x_n - \frac{h'(x_n)}{h(x_n)}\,. \tag{1.57}$$

A Figura 1.9 ilustra duas iterações do método de Newton para uma função arbitrária $h(x)$. Utilizamos a mediana como estimativa inicial para calcular o valor do parâmetro de localização e o MAD como estimativa inicial para calcular o valor do parâmetro de escala. As equações de atualização dos parâmetros são dadas por

$$\begin{aligned} [\sigma_{n+1}]^2 &= \frac{1}{(n-1)a(c)} \sum_i \psi^2(y_{n,i})[\sigma_n]^2\,, \\ \mu_{n+1} &= \mu_n + \frac{\sum_i \psi(y_{n,i})\sigma_n}{\psi'(y_{n,i})}\,, \end{aligned} \tag{1.58}$$

em que ψ é a função escolhida por Huber, dada pela Equação 1.51, e a constante $a(c)$ é otimizada considerando a eficiência assimptótica de μ e o limite inferior da função de influência[5]. Em nossos resultados, utilizamos o pacote *statsmodels* [81] do *Python* para o cálculo dessas medidas. Por padrão, a constante $|c|$ é fixada em 1.5.

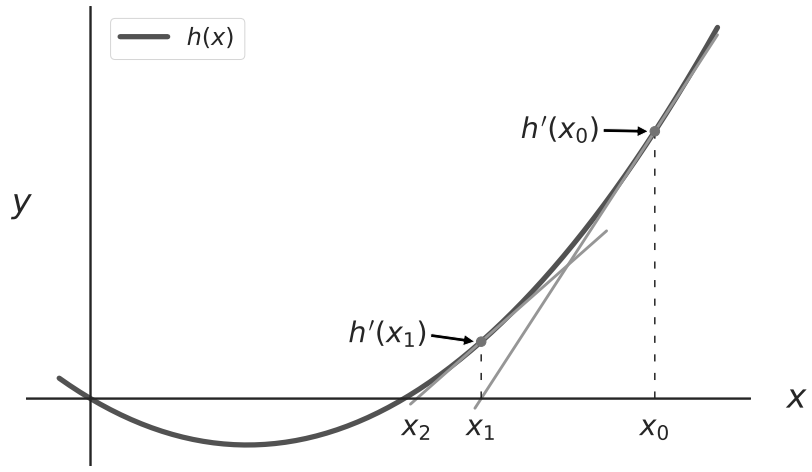

Figura 1.9: Ilustração do método de Newton para uma função arbitrária $h(x)$.

[5]Para mais detalhes recomendamos Staudte e Sheather [107].

1.6 Uniform Manifold Approximation and Projection

O *Uniform Manifold Approximation and Projection* (UMAP) é uma técnica de redução de dimensionalidade baseada em aprendizado de variedades (*manifold learning*) [109]. Essa técnica cria duas redes pesadas, por meio do algoritmo de grafo de k vizinhos mais próximos, associadas aos dados em alta e baixa dimensão. A projeção em baixa dimensão resultante é aquela que maximiza a similaridade entre as duas redes. O UMAP foi desenvolvido com base em conceitos da teoria da topologia algébrica e de categorias, cujos teoremas garantem que a representação de rede pesada derivada no procedimento é equivalente à estrutura topológica do dado (recomendamos as referências [109,110] para detalhes sobre o formalismo e os teoremas que suportam as escolhas algorítmicas adotadas no método). Existem duas suposições fundamentais na derivação do algoritmo UMAP [109]. A primeira é que existe uma variedade em que os dados encontram-se uniformemente distribuídos. A segunda é que a variedade de interesse é localmente conectada. No decorrer desta seção, vamos explicar e explicitar como essas suposições são introduzidas no método.

Começamos considerando uma variedade riemanniana $(\mathcal{M}, g)$ que contém um conjunto de dados $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \in \mathbb{R}^{n \times N}$, em que N é o tamanho do conjunto de dados e n é o número de dimensões. A representação dos dados multidimensionais é ilustrada na Figura 1.10A. A métrica riemanniana determina o produto interno g_p num ponto p, com $p \in \mathcal{M}$, em cada espaço tangente T_pM. Inicialmente, supomos que os dados $\mathbf{X}$ estão uniformemente distribuídos em $\mathcal{M}$. Em outras palavras, longe das regiões de borda, uma esfera de tamanho fixo posicionada em qualquer região de $\mathcal{M}$ engloba a mesma quantidade de pontos ou, alternativamente, uma esfera contendo os k primeiros vizinhos centrada em $\mathbf{x}_i$ deve manter o volume constante independentemente da escolha de $\mathbf{x}_i \in \mathbf{X}$. A suposição de uniformidade garante que a estrutura topológica dos dados na representação de rede é derivada de forma homogênea para todos os pontos em $\mathbf{X}$. Na prática, entretanto, a distribuição dos dados geralmente não é uniforme em $\mathbb{R}^n$.

Para assegurar a validade da suposição de uniformidade em $\mathcal{M}$, podemos introduzir uma medida de distância particular associada a cada dado $\mathbf{x}_i$, normalizando as distâncias geodésicas em relação ao k-ésimo primeiro vizinho de $\mathbf{x}_i$. Assim, cada ponto $\mathbf{x}_i$ associa-se com uma esfera unitária em $\mathbb{R}^n$ que se estende até seu k-ésimo vizinho. A Figura 1.10B mostra as esferas unitárias obtidas pelas definições locais de distância com $k = 2$ vizinhos. Considerando todos os pontos $\mathbf{x}_i$, temos uma família de métricas, composta pelas definições individuais de distância para cada $\mathbf{x}_i$, que podem ser agregadas numa estrutura global consistente para capturar a estrutura topológica do dado. A escolha do parâmetro k determina como a variedade $\mathcal{M}$ representa a estrutura topológica dos dados. Valores pequenos de k fazem com que a métrica capture a estrutura local do dado. Por outro lado, valores grandes de k fazem com que a métrica capture a estrutura global do dado, perdendo detalhes sobre a estrutura

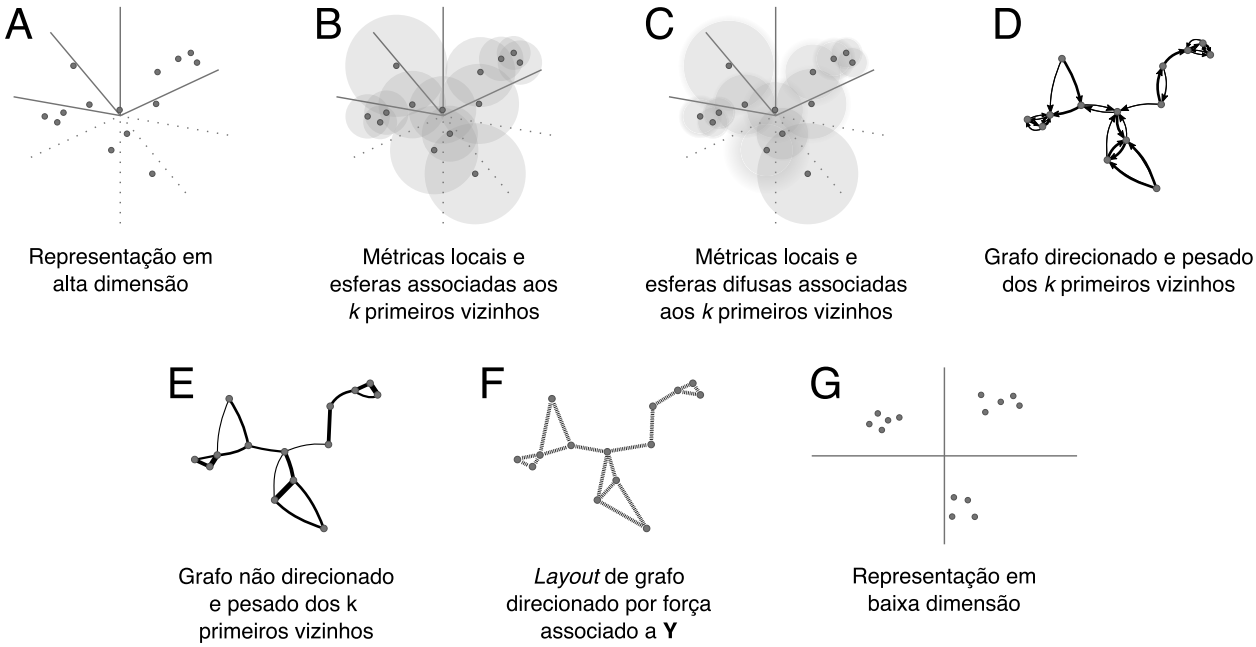

Figura 1.10: Ilustração do método *Uniform Manifold Approximation and Projection* (UMAP). Os painéis mostram as sete etapas que compõem o método UMAP de redução de dimensionalidade. (A) Representação multidimensional dos dados $\mathbf{X}$. (B) Métricas locais e esferas que englobam os $k = 2$ primeiros vizinhos para cada ponto $\mathbf{x}_i \in \mathbf{X}$. (C) Métricas locais e esferas difusas, representando a similaridade entre os pontos, definidas para cada ponto $\mathbf{x}_i \in \mathbf{X}$ via Equações 1.59 e 1.60. (D) Grafo direcionado e pesado construído a partir dos pesos calculados via Equação 1.59. (E) Grafo não direcionado e pesado construído a partir da simetrização da matriz de adjacência A associada ao grafo do painel (D). (F) Aplicação do algoritmo de *layout* de grafo direcionado por força, associado aos dados projetados em baixa dimensão $\mathbf{Y}$. (G) Representação final dos dados em baixa dimensão obtida após a aplicação do algoritmo do painel (F).

local, mas baseando as estimativas em mais informações.

Para compensar discrepâncias das distâncias entre os vizinhos dentro da esfera (principalmente em regiões esparsas) [111], estimamos a noção local de distância entre a observação x_i e seus vizinhos x_j a partir uma função de base radial, também conhecida como *kernel* gaussiano, dada por

$$\mu(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} \exp\left[\dfrac{-\max(0, d_{\mathbb{R}^n}(\mathbf{x}_i, \mathbf{x}_j) - \rho_i)}{\sigma_i}\right], & \text{para } \mathbf{x}_j \in C(\mathbf{x}_i; k) \\ 0, & \text{caso contrário} \end{cases}, \tag{1.59}$$

em que $\rho_i = \min\{d_{\mathbb{R}^n}(\mathbf{x}_i, \mathbf{x}_{ij}) \mid 1 \leq j \leq k, d_{\mathbb{R}^n}(\mathbf{x}_i, \mathbf{x}_{ij}) > 0\}$ é a distância ao vizinho mais próximo de $\mathbf{x}_i$ e σ_i é um fator de normalização. A adição de ρ_i na expressão impõe que $\mathbf{x}_i$ tem similaridade unitária com pelo menos outra observação $\mathbf{x}_j$, garantindo a suposição de conectividade local. O fator de normalização σ_i está associado com a densidade local dos pontos ao redor de $\mathbf{x}_i$, normalizando as similaridades de cada $\mathbf{x}_i$ e tornando-as comparáveis

entre todas as variedades. O fator σ_i é calculado por meio de um algoritmo de busca binária usando a expressão [110]

$$\sum_{j=1}^{k} \exp\left[\frac{-\max(0, d_{\mathbb{R}^n}(\mathbf{x}_i, \mathbf{x}_j) - \rho_i)}{\sigma_i}\right] = \log_2(k)\,, \tag{1.60}$$

em que a igualdade com $\log_2(k)$ foi fixada com base em experimentos empíricos [109]. Cada $\mathbf{x}_i$ está associado com uma esfera unitária em $\mathbb{R}^n$ até o primeiro vizinho, que representa similaridade unitária. Para além dessa esfera, as similaridades decrescem exponencialmente com escala σ_i definida pela densidade estimada localmente. A Figura 1.10C mostra essa representação difusa das similaridades locais.

Em posse da nova definição de similaridade de cada variedade, podemos construir a rede pesada que representa a estrutura topológica dos dados multidimensionais. Nessa representação, cada ponto $\mathbf{x}_i$ conecta-se com seus k vizinhos mais próximos. Consideramos um grafo pesado direcionado $G = (V, E, \mu)$, em que V são os vértices correspondendo aos dados $\mathbf{X}$, E são as arestas dadas por $E = \{(\mathbf{x}_i, \mathbf{x}_{ij}) | 1 \leq i \leq N, 1 \leq j \leq k\}$ e os pesos μ são dados por $\mu = \{\mu(\mathbf{x}_i, \mathbf{x}_{ij}) | 1 \leq i \leq N, 1 \leq j \leq k\}$ via Equações 1.59 e 1.60. Podemos criar uma matriz de adjacência A contendo os pesos de todas as arestas da rede. Porém, notamos que os pesos das arestas conectando $\mathbf{x}_i$ a $\mathbf{x}_j$ e $\mathbf{x}_j$ a $\mathbf{x}_i$ não necessariamente são iguais, isto é, $\mu(\mathbf{x}_i, \mathbf{x}_j) \neq \mu(\mathbf{x}_j, \mathbf{x}_i)$. Assim, matriz de adjacência A associada ao grafo é assimétrica como mostra a Figura 1.10D. Para unir as definições dos N espaços métricos, calculamos a matriz simétrica B, construindo um grafo simétrico, ilustrado na Figura 1.10E, a partir de A por meio da expressão

$$B = A + A^\top - A \circ A^\top\,,$$

em que $\circ$ é o produto de Hadamard definido por $(A \circ A^\top)_{ij} = A_{ij} A_{ij}\ \forall i, j$. Considerando que o elemento A_{ij} é a probabilidade de que a aresta com direção de $\mathbf{x}_i$ a $\mathbf{x}_j$ exista, o elemento B_{ij} representa a probabilidade de pelo menos uma das duas arestas direcionadas (de $\mathbf{x}_i$ a $\mathbf{x}_j$ ou de $\mathbf{x}_j$ a $\mathbf{x}_i$) existir.

No próximo passo, buscamos uma representação em baixa dimensão dos dados. Consideramos um conjunto de pontos $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\} \in \mathbb{R}^{d \times N}$ que corresponde à representação em baixa dimensão ($d \ll n$) do dado $\mathbf{X} \in \mathbb{R}^{n \times N}$. Para garantir maior convergência e agilidade do algoritmo, iniciamos as coordenadas $\mathbf{Y}$ a partir do resultado de uma redução de dimensionalidade espectral [112], realizada a partir da rede pesada construída no passo anterior. Aos dados de baixa dimensão, associamos uma rede pesada $H = (V', E', \nu)$ com peso da aresta entre $\mathbf{y}_i$ e $\mathbf{y}_j$ representado por $\nu(\mathbf{y}_i, \mathbf{y}_j)$. Aproximamos cada $\nu(\mathbf{y}_i, \mathbf{y}_j)$ definindo uma função $\Phi : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ tal que [109]

$$\Phi(\mathbf{y}_i, \mathbf{y}_j) = (1 + a(\|\mathbf{y}_i - \mathbf{y}_j\|_2^2)^b)^{-1}\,, \tag{1.61}$$

em que a e b são parâmetros calculados por meio de um ajuste de mínimos quadrados não linear da curva $\Psi : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$, que delimita a distância mínima `mindist` entre dois pontos no espaço de baixa dimensão como [109]

$$\Psi(\mathbf{y}_i, \mathbf{y}_j) = \begin{cases} 1, & \text{se } \|\mathbf{y}_i - \mathbf{y}_j\|_2 \leq \texttt{mindist} \\ \exp\left(-\|\mathbf{y}_i - \mathbf{y}_j\|_2 + \texttt{mindist}\right), & \text{caso contrário} \end{cases} . \tag{1.62}$$

O parâmetro `mindist` afeta apenas a distribuição dos pontos na projeção em baixa dimensão. Valores pequenos do `mindist` podem resultar em uma projeção com pontos densamente agrupados. Valores grandes do `mindist` forçam os pontos a se espalhar, ajudando na visualização dos dados projetados. Esse parâmetro é essencialmente estético, afetando apenas a aparência da projeção. Os parâmetros padrões do UMAP ($k = 15$ e `mindist` $= 0.1$) resultam em $a \approx 1.929$ e $b \approx 0.7915$ [109].

Para encontrar a rede pesada em baixa dimensão representativa do grafo em alta dimensão, utilizamos a entropia cruzada entre as duas representações como a função de custo no método de otimização de gradiente descendente, com $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ como as variáveis a serem otimizadas. A entropia cruzada pode ser definida como [109]

$$C(\mu, \nu) = \sum_{V,V'} \mu(\mathbf{x}_i, \mathbf{x}_j) \log\left(\frac{\mu(\mathbf{x}_i, \mathbf{x}_j)}{\nu(\mathbf{y}_i, \mathbf{y}_j)}\right) + (1 - \mu(\mathbf{x}_i, \mathbf{x}_j)) \log\left(\frac{1 - \mu(\mathbf{x}_i, \mathbf{x}_j)}{1 - \nu(\mathbf{y}_i, \mathbf{y}_j)}\right) . \tag{1.63}$$

Após descartar os termos que não dependem de $\mathbf{Y}$, a Equação 1.63 pode ser reescrita como

$$C(\mu, \nu) = \sum_{V,V'} \mu(\mathbf{x}_i, \mathbf{x}_j) \log\left(\nu(\mathbf{y}_i, \mathbf{y}_j)\right) + (1 - \mu(\mathbf{x}_i, \mathbf{x}_j)) \log\left(1 - \nu(\mathbf{y}_i, \mathbf{y}_j)\right) .$$

O problema de minimização da entropia cruzada, e consequentemente de redução de dimensionalidade, pode ser interpretado como a busca de um *layout* de grafo direcionado por força [109, 110] como mostra a Figura 1.10F. Para interpretar o problema dessa forma, definimos

$$\begin{aligned} C^{\text{atr}}(\mathbf{y}_i, \mathbf{y}_j) &= -\log\left(\nu(\mathbf{y}_i, \mathbf{y}_j)\right) , \\ C^{\text{rep}}(\mathbf{y}_i, \mathbf{y}_j) &= -\log\left(1 - \nu(\mathbf{y}_i, \mathbf{y}_j)\right) , \end{aligned}$$

em que $C^{\text{atr}}(\mathbf{y}_i, \mathbf{y}_j)$ é a parte da entropia relacionada com a força atrativa e $C^{\text{rep}}(\mathbf{y}_i, \mathbf{y}_j)$ é a parte da entropia relacionada com a força repulsiva. Assim, rescrevemos a entropia cruzada como

$$C(\mu, \nu) = \sum_{V,V'} \mu(\mathbf{x}_i, \mathbf{x}_j) C^{\text{atr}}(\mathbf{y}_i, \mathbf{y}_j) + (1 - \mu(\mathbf{x}_i, \mathbf{x}_j)) C^{\text{rep}}(\mathbf{y}_i, \mathbf{y}_j) . \tag{1.64}$$

Do primeiro termo da Equação 1.64, deriva-se uma força atrativa $F^{\text{atr}}(\mathbf{y}_i, \mathbf{y}_j)$ entre os pontos

$\mathbf{y}_i$ e $\mathbf{y}_j$ dada por

$$F^{\text{atr}}(\mathbf{y}_i, \mathbf{y}_j) = \frac{\partial C^{\text{atr}}(\mathbf{y}_i, \mathbf{y}_j)}{\partial \mathbf{y}_i} = \frac{-ab\|\mathbf{y}_i - \mathbf{y}_j\|_2^{2(b-1)}}{1 + \|\mathbf{y}_i - \mathbf{y}_j\|_2^2}\mu(\mathbf{x}_i, \mathbf{x}_j)(\mathbf{y}_i - \mathbf{y}_j)\,.$$

Esse termo apenas aparece quando $\mu(\mathbf{x}_i, \mathbf{x}_j) \neq 0$, o que significa que $\mathbf{x}_i$ é vizinho de $\mathbf{x}_j$, $\mathbf{x}_j$ é vizinho de $\mathbf{x}_i$ ou ambos os casos acontecem. Essa força é responsável pela atração de pontos vizinhos na projeção em baixa dimensão. Por outro lado, do segundo termo, deriva-se uma força repulsiva $F^{\text{rep}}(\mathbf{y}_i, \mathbf{y}_j)$ entre os pontos $\mathbf{y}_i$ e $\mathbf{y}_j$ dada por

$$F^{\text{rep}}(\mathbf{y}_i, \mathbf{y}_j) = \frac{\partial C^{\text{rep}}(\mathbf{y}_i, \mathbf{y}_j)}{\partial \mathbf{y}_i} = \frac{2b}{(\varepsilon + \|\mathbf{y}_i - \mathbf{y}_j\|_2^2)(1 + a\|\mathbf{y}_i - \mathbf{y}_j\|_2^{2b})}(1 - \mu(\mathbf{x}_i, \mathbf{x}_j))(\mathbf{y}_i - \mathbf{y}_j)\,,$$

em que ε é um número pequeno para evitar a divisão por zero quando $\mathbf{y}_i = \mathbf{y}_j$. Essa força é responsável pelo afastamento das projeções em baixa dimensão de pontos não vizinhos para longe um do outro. Após resolvermos esse problema de otimização com o método de gradiente descendente, chegamos à representação $\mathbf{Y}$ em baixa dimensão dos dados $\mathbf{X}$ em alta dimensão, ilustrada na Figura 1.10G.

1.7 Infomap

O Infomap é uma técnica de agrupamento que emprega passeios aleatórios como *proxy* do fluxo de informação na rede [113]. A estrutura modular resultante é composta por arestas responsáveis por transmitir a informação na rede e por grupos que representam regiões da rede em que a informação flui rapidamente e facilmente [113]. No processo de agrupamento, o algoritmo codifica os vértices com o objetivo de descrever eficientemente os caminhos percorridos pelo caminhante aleatório usando uma linguagem que reflete a estrutura modular da rede. O método de codificação Huffman une naturalmente as características de compressão ótima e de codificação semanticamente compatível com a representação em rede [113]. O método designa códigos curtos para vértices percorridos com frequência alta e códigos longos para vértices percorridos com frequência baixa, considerando uma caminhada aleatória infinitamente longa [113]. A codificação Huffman funciona de maneira similar ao código morse, que usa códigos curtos para letras mais utilizadas e códigos longos para letras menos utilizadas [114]. O interesse está no limite teórico do quão concisamente se consegue especificar a trajetória do caminhante aleatório na rede. Esse limite é descrito pela entropia de Shannon. Considerando n códigos utilizados para descrever os n estados da variável aleatória X, que ocorrem com frequências p_i, o comprimento médio do código não pode ser menor do que a

entropia de Shannon [115] definida como

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i \,, \tag{1.65}$$

que descreve a quantidade de *bits* necessária para representar a distribuição de probabilidade $P = \sum_i p_i$. Na perspectiva de otimizar a quantidade de informação para descrever a rede, estamos interessados em separar as informações importantes, referentes às estruturas modulares, das informações secundárias, referentes às estruturas internas dos grupos. Assim, dividimos a descrição da rede em dois níveis de detalhamento. O primeiro nível consiste da codificação de Huffman única designada para cada grupo. O segundo nível consiste da codificação de Huffman dentro dos grupos que utiliza códigos que se repetem entre grupos. Emprestando a analogia utilizada por Rosvall e Bergstrom [113], essa abordagem é similar à designação de nomes para cidades, ruas e avenidas. Os nomes das cidades em um país são únicos, enquanto os nomes das ruas e avenidas em cada cidade podem se repetir, como é o caso de nomes comuns tais quais Brasil e Tiradentes.

De forma geral, o Infomap fundamenta-se no problema dual em que se busca pela estrutura de comunidade da rede, ao mesmo tempo em que se procura pela codificação mais eficiente da rede. Considerando uma partição modular $\mathbf{M}$ de n vértices divididos em m módulos, procuramos minimizar o comprimento de descrição médio por passo de uma caminhada aleatória infinita na rede. O comprimento de descrição médio por passo é calculado pela equação mapa, isto é,

$$L(\mathbf{M}) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{m} p^i_{\circlearrowright} H(\mathcal{P}^i) \,. \tag{1.66}$$

O primeiro termo descreve a entropia $H(\mathcal{Q})$ de movimento entre os módulos pesada pela probabilidade por passo $q_{\curvearrowright}$ de mudança de módulo pelo caminhante aleatório. O segundo termo descreve a entropia $H(\mathcal{P}^i)$ dentro dos módulos pesada pela fração $p^i_{\circlearrowright}$ dos movimentos dentro do módulo i, sujeito à condição $\sum_{i=1}^{m} p^i_{\circlearrowright} = 1 + q_{\curvearrowright}$. O primeiro termo informa a quantidade média de *bits* necessária para descrever o movimento entre grupos, enquanto o segundo termo informa a quantidade média de *bits* necessária para descrever o movimento dentro dos grupos. Assim, a equação mapa fornece a quantidade de *bits* média necessária para descrever uma caminhada aleatória infinita em uma rede particionada conforme $\mathbf{M}$. A minimização da equação reflete a estrutura que melhor representa os padrões de fluxo dentro da rede, com cada grupo representando regiões em que o caminhante passa mais tempo antes de mudar de grupo durante a caminhada.

A probabilidade por passo $q_{\curvearrowright}$ de mudança de módulo pelo caminhante aleatório é calcu-

lada por

$$q_{\curvearrowright} = \sum_{i=1}^{m} q_{i\curvearrowright} \,,$$

em que $q_{i\curvearrowright}$ é a probabilidade por passo de saída do módulo i. Para garantir uma distribuição única no estado estacionário e para lidar com a presença de vértices isolados na estrutura de rede, introduzimos uma probabilidade de teletransporte τ do caminhante aleatório para qualquer vértice da rede, fixando o valor de $\tau = 0.15$ de forma análoga ao algoritmo *PageRank* do Google [116]. Nessa configuração, o movimento do caminhante é descrito por uma cadeia de Markov aperiódica e irredutível com um estado estacionário único garantido pelo teorema de Perron-Frobineous [113]. Na descrição das equações, vamos reservar os índices i e j para grupos e os índices α e β para vértices. Dessa forma, a probabilidade $q_{i\curvearrowright}$ pode ser calculada por

$$q_{i\curvearrowright} = \tau \frac{n - n_i}{n - 1} \sum_{\alpha \in i} p_\alpha + (1 - \tau) \sum_{\alpha \in i} \sum_{\beta \notin i} p_\alpha w_{\alpha\beta} \,,$$

em que n_i é o número de vértices no módulo i, $w_{\alpha\beta}$ são os pesos associados às arestas saindo do grupo α para o grupo β com $\sum_\beta w_{\alpha\beta} = 1$, $\alpha \in i$ indica todos os vértices do módulo i e p_α é a fração ergódica de visita para o vértice α calculada pelo método das potências. O primeiro termo da equação refere-se à probabilidade de teletransporte. O segundo termo refere-se à probabilidade de saída do módulo i sem teletransporte.

Usando a definição da entropia de Shannon, podemos escrever entropia $H(\mathcal{Q})$ de movimentos entre módulos interpretando os grupos como m estados da variável aleatória X da Equação 1.65, que ocorrem com frequência relativa $\mathcal{Q} = q_{1\curvearrowright}/q_{\curvearrowright}, \ldots, q_{m\curvearrowright}/q_{\curvearrowright}$, resultando em

$$H(\mathcal{Q}) = -\sum_{i=1}^{m} \frac{q_{i\curvearrowright}}{q_{\curvearrowright}} \log \frac{q_{i\curvearrowright}}{q_{\curvearrowright}} \,.$$

Os pesos $p^i_{\circlearrowright}$ da entropia de movimento dentro de cada grupo i são definidos como a soma das probabilidades de saída e das frequências ergódicas dos vértices do módulo, isto é,

$$p^i_{\circlearrowright} = q_{i\curvearrowright} + \sum_{\alpha \in i} p_\alpha \,,$$

em que a probabilidade $q_{i\curvearrowright}$ de saída do módulo i é incluída porque a codificação de saída do módulo é necessária para distinguir movimentos dentro do módulo e entre módulos. O termo da entropia $H(\mathcal{P}^i)$ de movimento dentro de cada grupo i, por sua vez, é dado por

$$H(\mathcal{P}^i) = -\frac{q_{i\curvearrowright}}{q_{i\curvearrowright} + \sum_{\beta \in i} p_\beta} \log \left(\frac{q_{i\curvearrowright}}{q_{i\curvearrowright} + \sum_{\beta \in i} p_\beta} \right) - \sum_{\alpha \in i} \frac{p_\alpha}{q_{i\curvearrowright} + \sum_{\beta \in i} p_\beta} \log \left(\frac{p_\alpha}{q_{i\curvearrowright} + \sum_{\beta \in i} p_\beta} \right) ,$$

em que o primeiro termo é a entropia de saída do módulo i e o segundo termo é a entropia

dos vértices do módulo i.

Com todos os termos da Equação 1.66 descritos, o problema de agrupamento consiste em encontrar a partição $\mathbf{M}$ e a codificação associada que minimizam a equação mapa. Os algoritmos de otimização do tipo Monte Carlo são precisos mas geralmente lentos [114], enquanto algoritmos míopes (*greedy search*) são rápidos mas menos precisos [114]. O método Infomap faz uso de um algoritmo míope cujas desvantagens serão contornadas com a aplicação de recursos algorítmicos. Iniciamos o algoritmo Infomap calculando as probabilidades p_α para todos os vértices da rede via método das potências, designando um código único para cada vértice e derivando as probabilidades de saída. Em seguida, cada vértice é considerado individualmente como seu próprio grupo de forma que começamos a partição $\mathbf{M}$ com n grupos. Aplicamos o algoritmo míope que realiza a junção de grupos vizinhos associada com o maior decréscimo no comprimento de descrição médio da rede. Esse processo é repetido até que não haja mais decréscimo no comprimento de descrição médio. A desvantagem dessa otimização reside na rigidez da partição $\mathbf{M}$ gerada, pois dois vértices, uma vez designados a um mesmo módulo, não podem ser separados no decorrer do procedimento. Para lidar com esse problema, utilizamos dois recursos algorítmicos. O primeiro recurso consiste em permitir o movimento de subgrupos. Consideramos cada grupo como sua própria rede, aplicamos o algoritmo míope nessa rede e, como resultado, o módulo é particionado em submódulos. Na partição total $\mathbf{M}$ da rede, esses submódulos ficam livres para transitar entre os módulos e são redesignados se houver diminuição no comprimento de descrição médio. O segundo recurso consiste em permitir o trânsito de vértices individuais entre módulos caso haja diminuição no comprimento de descrição médio. Juntos, os dois recursos tornam o algoritmo mais preciso.

Infomap hierárquico

Em várias situações, as redes apresentam estruturas em múltiplas escalas, isto é, existem subgrupos dentro dos grupos, que também podem conter mais subgrupos, e assim por diante. Entretanto, a equação mapa apresentada anteriormente não consegue capturar estruturas hierárquicas desse tipo, pois é apenas capaz de identificar estruturas em dois níveis. Para generalizar o método de agrupamento Infomap e capturar estruturas em vários níveis, Rosvall e Bergstrom [117] propuseram modificar a equação mapa para conter codificações de grupos e subgrupos. A equação mapa hierárquica [117] é dada por

$$L(\mathbf{M}) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{m} L(\mathbf{M}^i)\,, \tag{1.67}$$

em que a partição $\mathbf{M}$ de n vértices é dividida em m grupos. Para cada um dos m grupos, pode existir uma subpartição $\mathbf{M}^i$ com m^i subgrupos, para os quais também pode existir uma subpartição $\mathbf{M}^{ij}$ com m^{ij} subgrupos e assim sucessivamente. O primeiro termo da

Equação 1.67 é o mesmo da equação mapa de dois níveis, isto é,

$$q_{\curvearrowright} H(\mathcal{Q}) = q_{\curvearrowright} \left(- \sum_{i=1}^{m} \frac{q_{i\curvearrowright}}{q_{\curvearrowright}} \log \frac{q_{i\curvearrowright}}{q_{\curvearrowright}} \right) ,$$

que representa a entropia de movimento entre os módulos pesada pela probabilidade por passo de mudança de módulo pelo caminhante aleatório. O segundo termo da Equação 1.67 corresponde ao comprimento de descrição médio por passo $L(\mathbf{M}^i)$ da subpartição $\mathbf{M}^i$ e pode ser escrito, para cada submódulo i, como

$$L(\mathbf{M}^i) = q^i_{\circlearrowright} H(\mathcal{Q}^i) + \sum_{i=1}^{m^i} L(\mathbf{M}^{ij}) \,. \tag{1.68}$$

Nessa expressão, o termo $q^i_{\circlearrowright}$ corresponde à soma da probabilidade por passo $q^i_{\curvearrowright}$ de mudança de módulo com a probabilidade $\{q^{ij}_{\curvearrowright}\}$ de mudança para um dos m_i submódulos e é definida como

$$q^i_{\circlearrowright} = q^i_{\curvearrowright} + \sum_{j=1}^{m^i} q^{ij}_{\curvearrowright} \,.$$

A entropia associada a essas mudanças é escrita a partir da entropia de Shannon associada às frações $\mathcal{Q}^i = q^i_{\curvearrowright}/q^i_{\circlearrowright}, q^{i1}_{\curvearrowright}/q^i_{\circlearrowright}, \ldots, q^{im^i}_{\curvearrowright}/q^i_{\circlearrowright}$, dada por

$$H(\mathcal{Q}^i) = -\frac{q^i_{\curvearrowright}}{q^i_{\circlearrowright}} \log \frac{q^i_{\curvearrowright}}{q^i_{\circlearrowright}} - \sum_{j=1}^{m^i} \frac{q^{ij}_{\curvearrowright}}{q^i_{\circlearrowright}} \log \frac{q^{ij}_{\curvearrowright}}{q^i_{\circlearrowright}} \,.$$

O primeiro termo do comprimento de descrição médio para os movimentos no submódulo i na Equação 1.68 pode ser escrito como

$$q^i_{\circlearrowright} H(\mathcal{Q}^i) = q^i_{\circlearrowright} \left(-\frac{q^i_{\curvearrowright}}{q^i_{\circlearrowright}} \log \frac{q^i_{\curvearrowright}}{q^i_{\circlearrowright}} - \sum_{j=1}^{m^i} \frac{q^{ij}_{\curvearrowright}}{q^i_{\circlearrowright}} \log \frac{q^{ij}_{\curvearrowright}}{q^i_{\circlearrowright}} \right) .$$

Em se tratando do movimento dentro dos submódulos sucessivos, o segundo termo da Equação 1.68 pode ser obtido recursivamente por meio da própria equação. Todavia, para o nível mais fino, a estrutura que o caminhante aleatório visita são vértices e não mais grupos. Portanto, as frações relativas de visita são representadas por $\mathcal{P}^{ij\ldots k} = q^{ij\ldots k}_{\curvearrowright}/p^{ij\ldots k}_{\circlearrowright}$, com

$$p^{ij\ldots k}_{\circlearrowright} = q^{ij\ldots k}_{\curvearrowright} + \sum_{\alpha \in ij\ldots k} p_\alpha \,,$$

em que $q^{ij\ldots k}_{\curvearrowright}$ é a frequência de saída do módulo $ij\ldots k$ e p_α é a frequência de visita ergódica ao vértice α. A entropia de movimentos na estrutura mais fina pesada pelas frações de

visitação é, assim, descrita por

$$L(\mathbf{M}^{ij\ldots k}) = p_{\circlearrowleft}^{ij\ldots k} H(\mathcal{P}^{ij\ldots k}) = p_{\circlearrowleft}^{ij\ldots k} \left(-\frac{q_{\curvearrowright}^{ij\ldots k}}{p_{\circlearrowleft}^{ij\ldots k}} \log \frac{q_{\curvearrowright}^{ij\ldots k}}{p_{\circlearrowleft}^{ij\ldots k}} - \sum_{\alpha \in ij\ldots k} \frac{p_\alpha}{p_{\circlearrowleft}^{ij\ldots k}} \log \frac{p_\alpha}{p_{\circlearrowleft}^{ij\ldots k}} \right) .$$

Para encontrar as partições e subpartições que mais efetivamente representam a estrutura modular da rede, utilizamos o algoritmo de dois níveis recursivamente nos módulos e submódulos. Inicialmente, aplicamos o algoritmo descrito na seção anterior à estrutura completa da rede, encontrando a estrutura de grupos em dois níveis que gera o menor comprimento de descrição médio. Em seguida, cada módulo é considerado como uma rede independente. Aplicamos novamente o algoritmo a cada um desses módulos, buscando as subpartições que geram uma redução no comprimento de descrição global. Esse procedimento recursivo é realizado para todos os submódulos subsequentes até que não haja maiores reduções na equação mapa hierárquica. Além dos dois recursos algorítmicos empregados no algoritmo de dois níveis, utiliza-se um terceiro recurso para permitir o movimento de vértices entre grupos e subgrupos da estrutura hierárquica. Para isso, notamos que a estrutura mais fina compõe um ramo da estrutura hierárquica. O recurso consiste em agregar submódulos de ramos diferentes num único módulo, criando uma codificação num nível anterior, se essa junção resulta numa compressão do comprimento de descrição médio. Dessa forma e de modo geral, a busca pela melhor partição pode ser interpretada como uma variação, positiva ou negativa, da profundidade de cada ramo.

A Figura 1.11 mostra uma comparação entre os agrupamentos via Infomap de dois níveis e Infomap hierárquico. Os painéis mostram uma rede ilustrativa não direcionada e não pesada com uma estrutura hierárquica de grupos, cujo comprimento de descrição médio na ausência de agrupamento é de 4.75 *bits*. A Figura 1.11A mostra os grupos resultantes da aplicação do Infomap em dois níveis (círculos pretos). Esse agrupamento resulta num comprimento de descrição médio de 3.57 *bits*. A codificação utilizada pode ser calculada pela soma e normalização dos graus de cada vértice. No Infomap de dois níveis, as probabilidades são dadas por $\mathcal{Q} = \{2/24, 3/24, 3/24, 3/24, 2/24, 3/24, 3/24, 3/24, 2/24\}$, $q_{\curvearrowright} = 24/78$ e $\mathcal{P} = \{2/10, 3/10, 3/10, 2/10\}$. Ilustrativamente, temos que $\mathcal{P}^1 = \{2/10, 3/10, 3/10, 2/10\}$ e $p_{\circlearrowleft}^1 = 10/78$ para o primeiro grupo. A Figura 1.11B mostra os grupos e subgrupos resultantes da aplicação do Infomap hierárquico (círculos pretos). Esse agrupamento resulta num comprimento de descrição médio de 3.48 *bits*, melhorando a descrição da estrutura modular da rede. As probabilidades nesse caso são dadas por $\mathcal{Q} = \{2/6, 2/6, 2/6\}$, $\mathcal{Q}^1 = \{2/10, 3/10, 3/10, 2/10\}$, $q_{\curvearrowright} = 6/78$ e $q_{\curvearrowright}^1 = 10/78$, enquanto as probabilidades dos subgrupos são as mesmas para o caso de dois níveis.

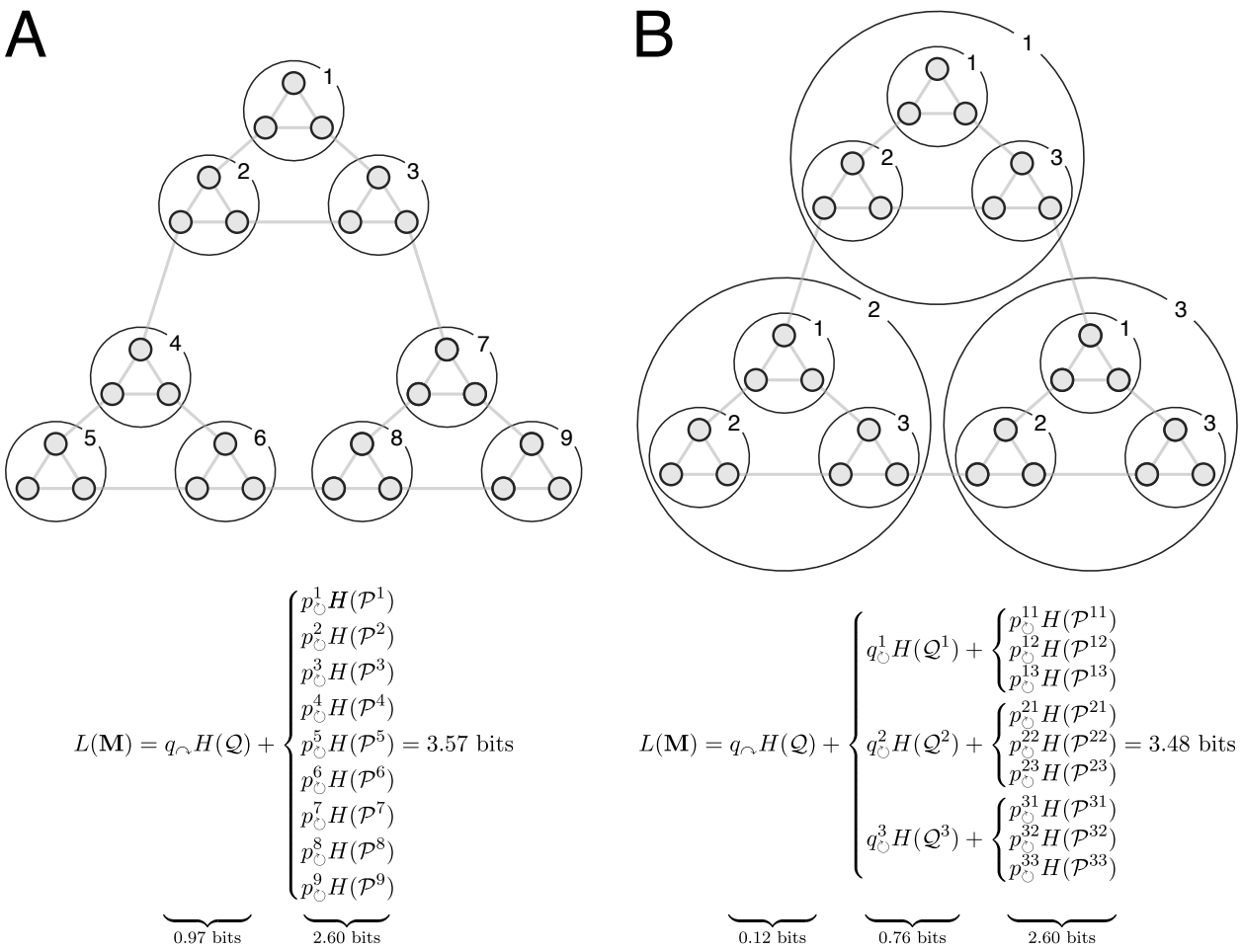

Figura 1.11: Comparação ilustrativa das performances do agrupamento em dois níveis e do agrupamento hierárquico. (A) Infomap de dois níveis. (B) Infomap hierárquico. Os círculos em preto acompanhados de números representam os grupos e subgrupos encontrados pela aplicação dos algoritmos de agrupamento. Na parte inferior, as equações detalham as contribuições para o comprimento de descrição médio de cada termo das equações mapa. A figura foi replicada da referência [117].

CAPÍTULO 2

Associação entre produtividade e impacto de jornal para diferentes disciplinas e estágios de carreira

Neste capítulo, estudamos aspectos variados da associação entre produtividade e impacto de jornal, considerando a inflação temporal desses indicadores, efeitos de disciplina, de estágio de carreira e de escala de produtividade, bem como a existência de pesquisadores que, em determinado momento da carreira, produzem em quantidade muito acima da média ou em revistas de altíssimo impacto [59].

2.1 Apresentação dos dados

A Plataforma Lattes [118] é a base de dados primária utilizada nesta tese. A plataforma, mantida pelo governo brasileiro desde 1999, contém o currículo oficial dos acadêmicos brasileiros – o currículo Lattes – que fornece, publicamente, uma ampla variedade de informações, como a disciplina acadêmica de atuação, relações de orientação e produção acadêmica detalhada. O currículo Lattes é amplamente utilizado para avaliação tanto individual quanto institucional. Por esse motivo, os pesquisadores precisam manter seus registros atualizados. Inicialmente, selecionamos todos os 14 487 pesquisadores, de 88 disciplinas, que possuíam bolsa produtividade do CNPq em maio de 2017. Obtivemos seu registro completo de publicações, resultando num total de 1 121 652 artigos científicos. A bolsa produtividade tem sido concedida em reconhecimento à eminente produção e impacto científico de pesquisadores desde os anos 70 pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). No Brasil, os bolsistas do CNPq são considerados como parte da elite científica do país. Filtramos os pesquisadores cujos currículos não foram atualizados pelo menos desde 1

de janeiro de 2016 e também aqueles que não incluíram informações sobre sua disciplina ou data de doutoramento, reduzindo o número para 14 146 pesquisadores. Para completar essa base de dados, incluímos informações faltantes sobre o ano de publicação e nome do jornal utilizando o código DOI de referência com a *API CrossRef*.

A fim de definir o prestígio do jornal dessas publicações, coletamos o fator de impacto de jornal (JIF, *Journal Impact Factor*) para todos os jornais científicos disponíveis nos relatórios de citação de jornais da Clarivate (*Clarivate's Journal Citation Reports*) entre 1997 e 2015. Combinamos ambos os conjuntos de dados para associar os valores variáveis no tempo do JIF aos artigos publicados pelos bolsistas do CNPq. Para cada um dos pesquisadores, calculamos o número de artigos publicados por ano (produtividade) e o valor médio do JIF dessas publicações (prestígio médio de jornal). Finalmente, agrupamos as séries temporais por disciplina, selecionando as 14 disciplinas com pelo menos 50 pesquisadores com artigos publicados em cada ano entre 1997 e 2015, condição necessária para o processo de deflação que será descrito na próxima subseção. Esses filtros nos levam ao nosso conjunto final de dados com 6 028 pesquisadores de 14 disciplinas e 312 881 artigos como mostra a Figura 2.1. Para além do conjunto de dados JIF, consideramos o ranque de jornais SCImago da Scopus (SJR, *Scopus' SCImago Journal Rank*) como medida de prestígio de jornal. Optamos por apresentar os resultados do JIF no texto principal e nos referimos ao Apêndice A para comparações com o SJR. Obtivemos todos os valores do SJR para os jornais disponíveis na Scopus entre os anos de 1999 e 2015. Usando o mesmo procedimento adotado para o JIF, definimos a produtividade e SJR médio para 448 959 artigos publicados por 8 465 pesquisadores de 25 disciplinas (Figura A.1).

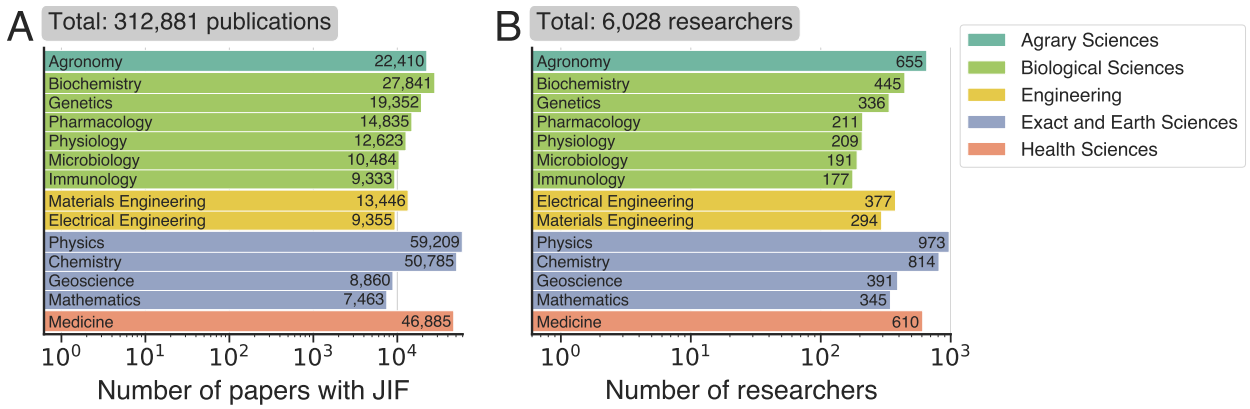

Figura 2.1: Número de publicações e pesquisadores no conjunto de dados JIF. O painel (A) mostra o número total de artigos e o painel (B) mostra o número total de pesquisadores para cada disciplina no conjunto de dados JIF. As cores das barras representam os diferentes campos da ciência em nosso conjunto de dados.

Enquanto o JIF de um jornal em dado ano é simplesmente definido como o número de citações recebidas por artigos publicados nos dois anos anteriores dividido pelo número total de artigos publicados nesses mesmos anos, o SJR é uma medida baseada em rede [119]

(especificamente, uma variante da métrica *eigenfactor* do algoritmo *PageRank*) de caráter mais complexo. Apesar dessa diferença, o JIF e o SJR são fortemente correlacionados, com coeficiente de correlação de Pearson $r = 0.85$ como mostra a Figura 2.2. Ambos os conjuntos de dados são formados por disciplinas de ciência, tecnologia, engenharia e matemática (disciplinas STEM, *Science, Technology, Engineering, and Mathematics*), o que reflete a maior consolidação desses ramos do conhecimento no Brasil e a consequente predominância de concessões de bolsas produtividade para pesquisadores dessas disciplinas.

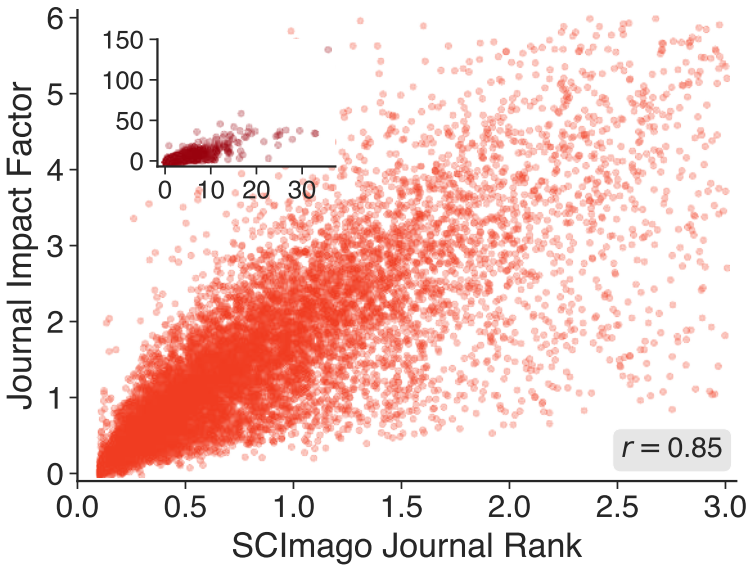

Figura 2.2: Fator de impacto de jornais (JIF) e ranque de jornais SCImago (SJR) são correlacionados. Gráfico de dispersão do SJR *versus* JIF para 11 055 jornais contidos em ambos os conjuntos de dados para o ano de 2015. A inserção mostra o gráfico de dispersão considerando o intervalo completo em que os dados estão disponíveis. O coeficiente de correlação de Pearson entre essas duas variáveis é $r = 0.85$, indicando uma correlação significativa entre essas medidas de prestígio de jornal. Os resultados são similares para outros anos em nosso conjunto de dados.

2.2 Inflação e medidas robustas de padronização

O volume de produção científica tem crescido com o tempo em nível global e individual [71, 72]. Esse crescimento acarreta um efeito de inflação temporal na produtividade e no impacto médio de jornal, impossibilitando a comparação de observações de períodos distintos. Nossas análises indicam que o JIF médio das publicações em nosso conjunto de dados tem crescido ≈ 0.72 unidades por década, como mostram as Figuras 2.3A e 2.3C. Similarmente, a produtividade média dos bolsistas CNPq tem crescido numa taxa de ≈ 1.57 artigos/ano por década, como mostram as Figuras 2.3B e 2.3D (veja a Figura A.2 para comparação com o conjunto de dados SJR). As disciplinas também apresentam volumes de publicação e dinâmicas de citações distintos [120, 121]. Como consequência, também existe

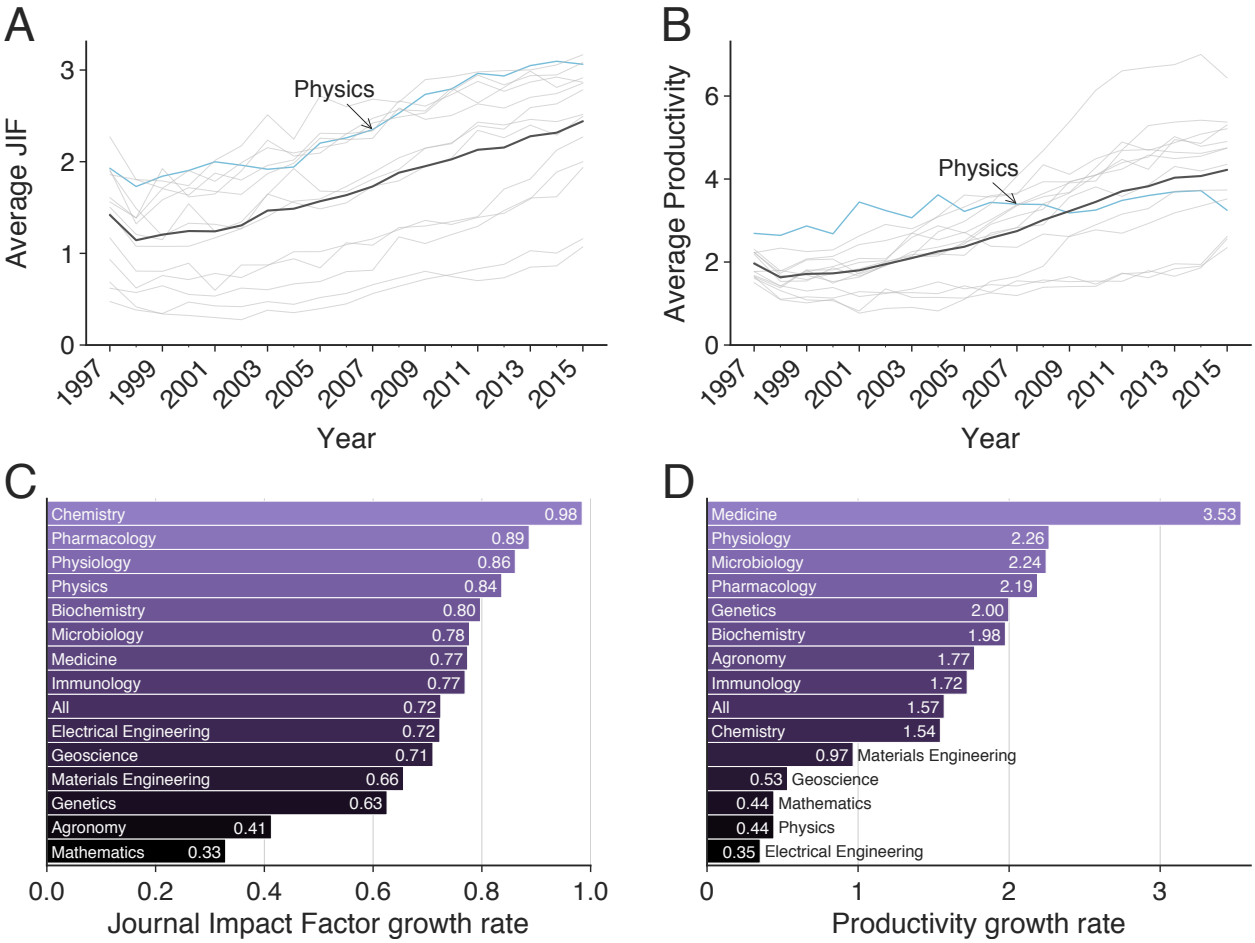

Figura 2.3: Evolução temporal do prestígio médio de jornal e da produtividade. Os painéis (A) e (B) mostram a evolução temporal dos valores médios do prestígio médio de jornal e da produtividade, respectivamente, para o conjunto de dados JIF. As curvas em cinza mostram o comportamento médio das disciplinas separadamente, as curvas em preto representam o comportamento médio agregado de todas as disciplinas e as curvas em azul ilustram o comportamento médio da disciplina de física. Os valores médios foram estimados utilizando o estimador de localização Huber. Os painéis (C) e (D) mostram as taxas de crescimento por década do prestígio médio de jornal e da produtividade, respectivamente, estimadas a partir do conjunto de dados JIF. Estimamos as taxas de crescimento ajustando um modelo linear à evolução temporal reportada nos painéis (A) e (B) para cada disciplina. Além disso, estimamos a taxa de crescimento agregando os dados de todas as disciplinas (indicado por *All* nos gráficos de barra).

dificuldade em comparar a produtividade e o impacto médio de jornal entre diferentes disciplinas. Em nosso conjunto de dados, por exemplo, a produtividade de pesquisadores da medicina tem crescido ≈ 3.5 artigos/ano por década, enquanto aqueles trabalhando com engenharia elétrica vivenciaram um aumento na produtividade de apenas ≈ 0.3 artigos/ano por década. Além disso, precisamos considerar que o impacto médio de jornal sofre um efeito de escala. Especificamente, a variabilidade do valor médio do prestígio diminui com o aumento da produtividade. De fato, em nosso conjunto de dados, pesquisadores pouco prolíficos

apresentam alta variabilidade nos valores médios de prestígio de jornal, enquanto pesquisadores muito prolíficos apresentam variabilidade significativamente menor, como mostra a Figura 2.4 (veja a Figura A.3 para comparação com o conjunto de dados SJR). Esse efeito foi anteriormente verificado em estudos sobre os fatores de impacto de jornais com diferentes números totais de publicação [122,123] e, como argumenta Antonoyiannakis [122,123], é uma consequência direta do Teorema Central do Limite.

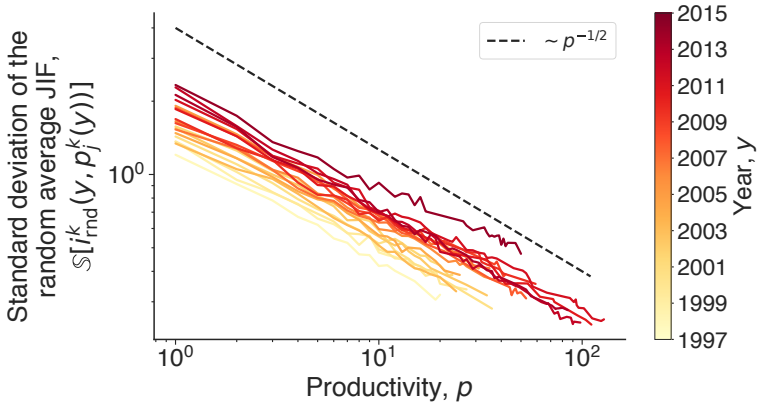

Figura 2.4: Efeito do tamanho da produtividade na dispersão do prestígio médio de jornal. Desvio padrão $\mathbb{S}[i^k_{\mathrm{rnd}}(y, p^k_j(y))]$ do valor médio do fator de impacto do jornal (JIF) para 1000 amostras aleatórias de p publicações de pesquisadores da física como uma função de p em todos os anos disponíveis no conjunto de dados JIF. O código de cor refere-se a cada ano do conjunto de dados e a linha tracejada representa o comportamento esperado pelo Teorema Central do Limite.

Para levar em consideração essas questões, usamos medidas z-score relativas ao ano e disciplina para produtividade e medidas z-score relativas ao ano, disciplina e nível de produtividade para o prestígio médio de jornal. Vamos assumir que $p^k_j(y)$ e $i^k_j(y)$ representam, respectivamente, o número de artigos e o prestígio médio de jornal de publicações de um pesquisador j da disciplina k no ano y. Com isso, calculamos os z-scores de produtividade como

$$P^k_j(y) = \frac{p^k_j(y) - \mathbb{E}[p^k_j(y)]}{\mathbb{S}[p^k_j(y)]},$$

em que $\mathbb{E}[p^k_j(y)]$ e $\mathbb{S}[p^k_j(y)]$ são, respectivamente, a média e o desvio padrão da produtividade dos pesquisadores da disciplina k no ano y. Similarmente, calculamos o z-score do prestígio de jornal como

$$I^k_j(y) = \frac{i^k_j(y) - \mathbb{E}[i^k_{\mathrm{rnd}}(y, p^k_j(y))]}{\mathbb{S}[i^k_{\mathrm{rnd}}(y, p^k_j(y))]},$$

em que $i^k_{\mathrm{rnd}}(y, p)$ é o impacto médio de jornal de uma amostra aleatória de p publicações da disciplina k no ano y e $\mathbb{E}[i^k_{\mathrm{rnd}}(y, p)]$ e $\mathbb{S}[i^k_{\mathrm{rnd}}(y, p)]$ são, respectivamente, a média e o desvio padrão de $i^k_{\mathrm{rnd}}(y, p)$ estimadas a partir de 1 000 realizações independentes. Esta última definição é uma adaptação do índice Φ proposto por Antonoyiannakis [122,123] para ranquear

jornais de diferentes tamanhos.

Usamos os estimadores robustos de Huber (veja a Seção 1.5) para média (localização) e desvio padrão (escala) em vez dos estimadores usuais [104] devido à existência de valores extremos, isto é, observações *outliers* para $p_j^k(y)$ e $i_j^k(y)$. As observações *outliers* podem ser observadas nos diagramas de caixa da Figura 2.5 (veja a Figura A.4 para comparação com o conjunto de dados SJR). Dessa forma, $\mathbb{E}[\dots]$ e $\mathbb{S}[\dots]$ representam, respectivamente, os estimadores de localização e escala de Huber (conforme implementado no pacote de Python *statsmodels* [81]).

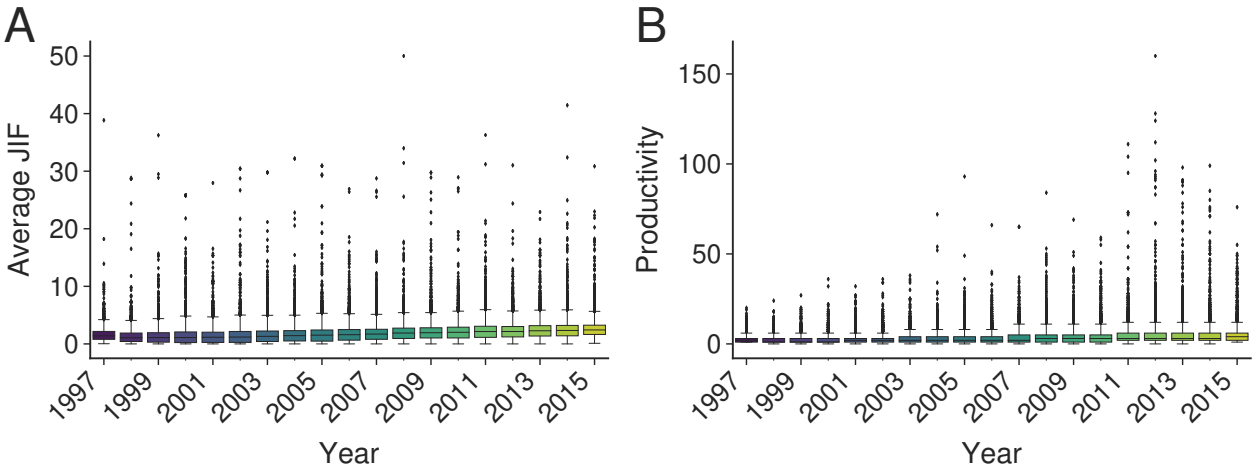

Figura 2.5: Valores *outliers* do prestígio médio de jornal e da produtividade. Os diagramas de caixa retratam o grau de dispersão do (A) prestígio médio de jornal (JIF) e da (B) produtividade dos pesquisadores no conjunto de dados JIF em cada ano. Existem observações extremas em todos os anos, que estão representados por marcadores pretos além dos bigodes (aqui definidos como 1.5 vezes o intervalo interquartil).

2.3 Plano prestígio de jornal *versus* produtividade

A Figura 2.6 mostra um diagrama de dispersão do prestígio de jornal *versus* produtividade para todos anos de carreira de pesquisadores em nosso conjunto de dados (veja a Figura A.5A para comparação com o conjunto de dados SJR). Nesse plano, uma unidade de produtividade indica uma performance de um desvio padrão acima (se positiva) ou abaixo (se negativa) da performance média de todos os acadêmicos de certa disciplina em dado ano. Similarmente, uma unidade de prestígio de jornal representa uma performance um desvio padrão acima (se positiva) ou abaixo (se negativa) da performance média aleatória em um dado nível de produtividade de certa disciplina e certo ano. Dividimos esse plano em quatro setores principais separando anos *outliers* dos pesquisadores (z-scores maiores do que 3.5) em relação à produtividade (P) e impacto de jornal (I). O setor $IP{+}{+}$ contém anos de carreira em que pesquisadores foram *outliers* simultaneamente em produtividade e prestígio de jornal ($I > 3.5$ e $P > 3.5$). Similarmente, os setores $I{+}{+}$ e $P{+}{+}$ indicam anos de carreira

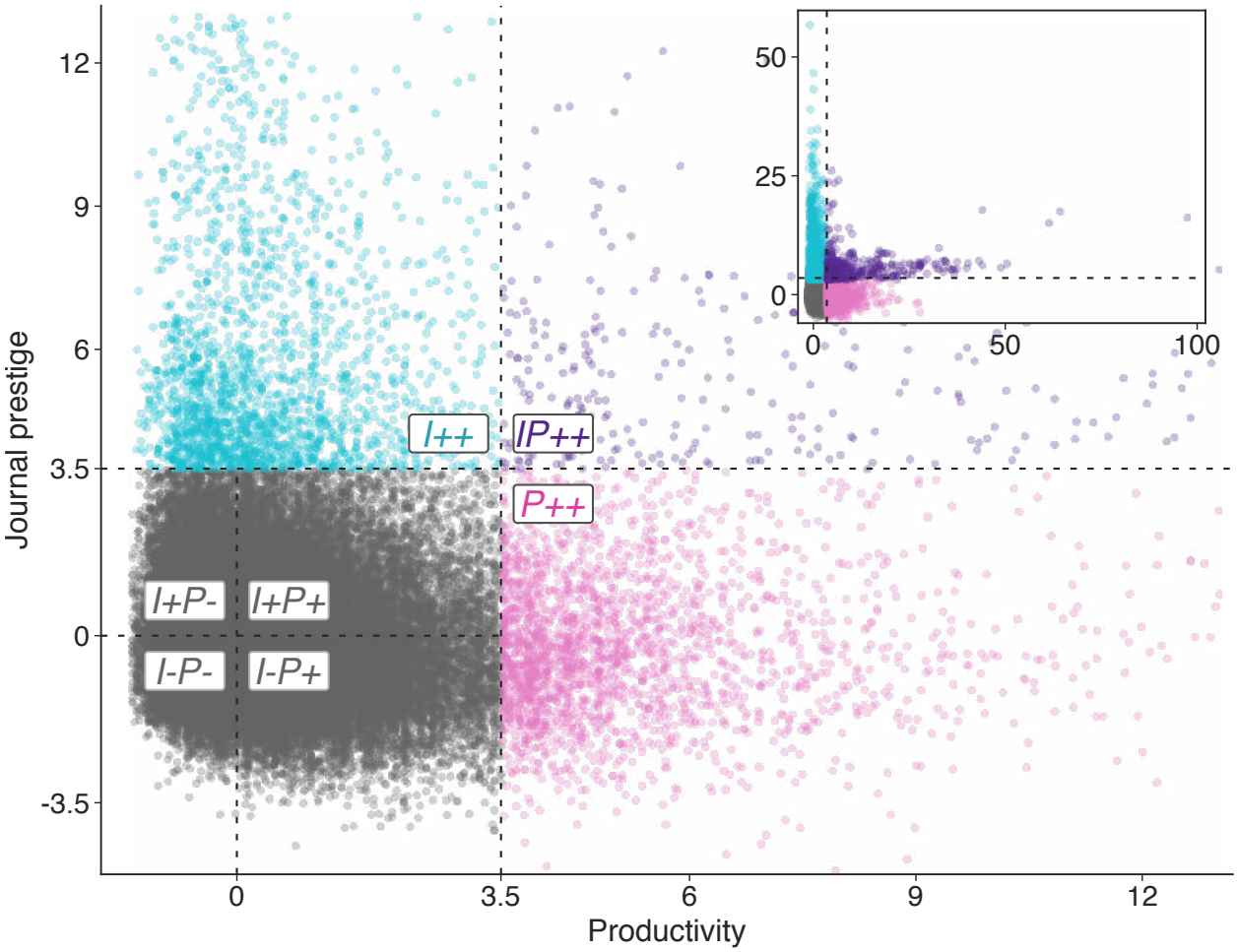

Figura 2.6: Plano prestígio de jornal *versus* produtividade em unidades padronizadas. A inserção mostra o intervalo completo do plano. Os marcadores representam anos de carreira de pesquisadores de 14 disciplinas em nosso estudo. O plano é dividido em sete setores. Três setores representam anos de carreira com níveis altíssimos de performance em impacto de jornal ($I{+}{+}$), produtividade ($P{+}{+}$) ou em ambos os indicadores ($IP{+}{+}$). Quatro setores não *outliers* representam anos de carreira com produtividade e impacto de jornal acima ($I{+}P{+}$) ou abaixo ($I{-}P{-}$) da média, impacto de jornal abaixo e produtividade acima da média ($I{-}P{+}$) e impacto de jornal acima e produtividade abaixo da média ($I{+}P{-}$).

outlier apenas em relação a prestígio de jornal ($I > 3.5$ e $P < 3.5$) e produtividade ($I < 3.5$ e $P > 3.5$), respectivamente. Para além da divisão entre *outliers*, separamos o setor não *outlier* ($I < 3.5$ e $P < 3.5$) em quatro outros setores: $I{+}P{+}$ para anos de carreira com prestígio de jornal e produtividade acima da média ($I > 0$ e $P > 0$); $I{+}P{-}$ para anos de carreira com prestígio de jornal acima e produtividade abaixo da média ($I > 0$ e $P < 0$); $I{-}P{+}$ para anos de carreira com prestígio de jornal abaixo e produtividade acima da média ($I < 0$ e $P > 0$); e $I{-}P{-}$ para anos de carreira com prestígio de jornal e produtividade abaixo da média ($I < 0$ e $P < 0$).

2.4 Pesquisadores *outliers* e não *outliers*

Uma das características mais marcantes do plano mostrado na Figura 2.6 é a existência de pesquisadores que, além de serem considerados parte da elite de pesquisadores brasileiros, se destacam exibindo níveis extremamente altos de produtividade ou prestígio de jornal (ou ambos) em anos específicos de suas carreiras. Os anos *outliers* da carreira são relativamente raros e representam apenas 7.7% do total de anos de carreira em nossa base de dados (76 454), como mostra a Figura 2.7. Entre os setores *outliers*, os números de anos de carreira em $P{+}{+}$ e $I{+}{+}$ representam 47% e 46% do total, respectivamente. Consequentemente, anos de carreira no setor $IP{+}{+}$ são ainda mais raros e correspondem a apenas 7% do total de anos *outliers*. Resultados similares foram obtidos para o conjunto de dados SJR (Figura A.6).

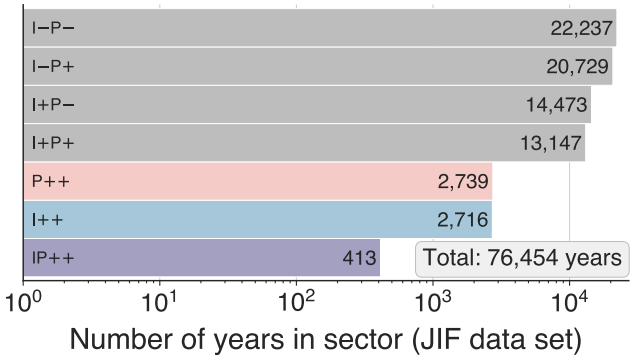

Figura 2.7: Demografia do plano prestígio de jornal *versus* produtividade para o conjunto de dados JIF. As barras mostram o número de anos de carreira em cada setor do plano prestígio de jornal *versus* produtividade.

Anos *outliers* também representam apenas uma pequena fração das carreiras dos pesquisadores da nossa base de dados, como pode ser observado na distribuição de probabilidade da Figura 2.8A. Mais de 47.6% desses pesquisadores são *outliers* em produtividade ou prestígio de jornal (ou ambos) em apenas um ano. Ainda, apenas 6.7% desses pesquisadores têm mais do que 50% de seus anos de carreira em setores *outliers*. O diagrama de Venn na Figura 2.8B ilustra o conjunto de relações entre pesquisadores categorizados como não *outliers* (todos os anos de carreira em setores não *outliers*), perfeccionistas (ao menos um ano de carreira no setor $I{+}{+}$), hiperprolíficos (ao menos um ano de carreira no setor $P{+}{+}$) e hiperprolífico-perfeccionistas (ao menos um ano de carreira no setor $IP{+}{+}$). Cerca de 30% de todos os pesquisadores conseguem ter ao menos um ano da carreira em setores *outliers*. Não existe pesquisador com todos os anos de carreira apenas no setor $IP{+}{+}$, $I{+}{+}$ ou $P{+}{+}$. Além disso, apenas sete pesquisadores (um químico, um agrônomo e cinco físicos) têm todos os anos de carreira cobertos por nosso conjunto de dados nos três setores *outliers*. Resultados similares foram encontrados para o conjunto de dados SJR (Figura A.7).

Dentre os 1 817 pesquisadores *outliers*, 1 556 (85.6%) são apenas hiperprolíficos ou apenas perfeccionistas ao longo de suas carreiras. Esse resultado indica que a maioria dos

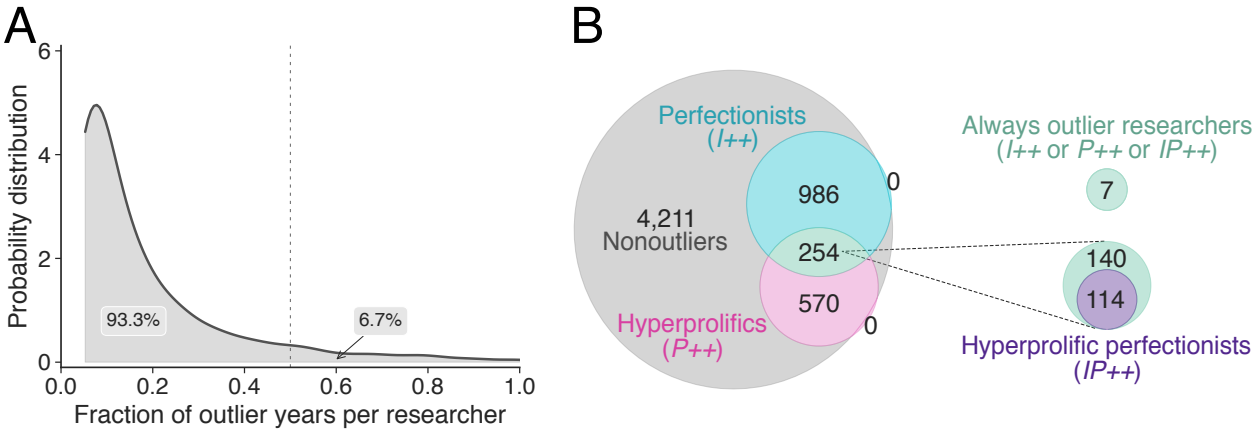

Figura 2.8: (A) Distribuição de probabilidade da fração de anos *outliers* na carreira de pesquisadores para o conjunto de dados JIF. (B) Diagrama de Venn mostrando o conjunto de relações entre as quatro categorias de pesquisadores (não *outliers*, perfeccionistas, hiperprolíficos e simultaneamente perfeccionistas e hiperprolíficos).

pesquisadores *outliers* apresenta um comportamento persistente quando são hiperprolíficos ou perfeccionistas. A existência de apenas 121 pesquisadores (6.7% dos *outliers*) simultaneamente *outliers* em ambas categorias (isto é, no setor *IP*++) corrobora essa clara distinção entre hiperprolíficos e perfeccionistas. Um padrão semelhante foi observado por Bornmann e Tekles [37] para a associação entre produtividade e o número de artigos no *top*-1% mais citados. Nosso resultado indica que é extremamente difícil publicar frequentemente em jornais de alto prestígio e, simultaneamente, manter altíssimos níveis de produtividade. De modo intrigante, observamos que anos de carreira extremamente hiperprolíficos ($P > 27.7$) estão todos no setor *IP*++. Esse resultado mostra que, apesar de muito raros, existem dezesseis pesquisadores capazes de manter performances em níveis elevadíssimos de produtividade e prestígio de jornal.

Para reforçar esse resultado, usamos uma regressão logística (veja a Seção 1.1) para estimar o efeito de anos hiperprolíficos na probabilidade de performar como um pesquisador perfeccionista. Nosso modelo é definido como

$$\Pi_{\text{perfectionist}} = \frac{e^{\alpha_0+\alpha_1 Y_P}}{1 - e^{\alpha_0+\alpha_1 Y_P}},$$

em que $\Pi_{\text{perfectionist}}$ é a probabilidade de ser um pesquisador perfeccionista dado que o acadêmico tem Y_P anos *outliers* em produtividade na carreira, α_0 é o intercepto e α_1 é o coeficiente de regressão logística. Valores positivos de α_1 indicam que um aumento em Y_P aumenta a probabilidade de performar como perfeccionista, enquanto valores negativos de α_1 indicam que um aumento em Y_P reduz a probabilidade de ser perfeccionista. Ajustamos esse modelo (conforme implementado no pacote de Python *statsmodels* [81]) aos nossos dados considerando todos os pesquisadores que foram *outliers* em impacto de jornal ou produtividade pelo menos em algum ano de suas carreiras. A Figura 2.9 mostra a probabilidade

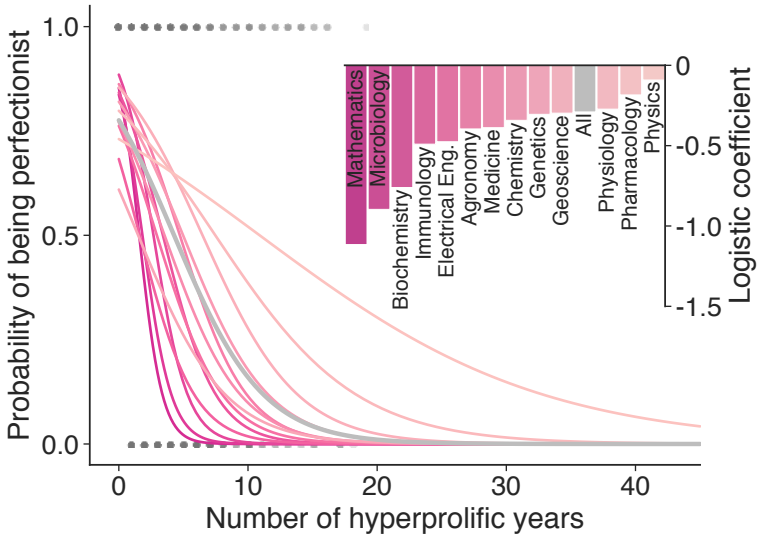

Figura 2.9: Probabilidade de ser um pesquisador perfeccionista tendo um determinado número de anos da carreira no setor hiperprolífico ($P{+}{+}$) estimada via regressão logística. A inserção mostra os coeficientes logísticos. As curvas e barras coloridas referem-se a diferentes disciplinas, enquanto a curva e a barra em cinza representam o resultado ao agregar todas as disciplinas. A disciplina de engenharia dos materiais (omitida nesse painel) é a única disciplina que não apresenta uma associação significativa.

de ser um pesquisador perfeccionista como função do número de anos hiperprolíficos e os respectivos coeficientes logísticos ao considerar todas as disciplinas conjuntamente e separadamente. A disciplina de engenharia dos materiais não mostra uma associação significativa (p-valor > 0.05), sendo assim omitida da Figura 2.9. Para as outras treze disciplinas e todas as disciplinas agregadas, os coeficientes são significativos e negativos, estabelecendo que um aumento no número de anos hiperprolíficos diminui a chance de performar como um pesquisador perfeccionista. Entretanto, esse efeito varia consideravelmente entre as disciplinas. Por exemplo, enquanto cinco anos hiperprolíficos praticamente evitam a existência de pesquisadores perfeccionistas na matemática, existe uma probabilidade de 63.2% de ser perfeccionista para o mesmo número de anos hiperprolíficos na física. Para o conjunto de dados SJR, 23 de 25 disciplinas mostram uma associação negativa e significativa entre o número de anos hiperprolíficos e a probabilidade de performar como perfeccionista (Figura A.5C), corroborando a existência de uma associação negativa entre esses dois comportamentos.

Os 261 pesquisadores que conseguem publicar como perfeccionistas e hiperprolíficos (simultaneamente ou não) é significantemente mais produtivo do que aqueles exclusivamente hiperprolíficos (z-score de produtividade de 2.71 ± 0.08 *versus* 2.06 ± 0.03; p-valor $< 10^{-16}$, teste de permutação) e exclusivamente perfeccionistas (z-score de produtividade de 2.71 ± 0.08 *versus* 0.54 ± 0.02; p-valor $< 10^{-16}$, teste de permutação). Além disso, esse grupo de pesquisadores publica em jornais de maior prestígio do que os exclusivamente hiperprolíficos (z-score médio do JIF de 1.89 ± 0.05 *versus* 0.23 ± 0.02; p-valor $< 10^{-16}$, teste de permutação) e exclusivamente perfeccionistas (z-score médio do JIF de 1.89 ± 0.05 *versus* 1.45 ± 0.02;

p-valor $< 10^{-16}$, teste de permutação). Encontramos resultados similares para o conjunto de dados SJR.

Quantificamos se pesquisadores *outliers* têm certa preferência por determinado setor *outlier*. Com esse fim, para cada pesquisador *outlier* em mais de uma categoria, consideramos apenas anos de carreira em setores *outliers*, estimamos as frações p_i em cada setor i dos n setores e calculamos a entropia normalizada de Shannon [115] $h = -\frac{1}{\log n}\sum_{i=1}^{n} p_i \log p_i$. Valores de entropia perto de um representam comportamento alternante, enquanto valores perto de zero indicam preferência por um dado setor *outlier*. A Figura 2.10 mostra que a distribuição dos valores da entropia para todos os pesquisadores *outliers* tem um pico ao redor de 0.6 (curva em roxo), sugerindo uma preferência por determinados setores *outliers*. No entanto, se não considerarmos o setor *IP++* (o setor menos povoado), a distribuição de entropia desloca na direção de valores mais elevados com pico próximo de um (curva em verde). Portanto, podemos inferir que não existe preferência entre os setores *I++* e *P++* para pesquisadores publicando em ambos os setores. Nesse aspecto, esses pesquisadores atípicos não são tão diferentes daqueles presentes apenas em setores não *outliers*. Como mostra a Figura 2.10 (curva em cinza), pesquisadores não *outliers* também não exibem uma forte preferência por qualquer setor ao longo de suas carreiras. Os mesmos padrões são observados para o conjunto de dados SJR (Figura A.5D).

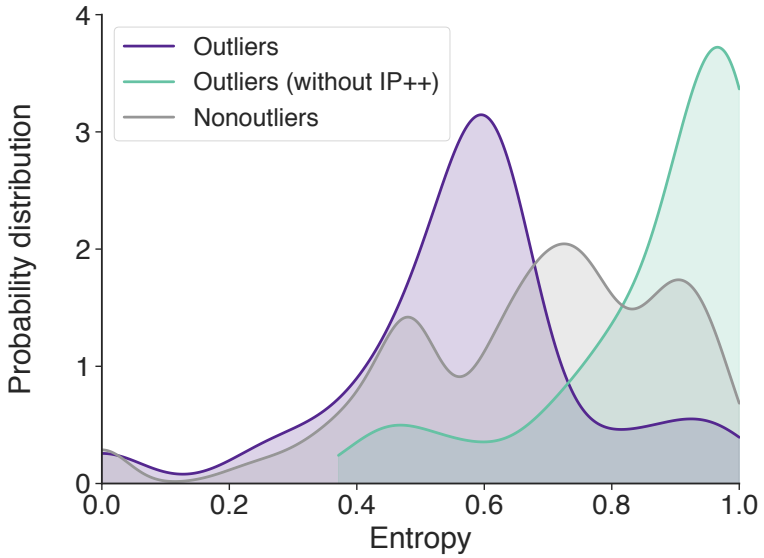

Figura 2.10: Distribuição de probabilidade da entropia normalizada de Shannon associada à ocupação dos setores do plano para as carreiras individuais dos pesquisadores. A curva em roxo mostra a entropia associada à ocupação de setores *outliers* por pesquisadores *outliers*, enquanto a curva em verde representa o mesmo mas ignorando o setor *IP++*. A curva em cinza mostra a distribuição da entropia para pesquisadores não *outliers*.

Outra questão intrigante é: existem transições mais frequentes entre setores do plano prestígio de jornal *versus* produtividade ao longo da carreira dos pesquisadores? Para investigar essa hipótese, estimamos o número de transições entre setores do plano (N_t^{rh}, com

$r, h \in \{IP{+}{+}, I{+}{+}, P{+}{+}, I{\pm}P{\pm}, I{\pm}P{\mp}\}$). Em seguida, definimos um modelo nulo como o número médio de transições entre setores do plano após misturar aleatoriamente as carreiras dos pesquisadores em 10 000 realizações ($\bar{N}_{ts}^{rh}$, com $r, h \in \{IP{+}{+}, I{+}{+}, P{+}{+}, I{\pm}P{\pm}, I{\pm}P{\mp}\}$). A partir desse processo, estimamos o excesso relativo para todas as transições possíveis via

$$\text{Excesso relativo} = \frac{N_t^{rh} - \bar{N}_{ts}^{rh}}{\bar{N}_{ts}^{rh}}\,.$$

A Figura 2.11 mostra duas matrizes de transições referentes a pesquisadores *outliers* e não *outliers*. A primeira característica que observamos é a simetria das matrizes, indicando que a maioria das transições não tem direção preferencial. Além disso, os elementos diagonais positivos estão entre os maiores valores absolutos. Em outras palavras, permanecer no mesmo setor é a principal tendência de curto prazo. Para pesquisadores *outliers*, as transições $IP{+}{+}{\bullet}{\rightarrow}IP{+}{+}$, $I{+}{+}{\bullet}{\rightarrow}I{+}{+}$ e $P{+}{+}{\bullet}{\rightarrow}P{+}{+}$ têm os maiores excessos dentre todas as autotransições. Para pesquisadores não *outliers*, $I{+}P{+}{\bullet}{\rightarrow}I{+}P{+}$ e $I{-}P{+}{\bullet}{\rightarrow}I{-}P{+}$ são as autotransições com os maiores excessos. Curiosamente, a autotransição $I{-}P{-}{\bullet}{\rightarrow}I{-}P{-}$ (setor de menor prestígio e menor produtividade) tem um excesso que é maior para pesquisadores *outliers* (23%) do que para pesquisadores não *outliers* (7%).

As transições entre setores não *outliers* são marcadas por um excesso negativo quando existe uma mudança simultânea de níveis de produtividade e impacto de jornal ($I{+}P{\pm}{\leftrightarrow}I{-}P{\mp}$). Essas transições, representadas pelos elementos antidiagonais da porção não *ou-*

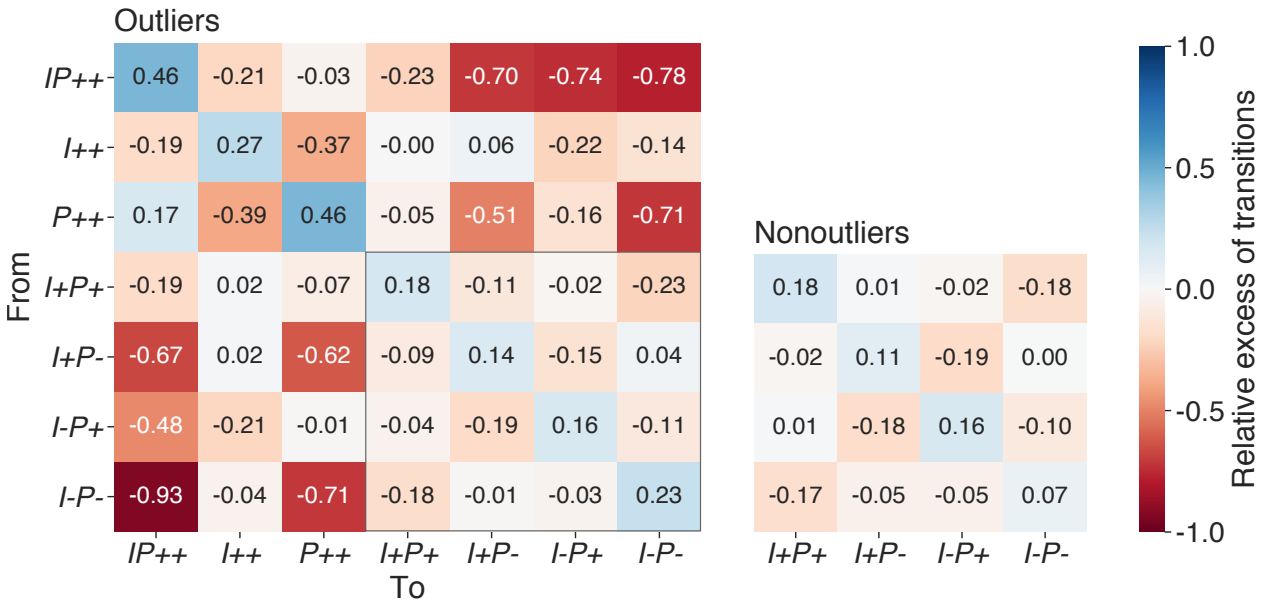

Figura 2.11: Matriz de transição entre setores do plano prestígio de jornal *versus* produtividade para pesquisadores *outliers* (esquerda) e não *outliers* (direita). Cada célula representa o excesso relativo de transições entre dois setores comparado com o modelo nulo. O modelo nulo fornece os valores médios de excesso para versões embaralhadas das carreiras dos pesquisadores considerando 10 000 realizações.

tlier das matrizes, são menos frequentes ao longo das carreiras de pesquisadores *outliers* e não *outliers*. Um padrão semelhante é observado para transições envolvendo setores *outliers* $I{+}{+}$ e $P{+}{+}$, ou seja, as transições $I{+}{+}{\leftrightarrow}P{+}{+}$, $P{+}{+}{\bullet\!\!\rightarrow}I{+}P{-}$ e $P{+}{+}{\bullet\!\!\rightarrow}I{-}P{-}$ são menos frequentes ao longo das carreiras de pesquisadores *outliers*. De maneira diferente, transições entre setores com níveis similares de produtividade ou prestígio de jornal (por exemplo, $I{+}P{+}{\leftrightarrow}I{-}P{+}$ e $I{-}P{-}{\leftrightarrow}I{+}P{-}$) geralmente têm excessos perto de zero e são, assim, tão frequentes quanto aquelas ocorrendo no modelo nulo. Juntamente com o excesso das autotransições, esses resultados sugerem uma aversão a mudanças simultâneas nos níveis de produtividade e prestígio de jornal, além de uma preferência pela manutenção desses níveis em anos consecutivos das carreiras dos pesquisadores.

Observamos que transições entre setores *outliers* e não *outliers* ocorrem muito menos ou tão frequentemente quanto ao acaso (excessos negativos ou perto de zero). Anos da carreira no setor $P{+}{+}$ usualmente não são precedidos nem seguidos por anos em setores de baixa produtividade ($I{+}P{-}$ e $I{-}P{-}$). Anos da carreira no setor $I{+}{+}$ são menos precedidos e seguidos por anos em setores com baixo prestígio de jornal ($I{-}P{+}$ e $I{-}P{-}$). Verificamos também que anos da carreira no setor $IP{+}{+}$ são mais frequentemente precedidos por anos no setor $P{+}{+}$ do que por anos no setor $I{+}{+}$, sugerindo que é mais fácil para hiperprolíficos se tornarem hiperprolífico-perfeccionistas do que para pesquisadores perfeccionistas.

No geral, encontramos resultados similares para o conjunto de dados SJR (Figura A.5E). As principais diferenças emergem para as transições envolvendo o setor $IP{+}{+}$. Fora da diagonal, as duas transições com maior excesso para pesquisadores *outliers* são $IP{+}{+}{\rightarrow}I{+}{+}$ e $IP{+}{+}{\rightarrow}P{+}{+}$, com excessos de 14% e 12%, respectivamente. Esse resultado sugere que anos $I{+}{+}$ e $P{+}{+}$ são comumente precedidos por anos $IP{+}{+}$ quando se considera o SJR como medida de prestígio de jornal. Além disso, apesar de o setor $IP{+}{+}$ ainda ser frequentemente mais precedido por anos hiperprolíficos ($P{+}{+}$) do que anos perfeccionistas ($I{+}{+}$), com 7% *versus* -3%, respectivamente, a diferença não é tão substancial quanto para o conjunto de dados JIF (17% *versus* -19%). Todas as outras transições apresentam aproximadamente o mesmo comportamento. Para testar a robustez dessas comparações, constatamos que os resultados para o conjunto de dados SJR são estáveis ao considerar apenas as disciplinas presentes na base de dados JIF (Figura A.8).

2.5 Efeitos do ano da carreira

Investigamos os efeitos do ano da carreira acadêmica no prestígio de jornal e na produtividade. Com esse objetivo, consideramos o ano após o doutoramento como o primeiro ano da carreira dos pesquisadores. Em seguida, calculamos os valores médios da produtividade e do prestígio de jornal em janelas móveis de 5 anos a partir do agrupamento das carreiras de cada disciplina. A Figura 2.12 mostra os valores médios como função do ano da carreira

dos pesquisadores. Observamos uma tendência crescente na produtividade média com a progressão da carreira para todas as disciplinas, seguida por um platô ou pequeno decréscimo no período final da carreira. Para o prestígio de jornal, observamos que esses valores médios são levemente maiores durante os primeiros anos da carreira e apresentam uma tendência decrescente sutil para a maioria das disciplinas. A Figura 2.13 mostra as estimativas das taxas de crescimento ou decrescimento da produtividade e do prestigío médio de jornal com a progressão da carreira calculadas a partir de regressões lineares com o dado agregado. As Figuras A.9 e A.10 mostram resultados similares para o conjunto de dados SJR. Entretanto, é importante pontuar que as tendências médias para as disciplinas podem não representar o comportamento individual dos pesquisadores como discutiremos mais detalhadamente no próximo capítulo.

Para continuar caracterizando os efeitos do ano da carreira na produtividade e prestígio de jornal, dividimos as carreiras acadêmicas em intervalos de cinco anos e estimamos a fração média dos anos das carreiras em cada setor do plano prestígio de jornal *versus* produtividade como uma função do ano da carreira. A Figura 2.14 mostra essas frações para todas as disciplinas de nossa investigação. Nessa representação matricial, colunas indicam intervalos da carreira, linhas indicam diferentes planos do setor e os códigos de cores representam a magnitude das frações. Em virtude da diferença no estágio da carreira dos pesquisadores do nosso conjunto de dados, essa análise abrange um intervalo temporal em anos de carreira

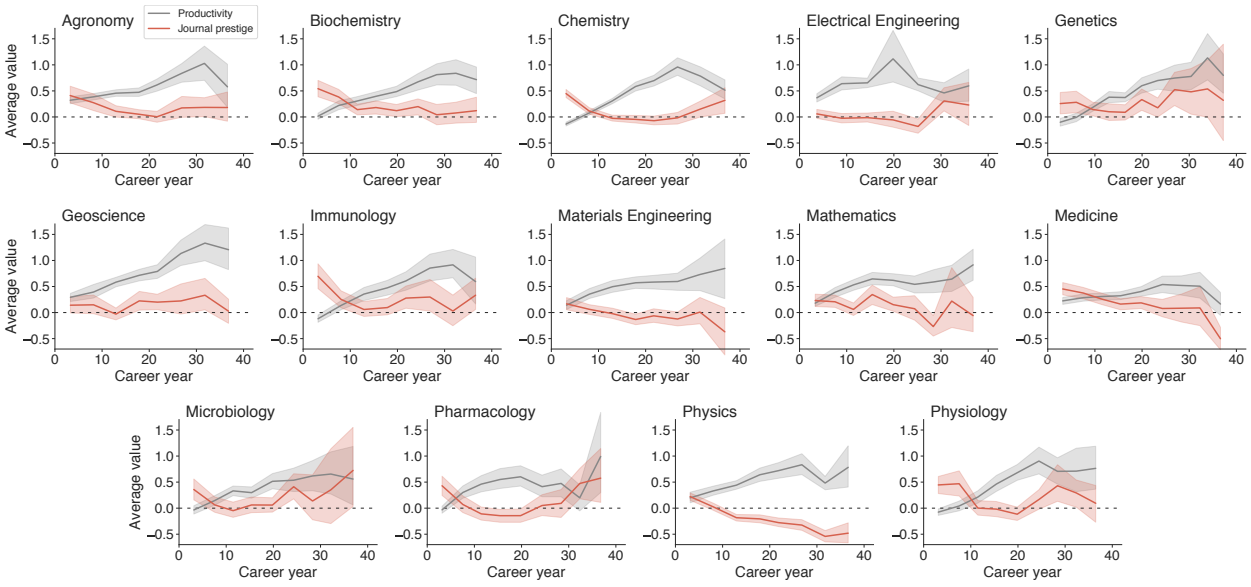

Figura 2.12: Valores médios da produtividade e do impacto de jornal ao longo da carreira dos pesquisadores para diferentes disciplinas considerando o conjunto de dados JIF. Essas visualizações mostram os valores médios da produtividade (curva em cinza) e do prestígio de jornal (curva em vermelho) calculados a partir de médias móveis de 5 anos ao longo dos anos da carreira para cada disciplina do conjunto de dados JIF. As regiões sombreadas correspondem a intervalos de confiança de 95% obtidos pelo método de *bootstrap*.

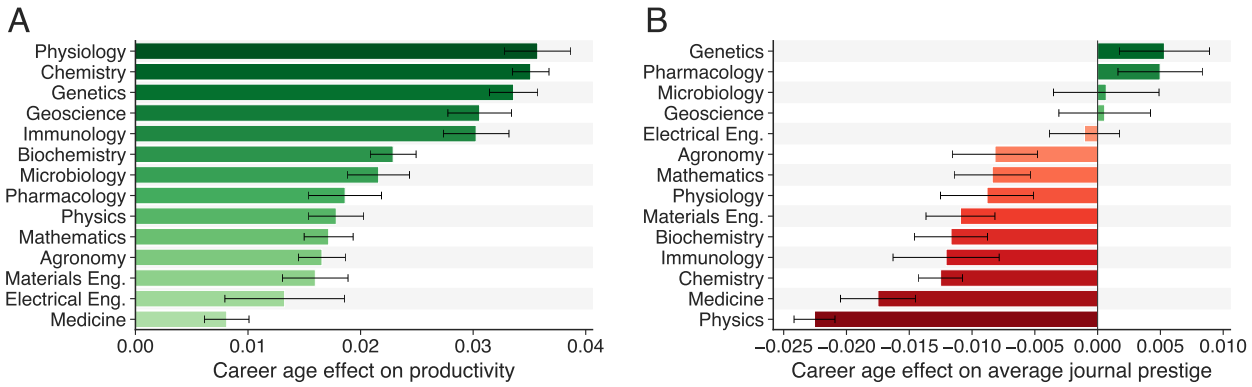

Figura 2.13: Efeito do ano da carreira na produtividade e no prestígio de jornal para diferentes disciplinas considerando o conjunto de dados JIF. Os gráficos de barra mostram o efeito do ano da carreira na (A) produtividade e no (B) prestígio de jornal para cada disciplina no conjunto de dados JIF. Estimamos os valores por meio de um modelo linear da associação média entre idade da carreira e produtividade e entre idade da carreira e prestígio de jornal (Figura 2.12) para cada disciplina. As barras de erro indicam o erro padrão dos coeficientes lineares.

maior do que o número de anos no conjunto de dados JIF (19 anos).

A Figura 2.14 indica que as tendências de ocupação no plano prestígio de jornal *versus* produtividade variam entre as disciplinas (veja a Figura A.11 para comparação com o conjunto de dados SJR). Contudo, alguns padrões de evolução são comuns. Para setores não *outliers*, observamos uma concentração em setores de baixa produtividade ($I{+}P{-}$ e $I{-}P{-}$) durante os anos iniciais da carreira e uma tendência de mudança para setores de alta produtividade ($I{+}P{+}$ e $I{-}P{+}$) em estágios posteriores da carreira de pesquisadores da maioria das disciplinas. Essa tendência é particularmente evidente na física e na química, para as quais observamos um crescimento mais pronunciado no setor $I{-}P{+}$. Para setores *outliers*, notamos uma baixa prevalência no setor $P{+}{+}$ durante estágios iniciais e uma tendência de aumento das frações em estágios posteriores para todas as disciplinas. O crescimento no nível de produtividade com o passar do tempo pode refletir a consolidação da carreira dos pesquisadores e o provável crescimento de suas redes de colaborações científicas. Além disso, os padrões para pesquisadores não *outliers* e *outliers* concordam com a tendência geral crescente na produtividade média para todas as disciplinas observada na Figura 2.12.

De modo oposto, é intrigante observar que o setor $I{+}{+}$ tende a ser mais povoado nos estágios iniciais da carreiras dos pesquisadores – um resultado que pode parcialmente explicar o valor médio ligeiramente maior do prestígio de jornal nos primeiros anos de carreira para a maioria das disciplinas, como mostra a Figura 2.12. Esse comportamento não apenas indica que é mais provável se tornar um *outlier* de impacto nos anos iniciais da carreira, bem como indica que pesquisadores mais jovens (com carreiras mais curtas) podem apresentar performance *outlier* nessa categoria mais frequentemente. Para investigar esse aspecto, utilizamos um modelo logístico para estimar a probabilidade de ser perfeccionista como uma função do

Figura 2.14: Tendências de ocupação do plano prestígio de jornal *versus* produtividade ao longo das carreiras dos pesquisadores considerando o conjunto de dados JIF. Os painéis mostram a fração dos anos das carreiras em cada setor não *outlier* e nos setores *outliers* $I{+}{+}$ e $P{+}{+}$ como uma função do ano da carreira dos pesquisadores de 14 disciplinas no conjunto de dados JIF. As colunas indicam intervalos de 5 anos e as linhas representam os diferentes setores. O código de cor indica as frações para setores não *outliers* (tons de cinza) e setores *outliers* para os setores $I{+}{+}$ (tons de azul) e $P{+}{+}$ (tons de rosa). O setor $IP{+}{+}$ foi omitido uma vez que anos de carreira nesse setor são muito raros. Apenas intervalos de 5 anos com pelo menos 20 pesquisadores são mostrados nessas visualizações.

comprimento da carreira dos pesquisadores (L). Nesse caso, o modelo pode ser escrito como

$$\Pi_{\text{perfectionist}} = \frac{e^{\theta_0+\theta_1 L}}{1-e^{\theta_0+\theta_1 L}},$$

em que θ_0 é o intercepto e θ_1 é coeficiente da regressão logística. Ajustamos esse modelo considerando todos os pesquisadores *outliers* nos conjuntos de dados JIF e SJR. A Figura 2.15 mostra $\Pi_{\text{perfectionist}}$ como função de L para o conjunto de dados JIF. Os parâmetros ajustados são $\theta_0 = 1.849 \pm 0.132$ e $\theta_1 = -0.051 \pm 0.006$ para o conjunto de dados JIF e $\theta_0 = 1.921 \pm 0.108$ e $\theta_1 = -0.054 \pm 0.005$ para o conjunto de dados SJR (Figura A.12). Em um exemplo

concreto, entre os pesquisadores *outliers*, a chance de encontrar pesquisadores perfeccionistas diminui de 79% para 58% quando o comprimento da carreira aumenta de 10 para 30 anos. É importante mencionar que a tendência de exibição de altos níveis de prestígio de jornal no início da carreira pode refletir um efeito de seleção já que nosso conjunto de dados inclui apenas pesquisadores pertencentes à elite científica brasileira. Os resultados para o conjunto de dados SJR (Figura A.12) corroboram esse resultado e indicam tendências muito similares não apenas para disciplinas presentes em ambos os conjuntos de dados mas também para disciplinas exclusivas do conjunto de dados SJR.

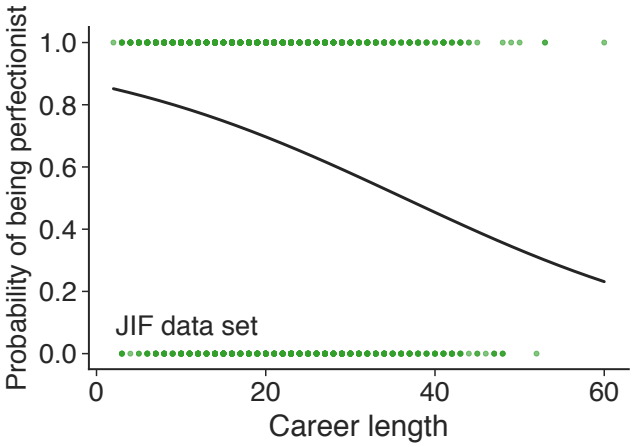

Figura 2.15: Efeito do comprimento da carreira na probabilidade de ser perfeccionista para o conjunto de dados JIF. Estimamos a probabilidade de ser perfeccionista como uma função do comprimento da carreira do pesquisador via modelo logístico (veja a Seção 1.1).

2.6 Quantificando o efeito da produtividade no prestígio de jornal

Apesar de nossos resultados indicarem uma associação negativa entre produtividade e prestígio de jornal em níveis altíssimos de ambas quantidades para a maioria dos pesquisadores, ainda precisamos investigar como essa relação se expressa para pesquisadores que nunca acessaram os setores *outliers*. Os acadêmicos não *outliers* representam a maior parte de nosso conjunto de dados, totalizando 70% dos pesquisadores. Individualmente, esses cientistas exibem comportamentos heterogêneos na associação entre a produtividade e impacto de jornal, o que acaba por limitar a emergência de uma clara associação no nível de disciplina em modelos lineares simples. Assim, para estimar o efeito da produtividade no impacto de jornal de pesquisadores não *outliers* levando em conta os diversos padrões individuais, aplicamos um modelo hierárquico bayesiano (veja a Seção 1.2) selecionando apenas anos produtivos de pesquisadores não *outliers* com carreiras maiores do que cinco anos (Tabelas 2.1

e A.1). Dada a disciplina k, consideramos que os dados estão estruturados hierarquicamente de tal forma que cada observação I_j e P_j pertence ao pesquisador j (suprimimos o índice k por simplicidade). Assumimos uma relação linear entre essas variáveis no nível individual, para a qual c_j e β_j são, respectivamente, o intercepto e a inclinação da associação linear do j-ésimo pesquisador. Consideramos os parâmetros c_j e β_j como variáveis aleatórias distribuídas de acordo com distribuições normais cujos parâmetros também são variáveis aleatórias. Em notação matemática, podemos escrever esse modelo como

$$I_j \sim \mathcal{N}(c_j + \beta_j P_j, \varepsilon)\,, \tag{2.1}$$

em que $\mathcal{N}(\mu, \sigma)$ representa uma distribuição normal com média μ e desvio padrão σ, ε leva em consideração os determinantes não observáveis de I_j e

$$\begin{aligned} c_j &\sim \mathcal{N}(\mu_c, \sigma_c) \\ \beta_j &\sim \mathcal{N}(\mu_P, \sigma_P) \end{aligned},$$

em que μ_c é a média e σ_c é o desvio padrão da distribuição normal associada com o intercepto c_j e μ_P e σ_P são os equivalentes para a distribuição associada a β_j. O processo de inferência bayesiana consiste em determinar as distribuições de probabilidade *a posteriori* dos parâmetros no nível da disciplina (μ_c, σ_c, μ_P e σ_P) e no nível do pesquisador (c_j e β_j para cada pesquisador j de dada disciplina).

Realizamos a regressão bayesiana para cada área separadamente e usamos distribuições a *priori* não informativas [124] a fim de não enviesar a estimativa da *posteriori*, isto é, consideramos

$$\begin{aligned} \varepsilon &\sim \mathcal{U}(0, 10^2) \\ \mu_c &\sim \mathcal{N}(0, 10^5) \\ \mu_P &\sim \mathcal{N}(0, 10^5) \\ \sigma_c &\sim \text{Inv-}\Gamma(10^{-3}, 1) \\ \sigma_P &\sim \text{Inv-}\Gamma(10^{-3}, 1) \end{aligned}\quad, \tag{2.2}$$

em que $\mathcal{U}(x_{\min}, x_{\max})$ representa uma distribuição uniforme entre $x_{\min}$ e $x_{\max}$ e Inv-$\Gamma(a, b)$ representa uma distribuição gama inversa com parâmetros a (forma) e b (escala). A Figura 2.16A mostra uma representação gráfica desse modelo.

Implementamos o modelo utilizando o pacote `PyMC3` via amostrador de Monte Carlo Hamiltoniano NUTS (*No-U-Turn-Sampler*, veja as seções 1.3 e 1.4) para amostrar as distribuições a *posteriori*. Utilizamos 8 cadeias paralelas com 10 000 iterações (das quais 5 000 eram amostras de aquecimento) para permitir uma boa mistura das cadeias do amostrador de Monte Carlo. Estimamos as estatísticas de convergência de Gelman-Rubin (R chapéu,

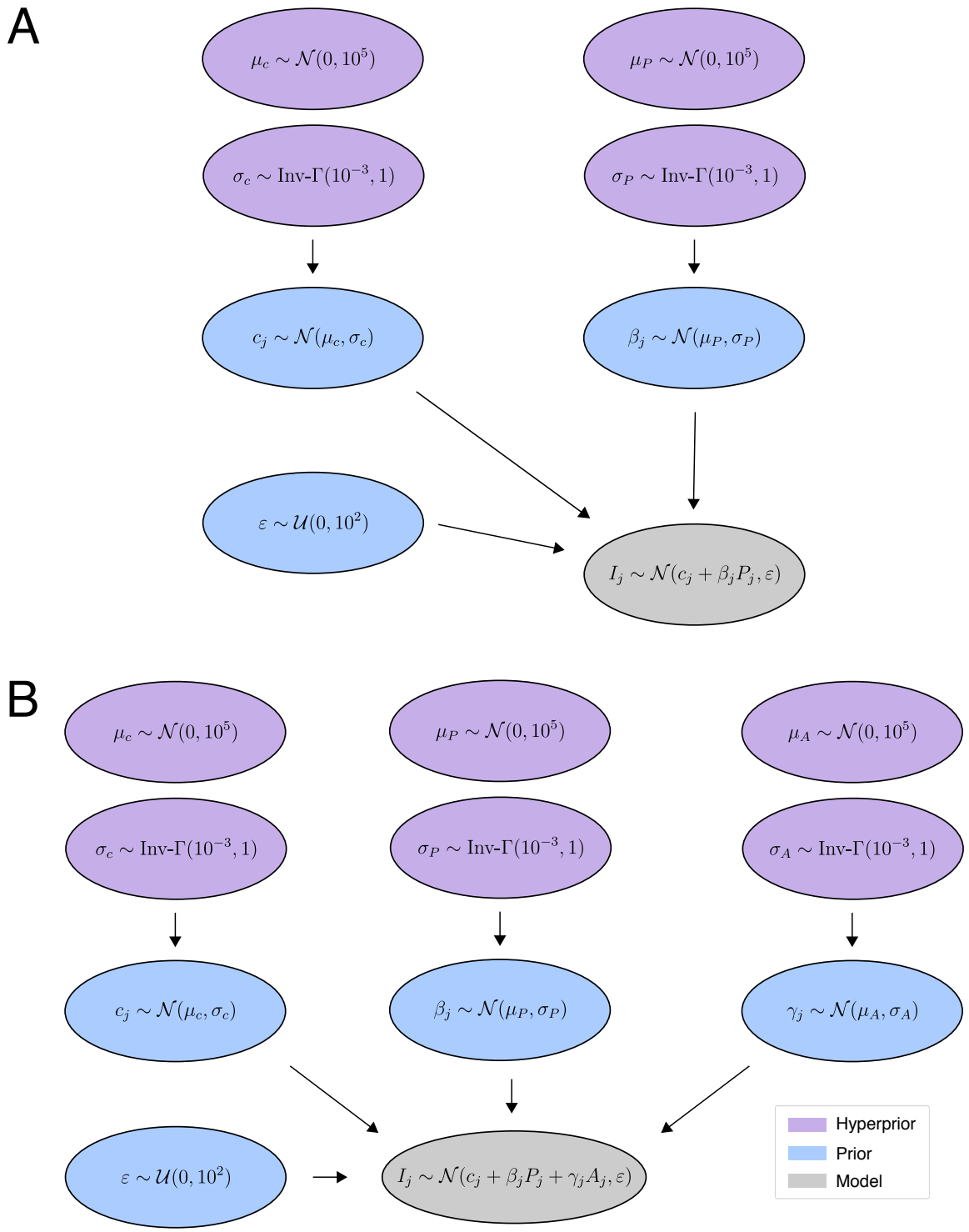

Figura 2.16: Representação visual dos modelos hierárquicos bayesianos. (A) Descrição esquemática do modelo hierárquico bayesiano da Equação 2.1 usado para estimar o efeito da produtividade no prestígio de jornal para pesquisadores não *outliers*. (B) Descrição esquemática do modelo hierárquico bayesiano da Equação 2.3 usado para estimar o efeito da produtividade e do ano da carreira no prestígio de jornal para pesquisadores não *outliers*. Formas em roxo representam distribuições a *hiperpriori*, formas em azul representam distribuições a *priori* e a forma em cinza representa a estrutura geral do modelo hierárquico.

Tabela 2.1: Descrição do conjunto de dados JIF usado na análise bayesiana hierárquica. Número de pesquisadores e de observações para cada disciplina no conjunto de dados JIF após filtrar pesquisadores com carreiras mais curtas do que cinco anos.

Disciplina	Número de pesquisadores	Número de observações
Agronomia	462	4 523
Bioquímica	258	3 482
Engenharia Elétrica	232	2 302
Engenharia de Materiais	210	2 496
Farmacologia	147	2 003
Fisiologia	136	1 757
Física	686	9 348
Genética	210	2 709
Geociências	229	2 195
Imunologia	109	1 415
Matemática	212	2 128
Medicina	357	4 765
Microbiologia	131	1 670
Química	577	7 701

veja a Seção 1.4) para a regressão e os resultados foram próximos de um para todos os parâmetros, indicando a convergência do método de amostragem.

Por meio da abordagem bayesiana, estimamos a distribuição de probabilidade a *posteriori* do coeficiente linear de cada pesquisador e a distribuição de probabilidade a *posteriori* de μ_P para cada área. Dessa forma, a distribuição de μ_P representa o efeito agregado da produtividade no impacto de jornal para pesquisadores não *outliers* em cada disciplina. As distribuições de μ_P deslocadas em direção a valores positivos representam disciplinas em que a maioria dos pesquisadores apresenta uma associação positiva entre produtividade e impacto de jornal. Em contraste, distribuições deslocadas em direção a valores negativos caracterizam disciplinas em que um aumento da produtividade é correlacionado com uma queda no impacto de jornal para maioria dos pesquisadores.

A Figura 2.17A mostra que a distribuição de μ_P (curvas coloridas preenchidas) variam significativamente entre as disciplinas. Entretanto, com exceção da matemática, todas as disciplinas têm distribuições quase inteiramente localizadas em valores de μ_P menores do que zero, sugerindo uma associação negativa entre produtividade e impacto de jornal para a maioria dos pesquisadores não *outliers*. No caso mais extremo, um aumento de uma unidade de produtividade para físicos associa-se com uma diminuição de ≈ 0.242 unidades de impacto de jornal de suas publicações (em unidades padronizadas). No outro extremo, a matemática apresenta distribuição localizada perto de zero. Esse resultado indica que produtividade apresenta um papel não tão significativo no impacto de jornal para a maioria dos matemáticos mesmo que alguns deles possam demonstrar associações mais intensas (positivas ou

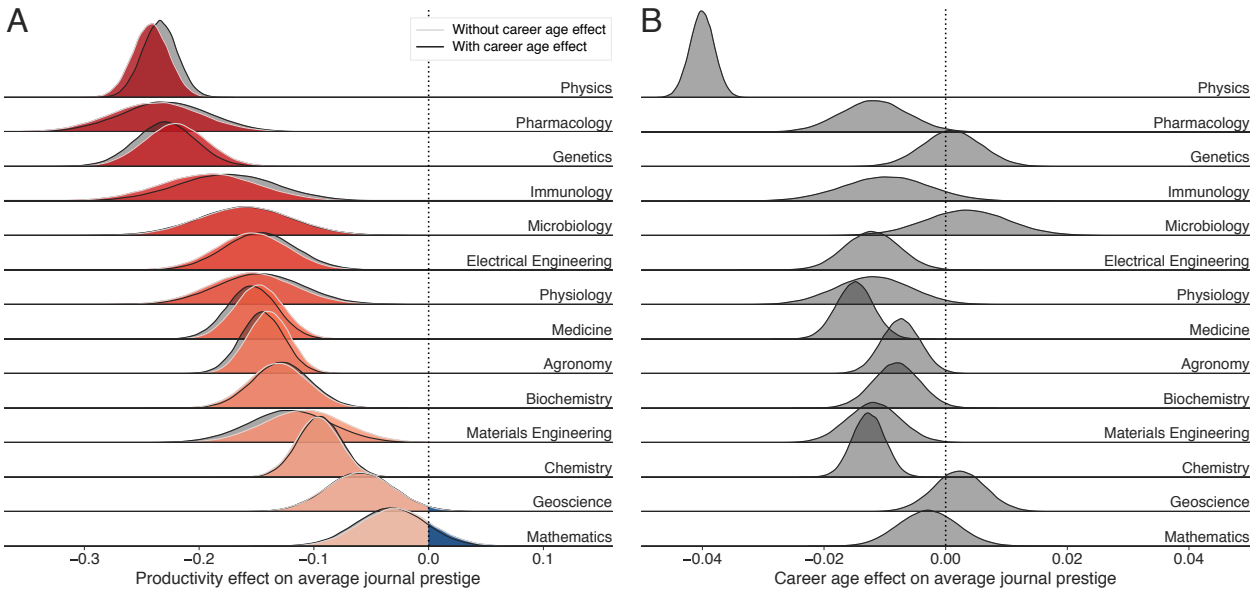

Figura 2.17: Efeito da produtividade no prestígio de jornal para pesquisadores não *outliers* considerando o conjunto de dados JIF. (A) Distribuições de probabilidade a *posteriori* do valor médio do coeficiente linear (μ_P) ao considerar a associação entre produtividade e impacto de jornal para pesquisadores não *outliers* de cada disciplina. As curvas coloridas preenchidas representam os resultados sem levar em consideração os efeitos do ano da carreira, enquanto as curvas preenchidas em cinza mostram as distribuições de μ_P após incluir o ano da carreira como fator de confusão no modelo bayesiano hierárquico. (B) Distribuições de probabilidade a *posteriori* do valor médio do coeficiente linear (μ_A) relacionado ao efeito do ano da carreira no impacto de jornal para pesquisadores não *outliers* de cada disciplina.

negativas).

Os resultados ilustrados nas Figuras 2.12 e 2.14 já demonstraram que o ano da carreira afeta os valores médios da produtividade e do prestígio de jornal quando se agregam pesquisadores por suas respectivas disciplinas. Sendo assim, podemos esperar que o ano da carreira afete a associação entre o prestígio de jornal e produtividade também no nível individual. Esse é um aspecto crítico uma vez que a associação negativa geral reportada na Figura 2.17A pode refletir uma mudança de um estágio inicial marcado por baixa produtividade e alto impacto para estágios posteriores marcados por alta produtividade e baixo impacto. Para considerar o possível efeito de confusão do ano da carreira na associação entre prestígio de jornal e produtividade, incluímos o ano da carreira como um preditor do impacto de jornal no modelo bayesiano hierárquico. Nesse caso, o ano da carreira A_j também é considerado como uma variável independente no modelo hierárquico, resultando na relação

$$I_j \sim \mathcal{N}(c_j + \beta_j P_j + \gamma_j A_j, \varepsilon)\,, \tag{2.3}$$

em que γ_j é a inclinação da associação linear entre o ano da carreira e o prestígio de jornal.

Assumimos que esse coeficiente é distribuído de acordo com uma distribuição normal

$$\gamma_j \sim \mathcal{N}(\mu_A, \sigma_A)\,,$$

em que μ_A é a média e σ_A é o desvio padrão. Ajustamos o modelo da Equação 2.3 com as mesmas distribuições não informativas a *priori* definidas na Equação 2.2 e usamos

$$\begin{aligned} \mu_A &\sim \mathcal{N}(0, 10^5) \\ \sigma_A &\sim \text{Inv-}\Gamma(10^{-3}, 1) \end{aligned}\,, \tag{2.4}$$

como as distribuições não informativas a *priori* para os parâmetros adicionais relacionados aos efeitos do ano da carreira. A Figura 2.16B mostra a representação gráfica desse modelo generalizado que leva em consideração possíveis efeitos de confusão do ano da carreira na associação entre impacto de jornal e produtividade. Também utilizamos o pacote `PyMC3` para amostrar as distribuições a *posteriori* com os mesmos métodos do modelo anterior. Novamente, as estatísticas de convergência de Gelman-Rubin de todos os parâmetros foram próximos de um, indicando sua convergência.

A Figura 2.17B mostra que distribuições de μ_A também variam entre disciplinas, apresentando valores médios negativos ou perto de zero em sua maioria. Esses resultados indicam uma redução no impacto de jornal ao longo das carreiras para maioria dos pesquisadores da maioria das disciplinas. Apesar da dificuldade em comparar diretamente os efeitos da mudança de produtividade com os efeitos da progressão da carreira, uma progressão de dez anos na carreira tem um efeito maior no prestígio de jornal do que aumentar uma unidade de produtividade (z-score) de um pesquisador típico apenas para as disciplinas química e física como mostra a Figura 2.18. Mais importante, a Figura 2.17A mostra que a distribuições de μ_P com (curvas coloridas) e sem (curvas em cinza) o efeito do ano da carreira mudam pouco. Dessa forma, o efeito de confusão do ano da carreira na associação geral negativa entre prestígio de jornal e produtividade é quase insignificante – ou seja, um aumento na produtividade associa-se com um decréscimo em prestígio de jornal independentemente do ano da carreira.

O conjunto de dados SJR (Figuras A.13 e A.14) estende essa análise para mais disciplinas e apresenta resultados semelhantes para as disciplinas presentes em ambos os conjuntos de dados.

Assim, encerramos a apresentação dos resultados referentes à investigação sobre a associação entre produtividade e impacto de jornal para diferentes disciplinas e estágios de carreira. No capítulo seguinte, apresentaremos os resultados referentes à investigação sobre padrões universais de produtividade em carreiras científicas. As discussões dos resultados desses dois capítulos serão realizadas de maneira conjunta no Capítulo 4.

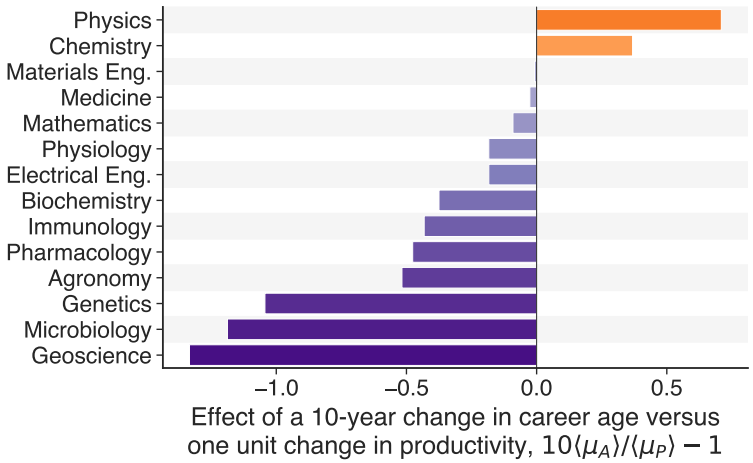

Figura 2.18: Comparação entre os efeitos do ano da carreira e produtividade no prestígio de jornal considerando o conjunto de dados JIF. As barras comparam o efeito de uma progressão de 10 anos na carreira com o efeito de aumentar uma unidade da produtividade (z-score) para um pesquisador típico de cada disciplina no conjunto de dados JIF. Esses valores representam a fração de quão maior ou menor é o efeito do ano da carreira comparado com o efeito da produtividade (isto é, $10\langle\mu_A\rangle/\langle\mu_P\rangle - 1$, em que $\langle\mu_A\rangle$ e $\langle\mu_P\rangle$ são os valores médios, respectivamente, de μ_A e μ_P para cada disciplina). Frações ao redor de zero indicam que um aumento de 10 anos na idade da carreira afeta o impacto de jornal de maneira similar ao aumento de uma unidade na produtividade. Valores positivos indicam que uma mudança de 10 anos na idade da carreira afeta mais o impacto de jornal do que o aumento de uma unidade de produtividade, enquanto valores negativos indicam que produtividade tem maior impacto no prestígio de jornal.

CAPÍTULO 3

Padrões universais de produtividade em carreiras científicas

Neste capítulo, concentramos nossa atenção sobre o indicador produtividade. Especificamente, investigamos as trajetórias acadêmicas de produtividade. Empregamos métodos de análise de séries temporais, redução de dimensionalidade e análise de redes para encontrar os padrões universais de produtividade de carreiras de cientistas brasileiros [73]. Diferentemente da abordagem empregada no capítulo anterior, em que estimamos o efeito do estágio da carreira por curvas médias, vamos considerar agora as carreiras de cada pesquisador individualmente na tentativa de determinar as diversas formas de curvas de produtividade em carreiras científicas.

3.1 Apresentação dos dados

Novamente, empregamos a Plataforma Lattes [118] como nossa fonte primária de dados. A Plataforma Lattes resolve um importante problema da literatura de ciência da ciência: a imprecisão associada a dados de carreiras científicas individuais [3]. Atualmente, não existem bases de carreiras científicas confiáveis e amplas. As bases de dados existentes são construídas manualmente, restringindo-se a determinadas disciplinas [65], ou são construídas de maneira indireta, conferindo imprecisão às carreiras. Para reconstruir carreiras de produtividade de maneira indireta, a produção científica em bases de dados gerais (que não apresentam uma identificação única para cada pesquisador) é sujeita a algoritmos de desambiguação de nome. Como resultado, esse procedimento não é capaz de reconstruir as trajetórias individuais de maneira exata [3]. Por outro lado, as carreiras de produtividade obtidas pela Plataforma Lattes são precisas, pois advêm dos currículos acadêmicos preenchidos diretamente pelos próprios pesquisadores. Além disso, nosso conjunto de dados garante uma cobertura abrangente

de diferentes campos do conhecimento.

Apesar de utilizarmos informações da Plataforma Lattes assim como no capítulo anterior, o conjunto de dados desta investigação é distinto, pois utilizamos diferentes critérios para selecionar os pesquisadores. Além disso, estamos interessados apenas no indicador produtividade. Os dados da produtividade de carreiras científicas estão disponíveis em sua totalidade na Plataforma Lattes e, neste estudo, não estão atrelados à disponibilidade de dados de outros indicadores. Em vista disso, vamos detalhar os procedimentos específicos empregados na construção deste conjunto de dados. Inicialmente, selecionamos os currículos acadêmicos dos mesmos 14 487 pesquisadores com bolsa produtividade do CNPq em maio de 2017. Para cada pesquisador, compilamos os registros de publicações anuais a partir do ano de doutoramento, definindo sua carreira científica. Incluímos informações faltantes das publicações utilizando o código DOI de referência com a *API CrossRef* e excluímos currículos de pesquisadores que não continham informações de ano de doutoramento ou de disciplina acadêmica. Além disso, consideramos apenas pesquisadores com pelo menos dez anos de carreira, o mesmo limiar utilizado num trabalho similar de Way *et al.* [65]. As carreiras de nosso conjunto de dados tem uma mediana de comprimento de 17 anos como mostra o histograma da Figura 3.1.

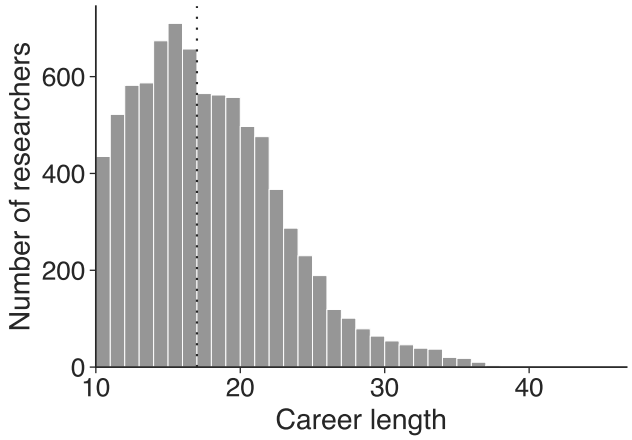

Figura 3.1: Distribuição dos comprimentos de carreira dos pesquisadores em nosso estudo. As barras representam um histograma de comprimentos de carreira, com a linha vertical pontilhada indicando o valor mediano da variável. A carreira dos pesquisadores começa após o ano de doutoramento. Todos os pesquisadores têm carreira de pelo menos dez anos.

3.2 Séries de produtividade deflacionadas, padronizadas e suavizadas

Para agrupar as séries de acordo com sua forma, precisamos lidar com três características de séries de produtividade que podem dificultar a identificação de padrões pelo nosso

método de agrupamento, que, como veremos, consiste em calcular as dissimilaridades entre cada par de séries. Precisamos lidar com a inflação temporal da produtividade, escalas de produtividade diferentes entre pesquisadores e disciplinas, e a natureza ruidosa dessas séries. Conforme já mencionamos no capítulo anterior, o volume de produção científica tem crescido consistentemente com o passar do tempo em níveis individual e agregado [59, 71, 72]. Esse crescimento da produtividade, ou inflação da produtividade, não acontece de maneira uniforme para todas as disciplinas acadêmicas, sendo influenciado pelas diversas práticas de publicação de cada área [59, 121, 125]. Para nosso conjunto de dados, os pesquisadores apresentam uma média de crescimento de 0.8 artigos/ano por década, como mostra a Figura 3.2, mas o crescimento é desigual entre as disciplinas acadêmicas. Por exemplo, enquanto a produtividade cresceu aproximadamente 2.1 artigos/ano por década para pesquisadores da medicina, ela cresceu apenas 0.7 artigos/ano por década para pesquisadores da física. A Figura 3.3 mostra as taxas de crescimento para todas as disciplinas presentes em nosso estudo. Para levar em consideração a inflação específica de cada disciplina, utilizamos o método desenvolvido por Petersen *et al.* [126] no contexto de citações para calcular uma medida deflacionada de produtividade, definida como

$$p_j(y) = \overline{p}_j(y)\frac{\mu_p(2015)}{\mu_p(y)},$$

em que $\overline{p}_j(y)$ é a produtividade bruta do pesquisador j no ano y e $\mu_p(y)$ é o valor médio da produtividade de sua disciplina no ano y. Assim, a produtividade deflacionada $p_j(y)$ representa o número de artigos anuais em um dado ano y como se tivessem sido produzidos nos níveis de produtividade de 2015. De maneira similar ao procedimento adotado no

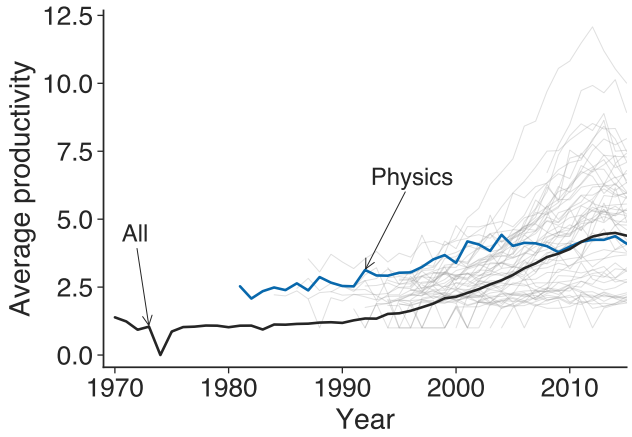

Figura 3.2: Evolução temporal da produtividade. As curvas em cinza mostram o comportamento médio das disciplinas separadamente, a curva em preto representa o comportamento médio agregado de todas as disciplinas e a curva em azul ilustra o comportamento médio da disciplina de física. Os valores médios foram estimados utilizando o estimador de localização Huber.

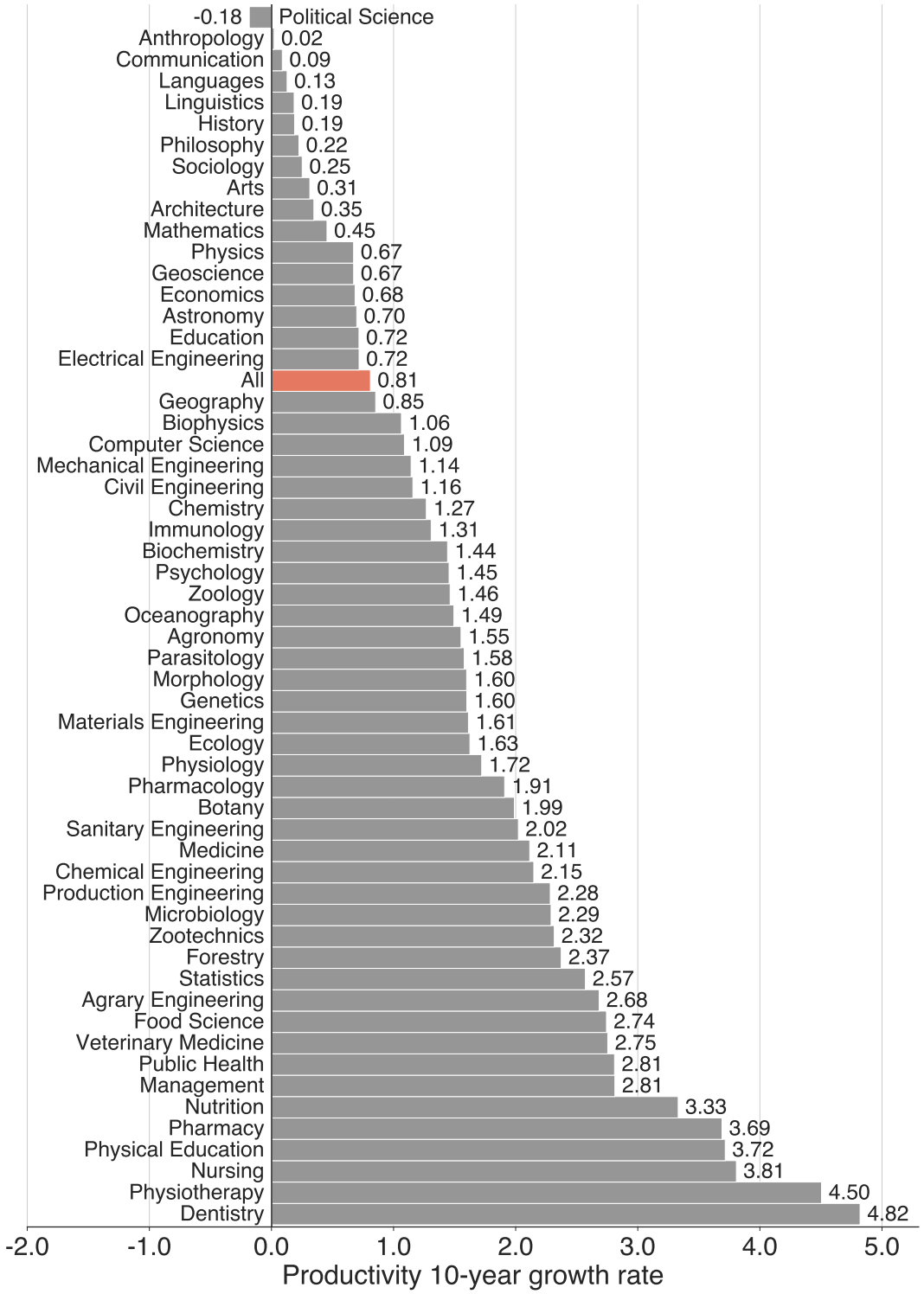

Figura 3.3: Taxas de crescimento para diferentes disciplinas. O gráfico de barras mostra as taxas de crescimento por década da produtividade. Estimamos as taxas de crescimento ajustando um modelo linear à evolução temporal reportada na Figura 3.2 para cada disciplina. Além disso, estimamos a taxa de crescimento agregando os dados de todas as disciplinas (indicado por *All* nos gráficos de barra).

capítulo anterior, utilizamos os estimadores robustos de Huber (conforme implementado no pacote de Python *statsmodels* [81], veja a Seção 1.5) para estimar a produtividade média de cada disciplina e considerar a presença dos *outliers*, ilustrados no diagrama de caixa da Figura 3.4. Estimamos a produtividade média de disciplinas apenas para anos com pelo menos cinquenta pesquisadores, descartando todos os pesquisadores que possuem pontos da carreira que não puderam ser deflacionados. Esse procedimento resulta no nosso conjunto de dados final contendo as trajetórias de produtividade deflacionadas de 8 493 pesquisadores divididos em 56 disciplinas acadêmicas, como detalha a Figura 3.5.

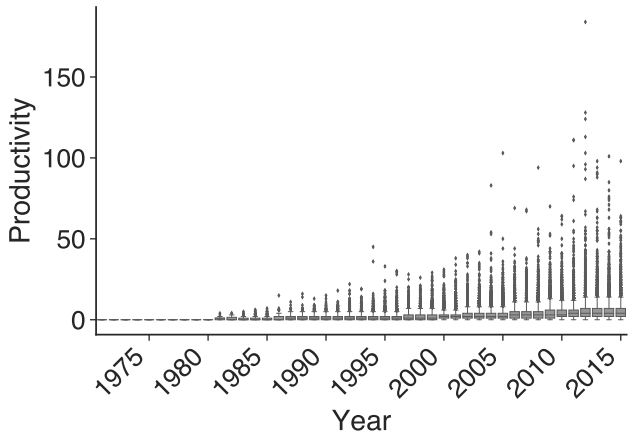

Figura 3.4: Valores *outliers* da produtividade. Os diagramas de caixa retratam o grau de dispersão da produtividade em cada ano. Existem observações extremas em todos os anos, que estão representados por marcadores pretos além dos bigodes (aqui definidos como 1.5 vezes o intervalo interquartil).

O segundo problema consiste na existência de trajetórias de produtividade com diferentes amplitudes. Nosso conjunto de dados apresenta séries com amplitudes que variam de 1 artigo por ano até 181 artigos por ano. Para tornar as trajetórias deflacionadas comparáveis em escala, padronizamos seus valores calculando o z-score de produtividade $P_j(y)$ para cada pesquisador j no ano y como

$$P_j(y) = \frac{p_j(y) - \mathbb{E}[p_j]}{\mathbb{S}[p_j]},$$

em que $\mathbb{E}[p_j]$ é a média e $\mathbb{S}[p_j]$ é o desvio padrão da produtividade da trajetória deflacionada do pesquisador j. O z-score quantifica quantos desvios padrões o pesquisador performa acima ou abaixo de sua própria produtividade média e faz com que todas as trajetórias sejam comparáveis em escala.

Por fim, a natureza ruidosa das trajetórias também se coloca como um desafio para estimar as dissimilaridades entre séries. Essas flutuações refletem a natureza intrínseca da produção científica, em que cada trabalho passa por um processo não determinístico e demorado de elaboração, experimentação, escrita e revisão por pares [127]. O ponto exato no tempo em que o artigo é publicado muitas vezes não reflete a data de finalização do

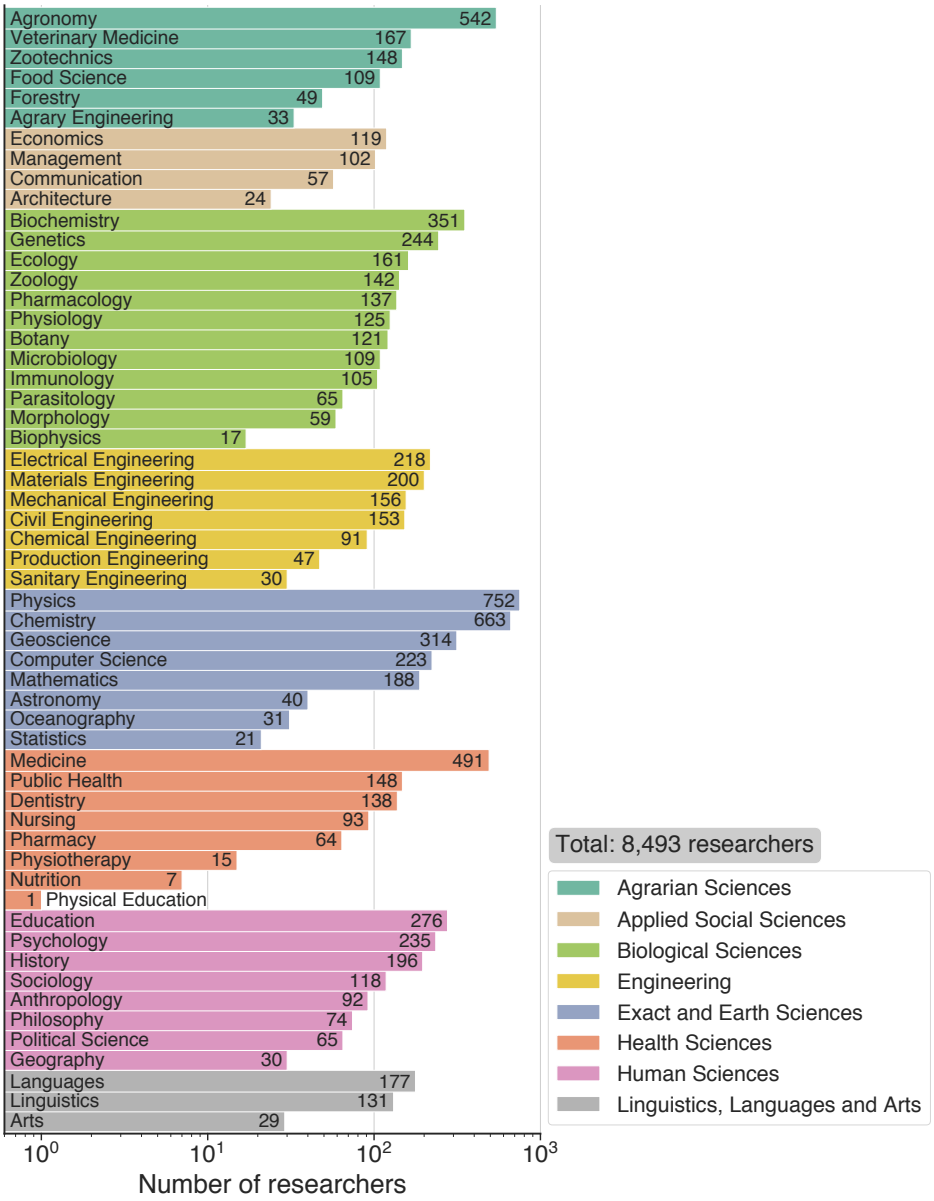

Figura 3.5: Número de pesquisadores no nosso conjunto de dados. O gráfico de barras mostra o número total de pesquisadores para cada disciplina. As cores das barras representam os diferentes campos da ciência em nosso conjunto de dados.

trabalho. Para lidar com esse problema, aplicamos um filtro gaussiano em todas as séries de produtividade padronizadas (conforme implementado no pacote `SciPy` [128]). Esse filtro designa pesos gaussianos com desvio padrão σ para cada ponto da série de produtividade e, posteriormente, utiliza esses pesos para obter trajetórias suavizadas por um processo de convolução. O parâmetro σ controla o grau de suavização, definindo a escala temporal em que o filtro é aplicado. A Figura 3.6 mostra o efeito da variação do parâmetro σ. Usamos $\sigma = 2$ anos para os resultados mostrado no texto principal, mas padrões similares de agrupamento são obtidos ao variar σ de 1.0 a 2.5 anos em intervalos de 6 meses (Figuras A.15, A.16 e A.17). Ao aplicar o filtro gaussiano nas trajetórias padronizadas, garantimos que o processo de

suavização foi aplicado uniformemente entre os pesquisadores com diferentes variabilidades de produtividade.

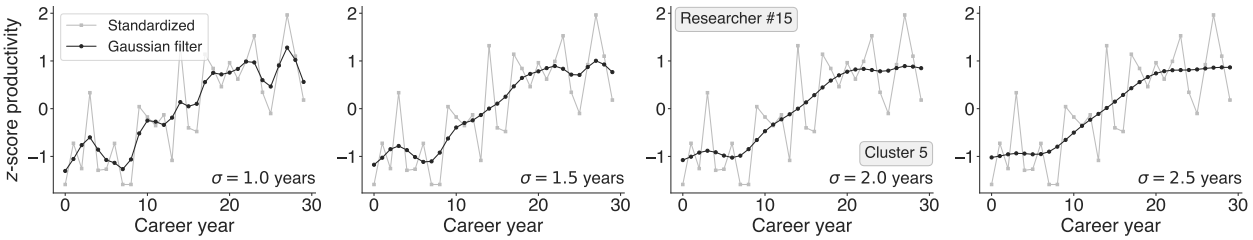

Figura 3.6: Ilustração dos diferentes graus de suavização das trajetórias de produtividade ao variar o desvio padrão do *kernel* gaussiano utilizado para filtrar as trajetórias padronizadas. O painel mostra as séries de z-score de produtividade (marcadores em cinza) e as séries suavizadas correspondentes (marcadores em preto) obtidas ao variar o desvio padrão σ de 1.0 a 2.5 anos em intervalos de 6 meses.

3.3 Agrupamento das séries temporais

Em seguida, estimamos as dissimilaridades entre todos os pares de trajetórias pré-processadas usando o algoritmo *dynamic time warping* [129] (DTW, conforme implementado no pacote `dtaidistance` [130]). Considerando duas sequências $S = s_1, \ldots, s_n$ e $T = t_1, \ldots, t_m$, podemos criar uma grade $n \times m$, em que cada ponto (i, j) corresponde a determinado alinhamento entre dois elementos s_i e t_j. Um caminho não linear $W = w_1, \ldots, w_k$, em que cada w_k é um elemento $(i, j)_k$ da grade, pode ser calculado para alinhar todos os elementos de forma a minimizar a "distância" entre as duas sequências. Para isso, definimos uma medida de distância como $\delta(i, j) = |s_i - t_j|$. A medida de dissimilaridade DTW é obtida pela minimização da distância cumulativa de todos os caminhos não lineares W, isto é,

$$\mathrm{DTW}(S, T) = \min_{W} \sum_{k} \delta(w_k)\,. \tag{3.1}$$

A medida garante maior flexibilidade, pois identifica padrões similares que podem estar deslocados temporalmente e pode ser calculada para séries de tamanhos diferentes. Após calcular a dissimilaridade entre todos os pares de trajetórias de produtividade, a matriz de dissimilaridade resultante é utilizada como uma métrica precomputada na técnica de redução de dimensionalidade *uniform manifold approximation and projection* (UMAP, conforme implementada no pacote `umap` [131], veja a Seção 1.6). De maneira simplificada, podemos afirmar que o primeiro passo do algoritmo cria uma representação de rede a partir da matriz de dissimilaridade, em que os vértices representam pesquisadores e as arestas pesadas conectam pesquisadores com trajetórias de produtividade parecidas. O segundo passo, por sua vez, projeta os dados em um espaço de baixa dimensionalidade por um algoritmo de *layout*

de grafo direcionado por força.

Em nosso procedimento, utilizamos apenas a estrutura topológica de rede e descartamos a representação em baixa dimensão produzida pelo UMAP[1]. Assim, podemos mapear o problema de agrupamento das séries em um problema de detecção de comunidades na rede UMAP. Para encontrar a estrutura modular, utilizamos a equação mapa [113, 114] e a equação mapa hierárquica [117], que definem o método de agrupamento Infomap. Conforme detalhado na Seção 1.7, o Infomap é uma técnica de detecção de comunidades em redes baseada em conceitos de teoria da informação e que utiliza caminhadas aleatórias como um *proxy* para o fluxo de informação na rede. A equação mapa e a equação mapa hierárquica representam os limites teóricos de quão concisamente se pode descrever uma caminhada aleatória infinita na rede (o comprimento de descrição) dada uma configuração particular de partição não hierárquica e hierárquica, respectivamente. Ao minimizar as duas equações, o Infomap revela a estrutura de comunidades da rede. Esse método apresenta performances excelentes em tarefas utilizando redes de *benchmark* com partições plantadas [132–134]. Utilizamos a implementação do Infomap do pacote `infomap` [135] com parâmetros padrões, testando o modelo de dois níveis e o modelo hierárquico. A equação mapa hierárquica resulta num comprimento de descrição menor, sendo escolhida como nosso método de agrupamento.

Enquanto os *embeddings* em baixa dimensão produzidos pelo UMAP não são determinísticos, a rede criada no primeiro passo é sempre igual para um conjunto de dados fixo. Entretanto, como o algoritmo do Infomap baseia-se em caminhadas aleatórias pela rede, cada execução do método resulta em partições similares mas que não são idênticas, impondo um caráter não determinístico ao processo de agrupamento. Para levar isso em consideração, executamos mil realizações do algoritmo Infomap e observamos que todas as partições são qualitativamente comparáveis. O número de comunidades detectadas varia de 7 a 14, mas cerca de 85% de todas as realizações apresenta de 9 a 11 comunidades, com 10 sendo o número mais comum de partições (34%). A Figura 3.7 mostra o histograma com a quantidade

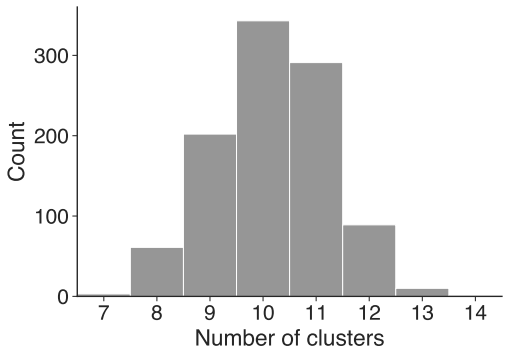

Figura 3.7: Histograma do número de comunidades detectadas na rede em mil realizações do algoritmo Infomap.

[1]Uma abordagem similar foi recentemente empregada por Lee *et al.* [111] no contexto de Neurociência.

de ocorrências para cada número de comunidades. Finalmente, escolhemos a partição final como aquela que maximiza o coeficiente de silhueta dentre as partições com número modal de dez grupos. A Figura 3.8 ilustra o procedimento completo utilizado para agrupar as séries de produtividade.

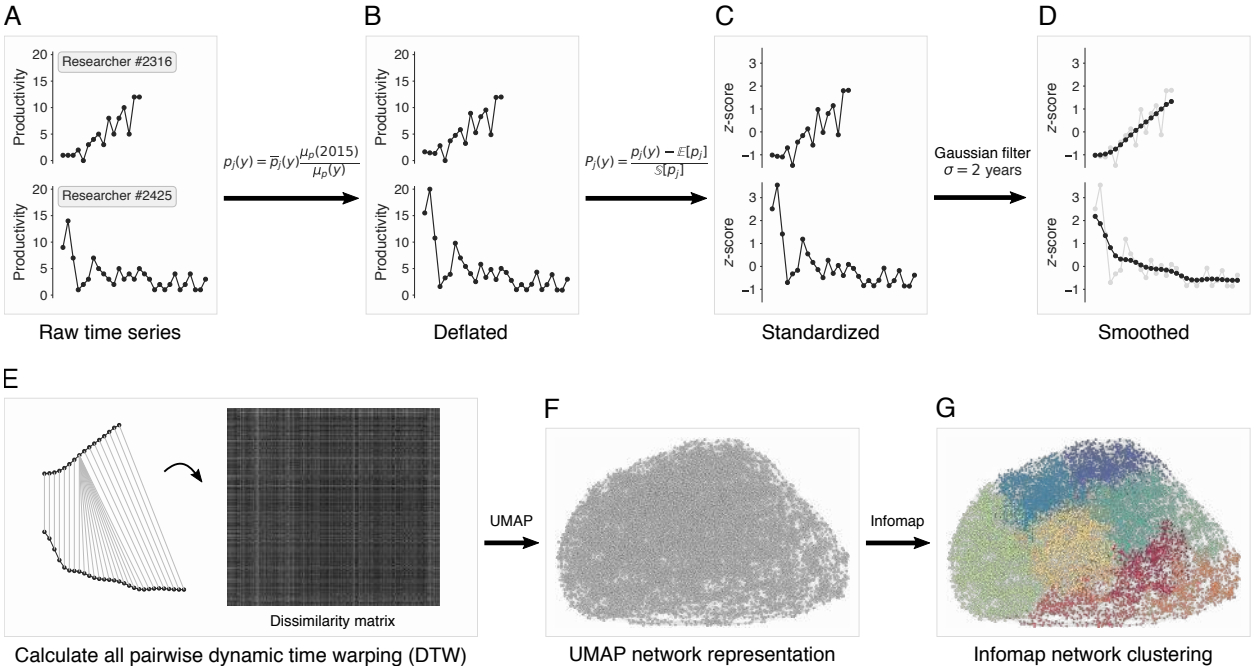

Figura 3.8: Descrição do método de agrupamento. (A) Trajetórias de produtividade brutas. (B) Trajetórias de produtividade deflacionadas. (C) Trajetórias de produtividade padronizadas. (D) Trajetórias de produtividade suavizadas. (E) Cálculo das dissimilaridades DTW entre todos os pares de trajetórias. (F) Representação em rede da matriz de dissimilaridade obtida no passo anterior por meio do algoritmo de redução de dimensionalidade UMAP. (G) Comunidades de padrões de produtividade obtidas por meio do algoritmo de agrupamento Infomap.

3.4 Padrões universais de produtividade

O painel central da Figura 3.9 mostra a representação de rede produzida pelo UMAP, com diferentes cores indicando as dez comunidades detectadas como melhor partição do Infomap, numeradas de 1 a 10. Ao redor da visualização da rede, incluímos as trajetórias de produtividade de todos os pesquisadores em cada grupo, bem como o comportamento médio de cada grupo. Além disso, desenvolvemos uma visualização interativa que pode ser acessada em: complex.pfi.uem.br/cluster. Reescalamos o comprimento das carreiras para o intervalo unitário para melhor visualizar trajetórias de diferentes tamanhos. As curvas de produtividade em cada grupo apresentam formas muito similares. Para averiguar de maneira quantitativa se os grupos contém séries similares, investigamos a estrutura da matriz de dissimilaridade que gera a representação de rede. A Figura 3.10A mostra a matriz

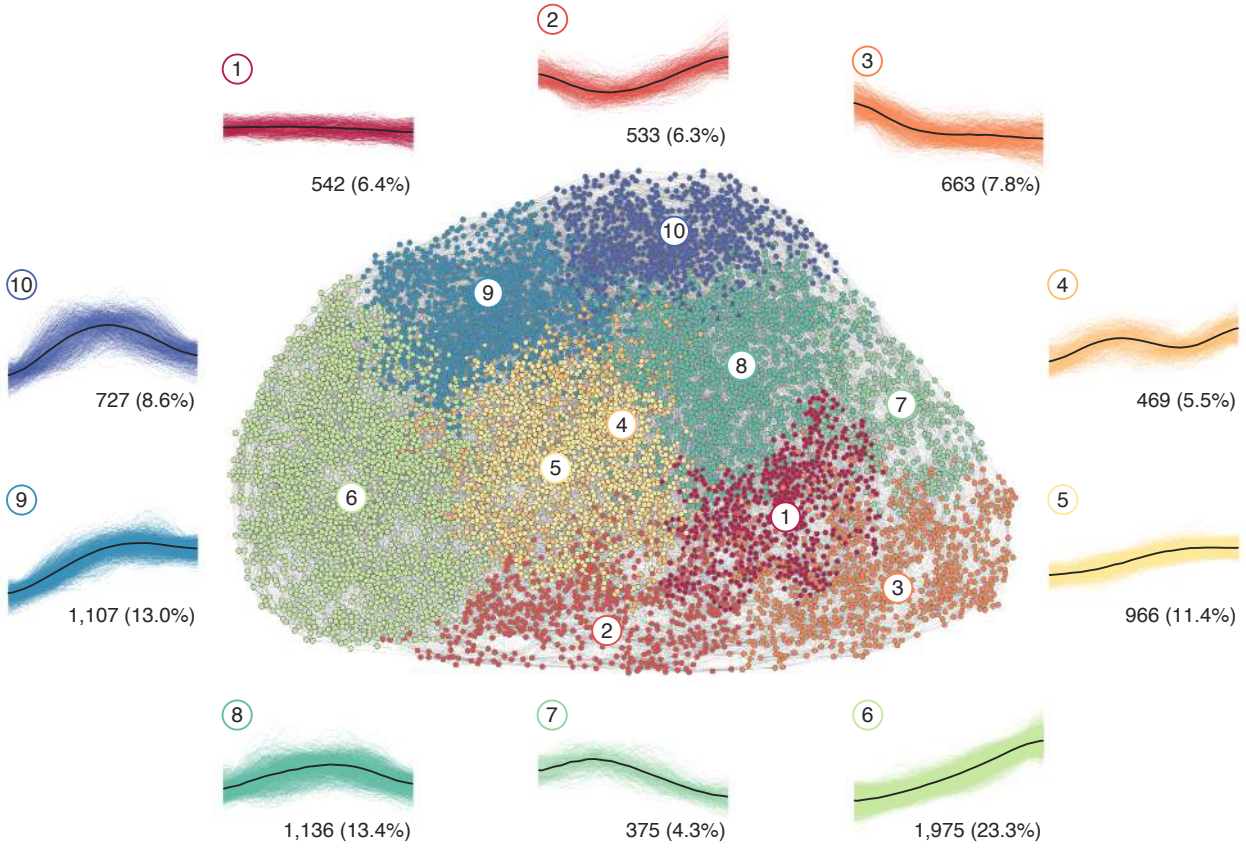

Figura 3.9: Padrões de trajetórias de produtividade. O painel central mostra a representação em rede, em que vértices representam pesquisadores e arestas pesadas conectam pesquisadores com curvas de produtividade similares. Dez comunidades distintas, representadas por diferentes cores e denotadas por números de 1 a 10, são identificadas e correspondem a grupos de pesquisadores com padrões de produtividade similares. Os painéis ao redor da rede mostram as curvas de produtividade de cada comunidade. As curvas em preto representam o comportamento médio da produtividade de cada grupo. Os comprimentos das carreiras de cada grupo foram reescaladas para o intervalo unitário e as frações de pesquisadores em cada grupo são mostradas em cada painel. Os dez grupos são agrupados em seis categorias de padrões: constante (grupo 1), em forma de U (grupo 2), decrescente (grupo 3), periódico (grupo 4), crescente (grupos 5 e 6) e com aspecto canônico (grupos 7 a 10).

de dissimilaridade sem ordenar os pesquisadores com respeito às comunidades, enquanto a Figura 3.10B mostra a matriz de dissimilaridade ordenada. As comunidades da rede UMAP formam um bloco diagonal na matriz ordenada, indicando que séries do mesmo grupo são mais similares entre si. Mais ainda, vértices e grupos que estão próximos na rede compartilham padrões de produtividade similares. Por exemplo, os grupos de 7 a 10 todos apresentam um comportamento médio marcado por um pico em produtividade e aparecem em posições adjacentes na rede. Em contraste, os grupos 3 e 6 representam comportamentos opostos (tendências crescentes *versus* decrescentes) e, dessa forma, estão em regiões opostas na rede. Ao inspecionar visualmente os padrões de produtividade na representação de rede

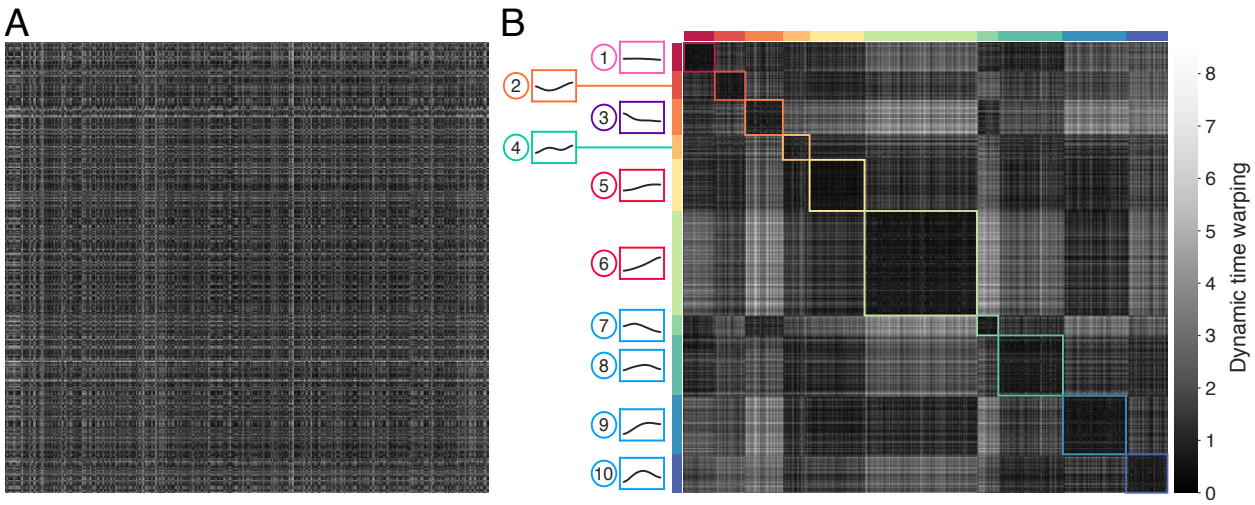

Figura 3.10: Cálculo da medida de dissimilaridade *Dynamic time warping* (DTW) para todos os pares de trajetórias de produtividade. (A) Matriz das medidas de dissimilaridade DTW sem ordenar os pesquisadores de acordo com a estrutura de comunidades. (B) Matriz das medidas de dissimilaridade DTW com a ordenação dos pesquisadores de acordo com a estrutura de comunidades. As cores em volta da matriz correspondem a cada uma das dez comunidades identificadas pelo Infomap. Os quadrados coloridos dentro da representação matricial indicam cada um dos grupos de trajetórias.

usando a visualização interativa, também observamos que vértices localizados na fronteira entre duas ou mais comunidades apresentam frequentemente padrões de produtividade mais complexos que podem remeter a uma mistura do comportamento médio de grupos adjacentes.

Nossa análise revela um diverso conjunto de trajetórias de produtividade que vão além da narrativa canônica, incluindo padrões que foram apenas hipotetizados ou observados em estudos usando dados agregados [53,54,58,61,62,64]. Uma examinação detalhada das trajetórias e suas derivadas, representadas na Figura 3.11, permite agrupar os dez grupos em seis categorias: constante (grupo 1), em forma de U (grupo 2), decrescente (grupo 3), periódica (grupo 4), crescente (grupos 5 e 6) e com aspecto canônico (grupos 7 a 10). As trajetórias constantes são caracterizadas por produtividade estável ou ligeiramente decrescente, representando 6.4% dos pesquisadores. As trajetórias em forma de U apresentam um decréscimo seguido por um crescimento na produtividade, representando 6.3% dos pesquisadores. As trajetórias decrescentes exibem um declínio brusco na produtividade seguido por um platô quase constante, representando 7.8% dos pesquisadores. As trajetórias periódicas têm um pico antes do meio da carreira seguido por um crescimento na produtividade, representando 5.5% dos pesquisadores. Juntos, esses padrões representam pouco mais de um quarto dos pesquisadores, sendo as trajetórias periódicas as menos frequentes. Como resultado, padrões crescentes e com aspecto canônico descrevem quase três quartos dos pesquisadores em nosso estudo. Especificamente, 35% dos pesquisadores têm carreiras crescentes, que são divididas em dois grupos: o primeiro em que a produtividade sempre cresce ao longo da carreira

(grupo 6) e o segundo em que a produtividade cresce com taxas decrescentes ou aproximando a um platô (grupo 5). As carreiras com aspecto canônico, definidas de maneira ampla como aquelas que apresentam um pico único em produtividade (grupos 7 a 10), são o tipo mais frequente de trajetória, representando 39% dos pesquisadores. A denominação "canônica" vêm de uma série de trabalhos na literatura de ciência da ciência, com o trabalho de Lehman [49] sendo a investigação seminal, que encontraram padrões de produtividade com um pico único por meio de análises utilizando dados agregados [49–60]. Entretanto, utilizamos o termo "aspecto canônico" visto que a definição de Lehman é mais restritiva, definindo a narrativa canônica como "curvas de criatividade que inicialmente crescem rapidamente e, após atingir um pico em seu início, diminuem lentamente" [49] (em tradução livre). Observamos que apenas o grupo 7 estritamente corresponde à definição de Lehman, pois é o único grupo que apresenta um pico antes do meio da carreira, conforme mostra a Figura 3.12. De fato, as posições do pico são um dos comportamentos que distinguem os grupos de aspecto canônico e, possivelmente, a causa por emergirem como grupos distintos. A Tabela 3.1 mostra as posições dos picos das carreiras de aspecto canônico.

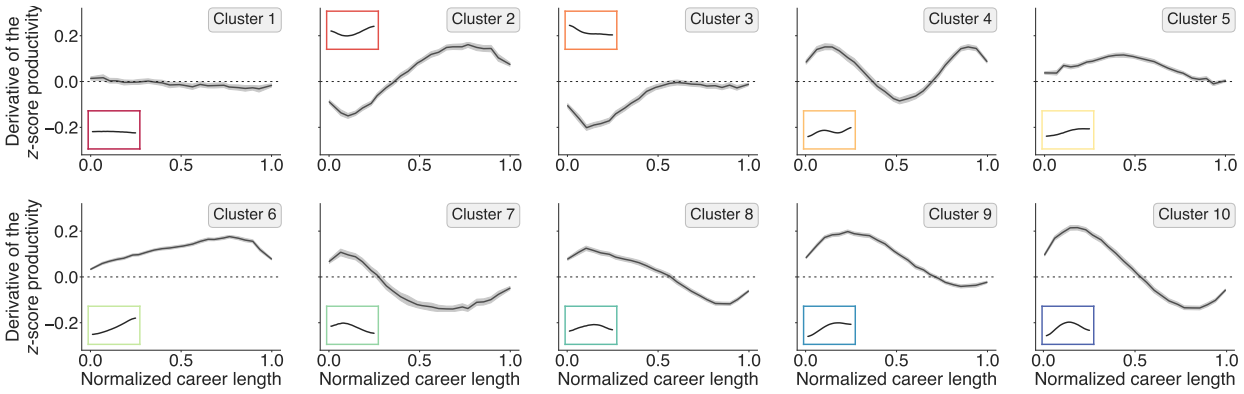

Figura 3.11: Derivadas das trajetórias de produtividade de cada grupo. As curvas em cada painel mostram a média móvel das trajetórias de produtividade diferenciadas, com as regiões sombreadas representando os intervalos de confiança de 95%. Os comprimentos das carreiras dos pesquisadores em cada grupo foram reescalados para o intervalo unitário antes da estimativa das médias. Valores positivos indicam taxas crescentes de produtividade, valores negativos indicam taxas decrescentes de produtividade e valores próximos de zero indicam produtividade constante em anos consecutivos da carreira. Com base nessas curvas e nos padrões médios de produtividade em cada grupo, definimos seis categorias de narrativas de produtividade: constante (grupo 1), em forma de U (grupo 2), decrescente (grupo 3), periódico (grupo 4), crescente (grupos 5 e 6) e com aspecto canônico (grupos 7 a 10).

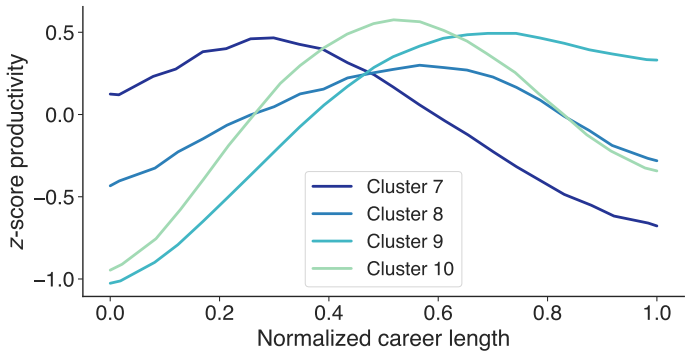

Figura 3.12: Diferenças na produtividade média dos grupos com aspecto canônico. As curvas coloridas representam as trajetórias médias de produtividade para cada um dos quatro grupos que compõem a categoria de trajetórias com aspecto canônico (grupos 7 a 10, como indicado na legenda). Os comprimentos das carreiras em cada grupo foram reescalados para o intervalo unitário.

Tabela 3.1: Estatísticas descritivas dos pesquisadores em cada grupo. A coluna da amplitude indica o valor médio da diferença entre os valores máximo e mínimo da produtividade em unidades padronizadas. A coluna da posição do pico indica o valor médio do ano de ocorrência da máxima produtividade para cada grupo da categoria de trajetórias com aspecto canônico. A coluna da posição normalizada do pico indica o valor médio do ano de ocorrência da máxima produtividade após reescalar os comprimentos de carreira para o intervalo unitário. Nessas três colunas, os valores após o símbolo ± representam um desvio padrão da quantidade correspondente.

Grupo	Número de pesquisadores	Categoria	Comprimento mediano (anos)	Amplitude (unidade padronizada)	Posição do pico (anos)	Posição normalizada do pico
1	542 (6.4%)	Constante	14	0.58 ± 0.19	-	-
2	533 (6.3%)	Em forma de U	16	1.27 ± 0.38	-	-
3	663 (7.8%)	Decrescente	16	1.37 ± 0.41	-	-
4	469 (5.5%)	Periódica	19	1.26 ± 0.29	-	-
5	966 (11.4%)	Crescente	14	1.04 ± 0.26	-	-
6	1,975 (23.3%)	Crescente	16	1.91 ± 0.32	-	-
7	375 (4.4%)	Canônica	17	1.37 ± 0.29	5.78 ± 2.97	0.33 ± 0.14
8	1,136 (13.4%)	Canônica	18	1.16 ± 0.30	10.12 ± 4.42	0.55 ± 0.15
9	1,107 (13.0%)	Canônica	18	1.72 ± 0.28	13.26 ± 4.54	0.73 ± 0.17
10	727 (8.6%)	Canônica	21	1.84 ± 0.27	11.64 ± 4.18	0.54 ± 0.13

3.5 Robustez dos seis padrões de produtividade

Para verificar a robustez de nosso procedimento, realizamos uma sequência de testes avaliando: a qualidade do agrupamento dos nossos dados em comparação com dados das trajetórias embaralhadas entre os grupos e de modelos nulos, a robustez em relação ao método de agrupamento, a consistência da classificação nas seis categorias e a qualidade semântica dos grupos. Primeiramente, comparamos o coeficiente de silhueta de nosso procedimento com os coeficientes calculados a partir do embaralhamento das trajetórias entre os grupos. O coeficiente de silhueta mede o quão similar cada série de produtividade é em relação aos

membros de seu grupo quando comparado com membros de outros grupos. O coeficiente é calculado como a diferença normalizada média entre a coesão (a distância média intra-grupo) e a separação (a distância média ao grupo mais próximo) para cada trajetória. O coeficiente varia de -1 a 1, com valores maiores indicando configurações com agrupamento de melhor qualidade. A Figura 3.13 mostra que o coeficiente de silhueta de nosso procedimento é significativamente maior do que os coeficientes obtidos ao misturar as trajetórias entre os grupos. Verificamos também que nossa melhor partição não apenas gera grupos internamente consistentes, bem como gera partições melhores do que partições associadas a dois modelos nulos para trajetórias de produtividade. O primeiro modelo nulo gera trajetórias de produtividade artificiais utilizando uma distribuição binomial com parâmetros que replicam a produtividade média de nosso dado (4.37 artigos/ano). O segundo modelo nulo gera carreiras sintéticas a partir do embaralhamento da produtividade das carreiras dos pesquisadores. Para cada um dos modelos nulos, criamos mil réplicas com mesmo número de carreiras e mesma distribuição de tamanho de carreira de nosso conjunto de dados. Para cada réplica, aplicamos os mesmos procedimentos descritos na Figura 3.8 e calculamos o coeficiente de silhueta associado à partição final. A Figura 3.13 mostra as distribuições de probabilidade

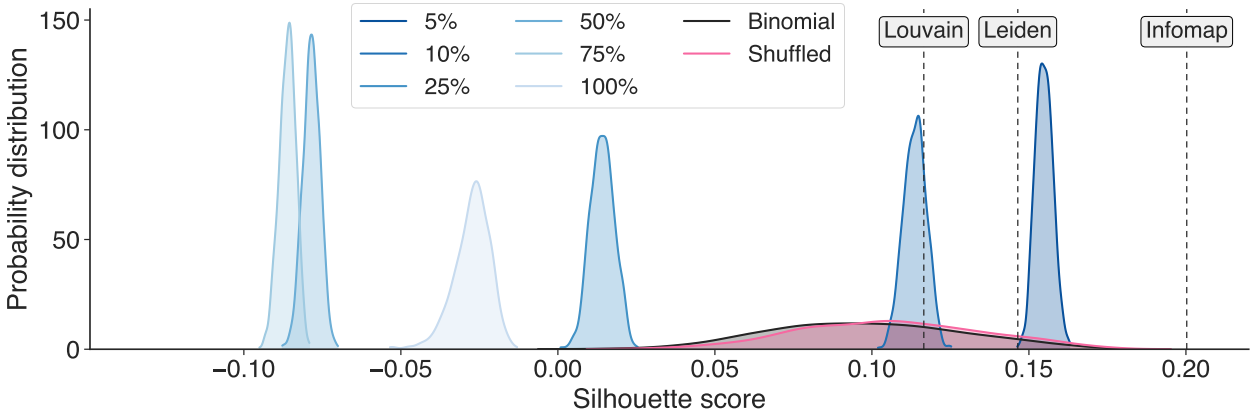

Figura 3.13: Verificação da consistência e da significância do agrupamento via coeficiente de silhueta. As linhas tracejadas verticais indicam o coeficiente de silhueta para a melhor partição do método Infomap, do método Louvain e do método Leiden de agrupamento. As curvas em tons azuis mostram as distribuições de probabilidade dos coeficientes de silhueta obtidas após embaralhar uma dada fração (indicada na legenda) das classificações da melhor partição do Infomap. As curvas em preto e rosa mostram as distribuições de probabilidade dos coeficientes de silhueta obtidos a partir da aplicação do nosso algoritmo de agrupamento, em mil realizações, a carreiras geradas a partir de dois modelos nulos. A curva em preto mostra a distribuição de probabilidade do coeficiente de silhueta para carreiras artificiais geradas com a mesma distribuição de comprimentos de carreira do nosso dado a partir de uma distribuição binomial, cujos parâmetros são escolhidos para replicar a produtividade média do dado (4.37 artigos/ano). A curva em rosa mostra a distribuição de probabilidade do coeficiente de silhueta para carreiras obtidas a partir do embaralhamento da produtividade nas carreiras do nosso conjunto de dados.

dos coeficientes de silhueta obtidos dos dois modelos nulos, em preto e rosa, cujos valores são significativamente menores do que para o dado original. Para verificar a robustez das categorias mediante a escolha de métodos de agrupamento alternativos, aplicamos também os algoritmos de detecção de comunidade determinísticos Louvain [136] e Leiden [137], que resultam em grupos de padrões similares (Figuras A.18 e A.19), todavia com coeficientes de silhueta menores, como mostra a Figura 3.13.

Para validar a robustez das seis categorias de padrões de produtividade, conduzimos dez realizações do procedimento de agrupamento com amostras obtidas pela divisão aleatória do dado completo em três partes iguais. Em cada realização, classificamos cada pesquisador dentre uma das seis categorias para verificar a consistência entre a classificação da amostragem com a classificação do dado original. Verificamos que os grupos obtidos podem ser consistentemente classificados nas mesmas seis categorias de padrões de produtividade encontradas para o dado completo. A Figura 3.14A mostra a matriz de confusão associada à classificação dos dados completos (linhas) e à classificação por amostragem (colunas) calculada a partir de trinta amostras. A acurácia média da classificação por amostragem (73%) é significativamente maior do que classificadores *dummy* com estratégias de moda (39%), de estratificação (29%) e uniforme (16%). A matriz exibe um padrão primordialmente diagonal, com diferenças ocorrendo principalmente quando trajetórias periódicas são classificadas como trajetórias crescentes ou canônicas. Calculamos ainda a entropia normalizada relacionada às probabilidades de atribuição de cada categoria para cada pesquisador considerando as dez realizações. Para calcular esses valores, estimamos as frações $[(p_1, p_2, \ldots, p_6)]$ de pertencimento a cada uma das seis categorias (constante, forma em U, decrescente, periódica, crescente e com aspecto canônico) para cada pesquisador considerando as dez realizações. A entropia é calculada pela fórmula padrão da entropia normalizada de Shannon [115], isto é, $h = -\frac{1}{\log 6} \sum_{i=1}^{6} p_i \log p_i$. A Figura 3.14B mostra o histograma da entropia normalizada. Aproximadamente 80% dos pesquisadores apresentam entropia normalizada abaixo de 0.5, indicando que as trajetórias são consistentemente classificadas como a mesma categoria nas dez realizações. Além disso, cerca de um terço dos pesquisadores têm entropia zero, isto é, eles sempre são classificados como a mesma categoria. A Figura 3.14C mostra a representação de rede com os grupos de padrões de produtividade (indicados por seus números e formas) delimitados por linhas pretas, enquanto a paleta azul corresponde aos valores de entropia normalizada. Notamos que pesquisadores com valores altos de entropia estão mais frequentemente localizados na fronteira entre dois ou mais grupos (onde os padrões tendem a ser mais complexos) ou na região de sobreposição entre os padrões periódico (grupo 4) e crescente com taxas decrescentes (grupo 5). Essas mesmas observações permanecem verdadeiras mudando a estratégia de amostragem dividindo o dado na metade (Figura A.20).

Por fim, para avaliar o desempenho semântico de nosso método, conduzimos uma validação humana em que um painel de dois especialistas categorizou 25% das trajetórias de nosso

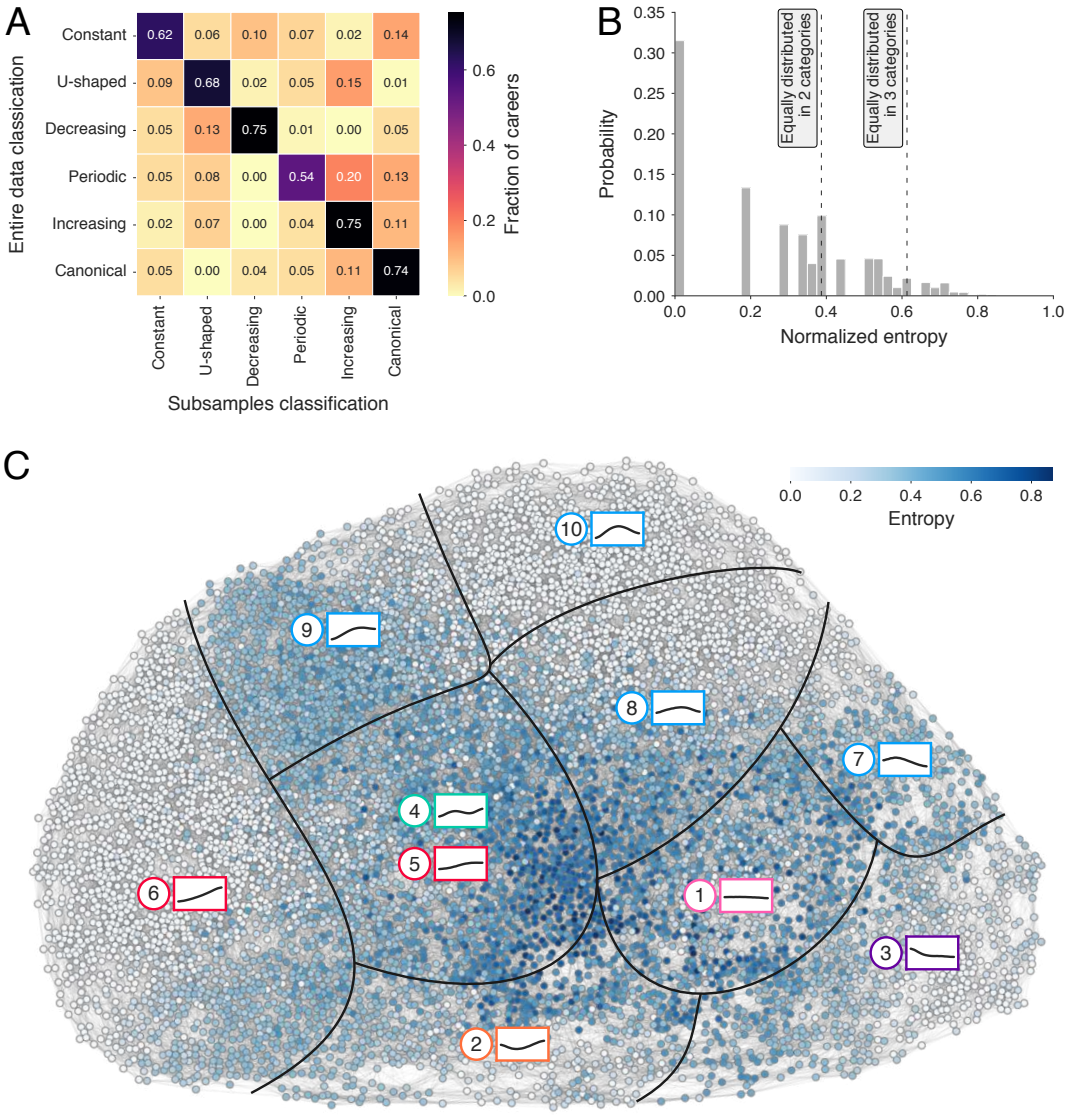

Figura 3.14: Validação da robustez das seis categorias de padrões de produtividade após amostrar os dados aleatoriamente em três partes iguais. (A) Matriz de confusão média associada à classificação dos dados completos (linhas) e à classificação por amostragem (colunas) calculada usando trinta amostras. (B) Histograma da entropia normalizada associada com as probabilidades de pertencimento às seis categorias de trajetória de produtividade. (C) Representação em rede em que vértices representam pesquisadores e arestas pesadas conectam pesquisadores com curvas de produtividade similares. As linhas em preto delimitam aproximadamente os dez grupos de curvas de produtividade (indicados no painel por seus números e padrões), enquanto os tons de azul correspondem aos valores normalizados de entropia.

conjunto de dados aleatoriamente amostradas de maneira estratificada. Eles são expostos a uma aplicação interativa, como ilustra a Figura 3.15, em que as trajetórias padronizadas e suavizadas são mostradas individualmente juntamente com os padrões médios de cada grupo. Seis botões são fornecidos para classificação de cada categoria e um botão extra é disponibilizado para situação em que há desacordo na classificação. A Figura 3.16A mostra

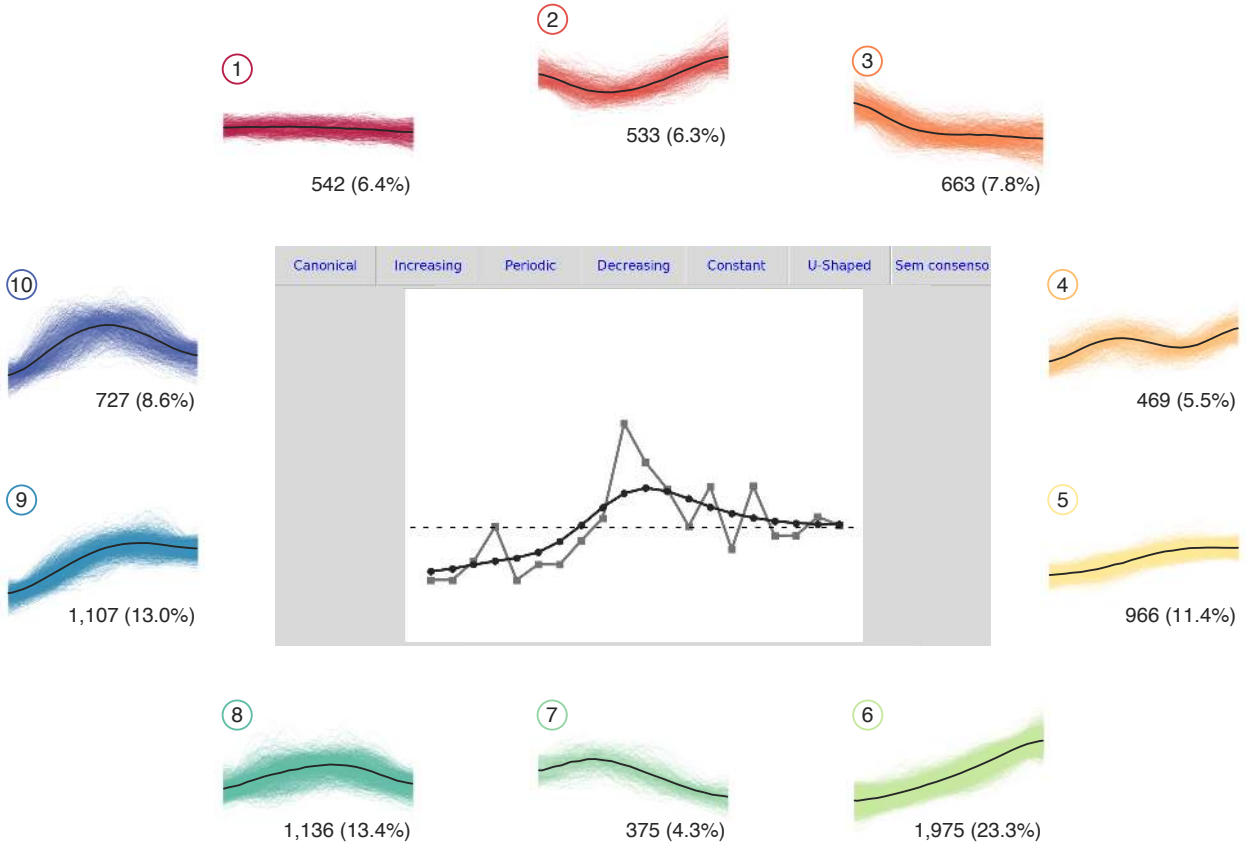

Figura 3.15: Captura de tela da aplicação interativa utilizada para realizar a validação humana. Para cada classificação, uma trajetória padronizada (em cinza) e suavizada (em preto) é mostrada no centro da aplicação. A linha horizontal tracejada mostra o valor médio da produtividade do pesquisador. Seis botões são disponibilizados para classificar o pesquisador dentre as seis categorias de trajetória e um botão adicional pode ser pressionado quando não há consenso entre os especialistas. As trajetórias de cada grupo e seus comportamentos médios também são disponibilizados como referência para classificação.

a matriz de confusão associada com a classificação algorítmica (linhas) e a classificação humana (colunas). A acurácia da classificação humana (73%) é significativamente maior do que classificadores *dummy* com estratégias de moda (39%), de estratificação (29%) e uniforme (16%). A matriz exibe um padrão primordialmente diagonal, com diferenças principalmente ocorrendo quando os especialistas classificam trajetórias em forma de U como decrescentes e periódicas como trajetórias com aspecto canônico. As trajetórias em forma de U e periódicas são as categorias que, proporcionalmente, causam maior discordância entre os dois especialistas (10% e 14%, respectivamente). A Figura 3.16B mostra a matriz de confusão associada à classificação algorítmica (linhas) e à classificação humana (colunas) para os grupos nas categorias crescente e com aspecto canônico. As curvas de produtividade de padrão crescente com taxas decrescentes (grupo 5) e de pico tardio (grupo 9) são mais confundidas entre si. A Figura 3.16C mostra a representação de rede com os grupos de padrões de produtividade

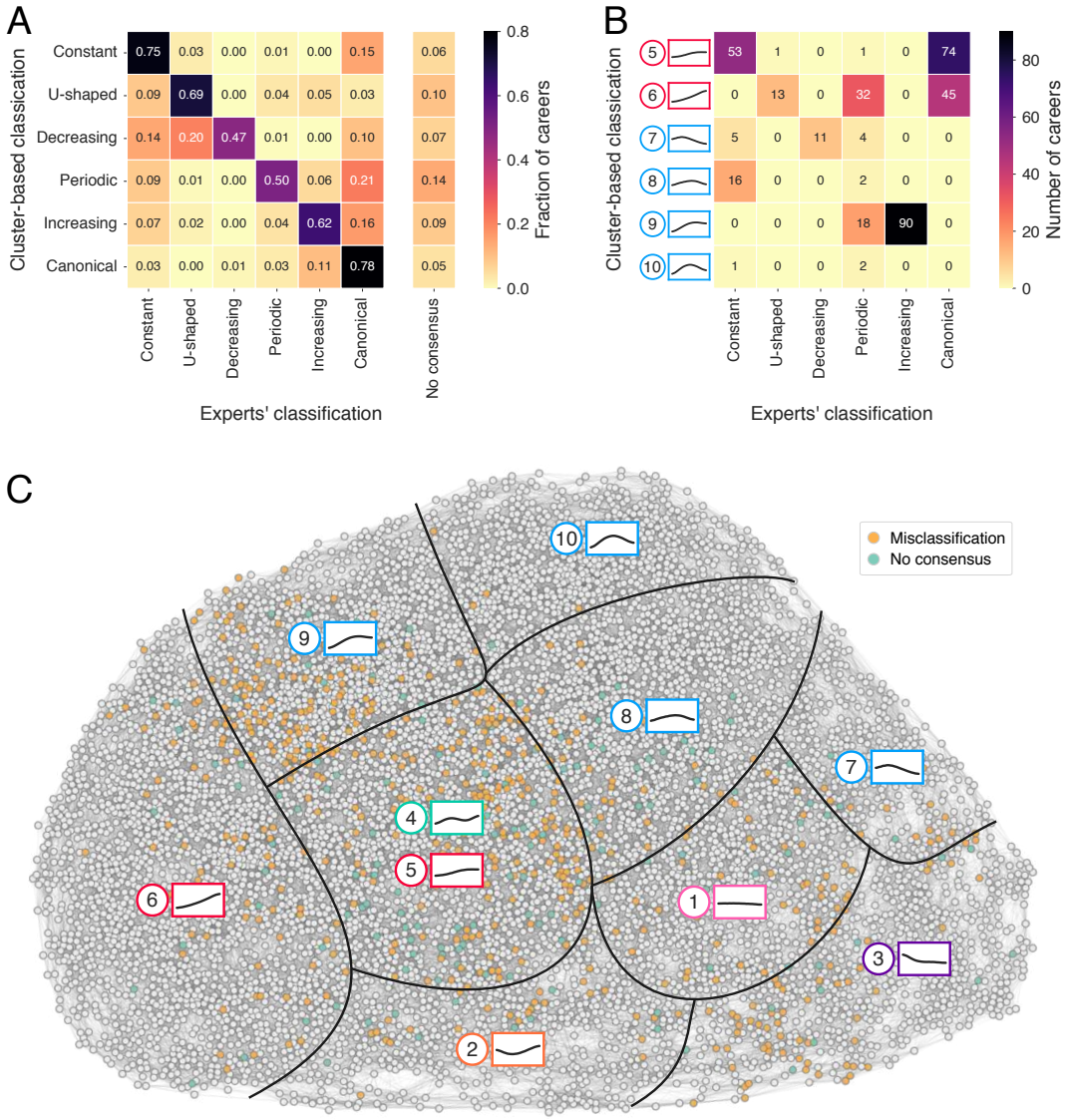

Figura 3.16: Validação humana das seis categorias de padrões de produtividade. (A) Matriz de confusão associada à classificação dos dados completos (linhas) e à classificação dos especialistas (colunas). (B) Matriz de confusão associada à classificação dos dados completos (linhas) e à classificação dos especialistas (colunas) para grupos nas categorias crescente e com aspecto canônico. (C) Representação em rede em que vértices representam pesquisadores e arestas pesadas conectam pesquisadores com curvas de produtividade similares. As linhas em preto delimitam aproximadamente os dez grupos de curvas de produtividade (indicados no painel por seus números e padrões). Os marcadores em laranja representam trajetórias com classificação divergente (a classificação dos especialistas não concorda com a classificação obtida por nosso algoritmo), enquanto os marcadores em verde indicam trajetórias para as quais não houve consenso entre os especialistas.

(indicados por seus números e formas) delimitados por linhas pretas. Os marcadores em laranja representam trajetórias com classificação divergente (a classificação dos especialistas não concorda com a classificação obtida por nosso algoritmo), enquanto os marcadores em

verde indicam trajetórias em que não houve consenso entre os especialistas. Similarmente à validação por amostragem, ambos os tipos de inconsistência são mais frequentemente observados na região de fronteira entre dois ou mais grupos e na região de sobreposição entre os padrões periódico (grupo 4) e crescente com taxas decrescentes (grupo 5).

3.6 Efeitos geracionais e de disciplina

A prevalência de cada padrão de produtividade pode variar entre carreiras acadêmicas de diferentes comprimentos. Para examinar esse potencial efeito de tamanho, estimamos as distribuições de probabilidade de tamanho de carreira dos pesquisadores de cada grupo. A Figura 3.17A mostra que todos os grupos abrangem um amplo intervalo de comprimentos de carreira, mas com valores distintos de mediana como detalha a Tabela 3.1. As trajetórias constantes e crescentes exibem os menores valores da mediana do comprimento de carreira (≈15 anos), enquanto carreiras canônicas e periódicas representam os maiores valores (≈20 anos). Para identificar os padrões de carreira mais comuns em cada estágio da carreira, agrupamos as carreiras acadêmicas em quatro intervalos de comprimento (10-14, 15-19, 20-24 e maiores do que 24 anos) e calculamos a prevalência de cada padrão. A Figura 3.17B mostra que a categoria crescente é a mais representativa para carreiras curtas, representando 45% dos pesquisadores. Entretanto, as curvas crescentes tornam-se menos prevalentes entre pesquisadores com as carreiras mais longas, representando apenas 19% dos pesquisadores.

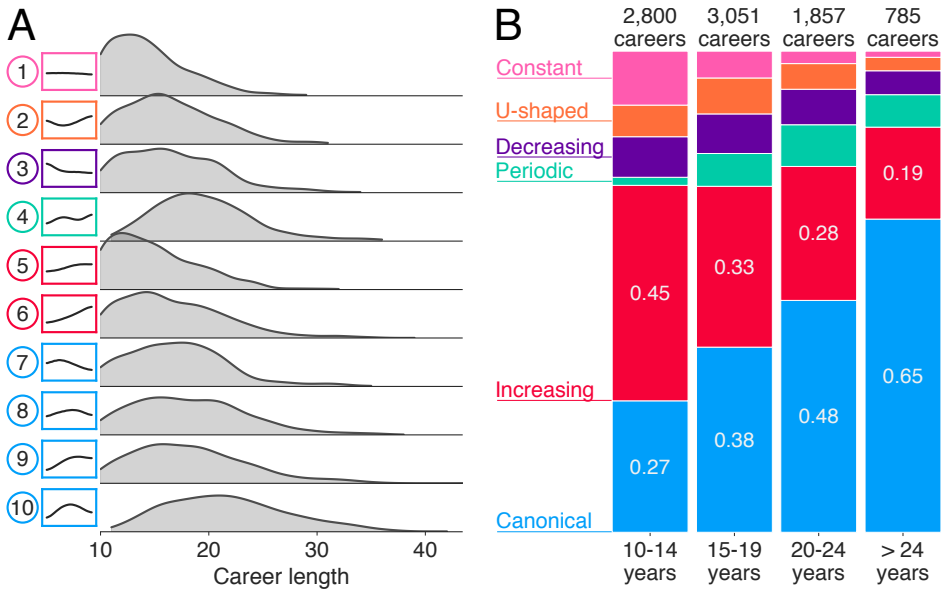

Figura 3.17: Efeito do comprimento da carreira na prevalência de padrões de produtividade. (A) Distribuições de probabilidade dos comprimentos de carreira de cada um dos dez grupos de trajetórias de produtividade. (B) Prevalência das seis categorias de padrão de produtividade considerando quatro intervalos de comprimento de carreira (10-14 anos, 15-19 anos, 20-24 anos e maior do que 24 anos).

As trajetórias com aspecto canônico apresentam o comportamento oposto. Apenas 27% dos pesquisadores com carreiras de 10-14 anos apresentam trajetórias de produtividade com aspecto canônico, enquanto esse padrão caracteriza 65% dos pesquisadores com mais de 24 anos de carreira. Mesmo combinados, carreiras constantes, em forma de U, decrescentes e periódicas ocorrem menos frequentemente do que carreiras crescentes e com aspecto canônico em todos os intervalos de comprimento. Mesmo assim, observamos que trajetórias constantes, em forma de U e decrescentes são relativamente mais comuns entre pesquisadores com carreiras curtas, enquanto carreiras periódicas são mais comuns entre pesquisadores com carreiras maiores do que 14 anos.

A Figura 3.18 mostra tendências de ocupação similares ao analisar os comportamentos individuais dos grupos contendo curvas crescentes e com aspecto canônico. Todavia, alguns grupos são mais prevalentes em determinados intervalos de comprimento. O padrão sempre crescente (grupo 6) é mais frequente do que o padrão crescente com taxas decrescentes (grupo 5) em todas os intervalos de comprimentos de carreira, mas especialmente entre os pesquisadores com as carreiras mais longas. A grande maioria dos pesquisadores que exibem carreiras de produtividade crescente com comprimento maior do que 24 anos pertence ao grupo 6. Entre as curvas com aspecto canônico, as trajetórias com pico no meio e no final da carreira (grupos 8 e 9, respectivamente) são os padrões mais comuns entre todos os intervalos de comprimento, exceto para as carreiras mais longas, para as quais trajetórias com pico pronunciado no meio da carreira (grupo 10) são as mais comuns. O padrão com pico no começo da carreira (grupo 8) é o mais raro entre todos os intervalos de comprimento,

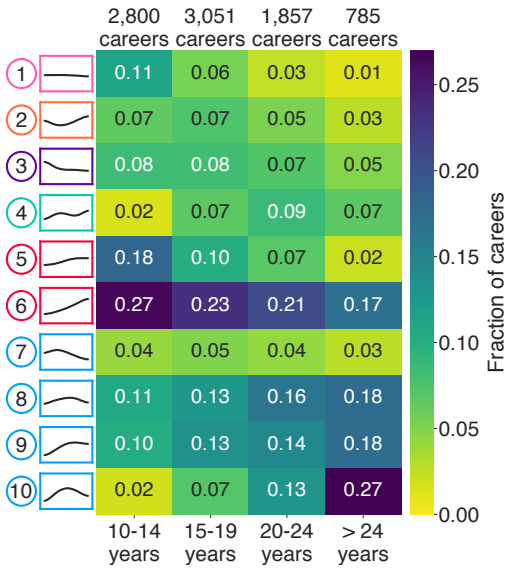

Figura 3.18: Efeitos de comprimento de carreira na prevalência de padrões de produtividade. A representação matricial mostra a prevalência dos padrões de produtividade em cada um dos dez grupos de padrão de produtividade considerando quatro intervalos de comprimento de carreira (10-14 anos, 15-19 anos, 20-24 anos e maior do que 24 anos).

exceto para os pesquisadores mais novos, sendo o único padrão com aspecto canônico cuja prevalência não cresce com o comprimento da carreira.

A Figura 3.19 mostra a prevalência das categorias de produtividade para oito disciplinas com mais de vinte pesquisadores em todos os intervalos de comprimento de carreira (agronomia, bioquímica, química, geociências, letras, matemática, medicina e física). A prevalência de curvas crescentes para os pesquisadores com as carreiras mais curtas é menor para pesquisadores da bioquímica, letras e física (34% a 41%) se comparada com a média geral de 45%. Para as outras cinco disciplinas (agronomia, química, geociências, matemática e medicina), a prevalência de padrões crescentes entre os pesquisadores com as carreiras mais curtas é ligeiramente maior (46% a 48%) do que a fração correspondente dos dados agregados. Entre os pesquisadores com as carreiras mais longas, observamos que cinco disciplinas

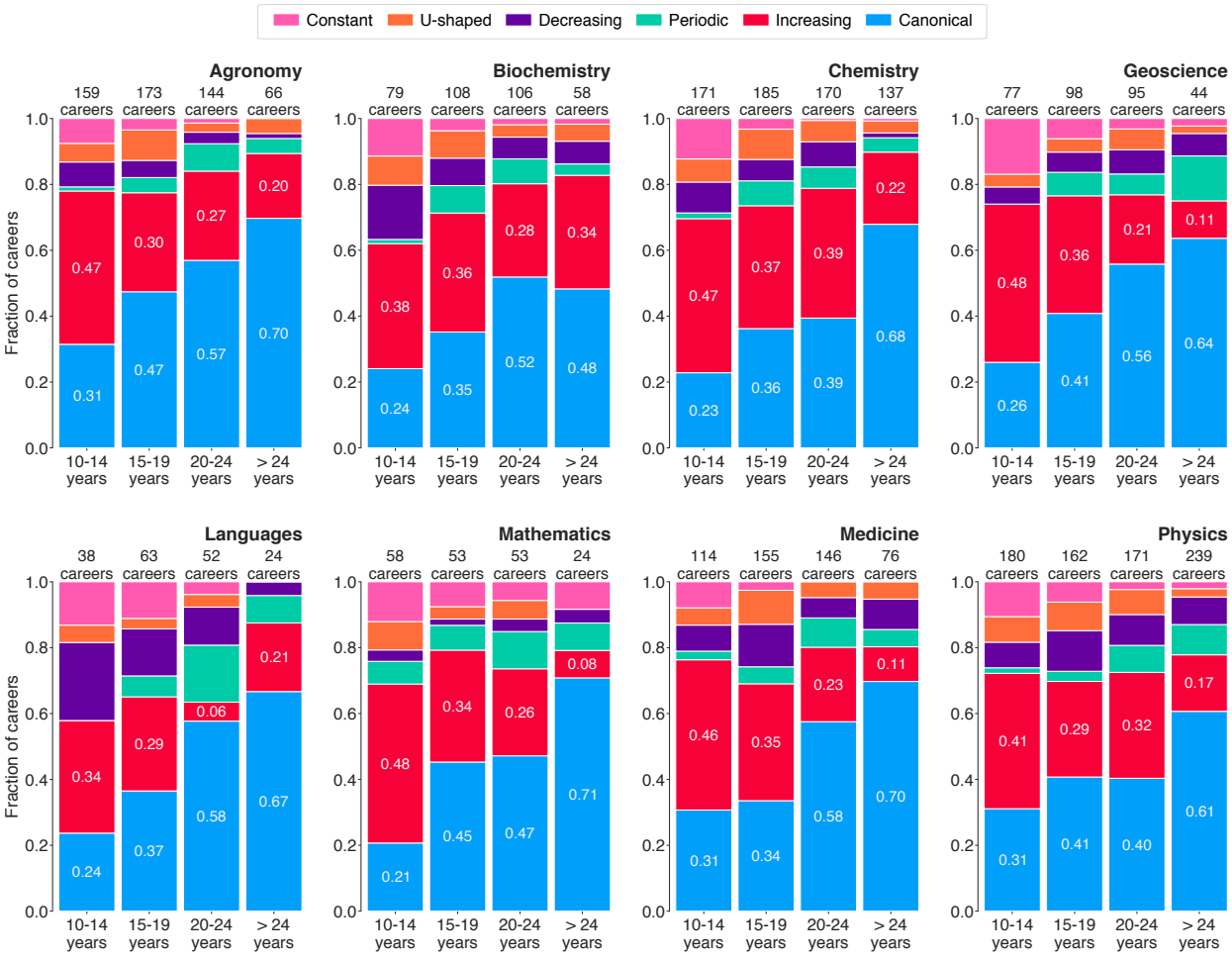

Figura 3.19: Prevalência das seis categorias de padrão de produtividade considerando diferentes disciplinas do nosso conjunto de dados. Os painéis mostram a prevalência dos padrões de produtividade considerando quatro intervalos de comprimento de carreira (10-14 anos, 15-19 anos, 20-24 anos e maior do que 24 anos) para oito disciplinas, que contêm cada uma pelo menos vinte pesquisadores em todos os intervalos de comprimento de carreira.

(agronomia, química, letras, matemática e medicina) têm maior prevalência (67% a 71%) e três disciplinas (bioquímica, geociências e física) têm menor prevalência (48% a 64%) de padrões com aspecto canônico do que a fração correspondente dos dados agregados (65%). A bioquímica é a única disciplina que não apresenta crescimento monotônico das frações da categoria de aspecto canônico com o comprimento de carreira, que está associado com frações constantes da categoria crescente entre os intervalos de comprimento de carreira. Exceto para matemática, que exibe uma maior prevalência de trajetórias constantes entre os intervalos de comprimento, todas as disciplinas seguem o mesmo padrão observado para o caso agregado. A prevalência de trajetórias em forma de U é similar ao padrão observado para o caso agregado, com apenas letras e matemática diferindo para o intervalo de carreiras mais longas. Trajetórias decrescentes mostram um declínio das frações similar ao caso agregado para quatro disciplinas (agronomia, bioquímica, química e letras), enquanto as outras quatro (geociências, matemática, medicina e física) apresentam frações aproximadamente constantes. A prevalência de curvas periódicas é menor para carreiras curtas e tende a crescer com o aumento do comprimento, de forma similar ao caso agregado.

A Figura 3.20 mostra os padrões para os grupos das categorias de padrões de produ-

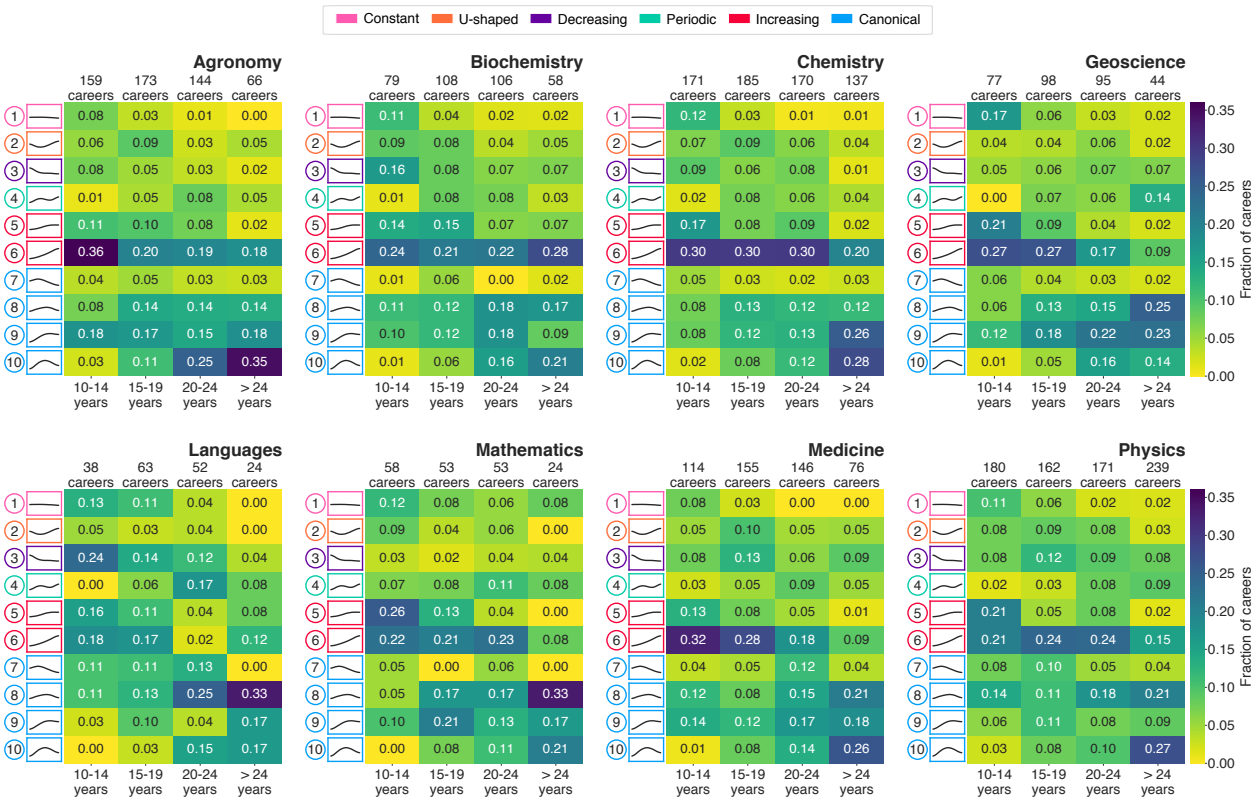

Figura 3.20: Prevalência dos dez grupos de padrão de produtividade considerando diferentes disciplinas do nosso conjunto de dados. Os painéis mostram a prevalência dos padrões de produtividade considerando quatro intervalos de comprimento de carreira (10-14 anos, 15-19 anos, 20-24 anos e maior do que 24 anos) para oito disciplinas, que contêm cada uma pelo menos vinte pesquisadores em todos os intervalos de comprimento de carreira.

tividade crescente e com aspecto canônico para as oito disciplinas com mais de vinte pesquisadores em todos os intervalos de comprimento de carreira. Similarmente aos padrões observados para o caso agregado, o padrão sempre crescente (grupo 6) é mais prevalente do que o padrão crescente com taxas decrescentes (grupo 5) entre todas os intervalos de comprimento e entre quase todas as disciplinas. A prevalência desses dois padrões também decresce entre os intervalos de comprimento para todas as disciplinas menos bioquímica. Na categoria com aspecto canônico (grupos 7 a 10), similarmente ao caso agregado, observamos que os grupos 8 e 9 tendem a ser os padrões com maior prevalência entre todos os intervalos de comprimento e entre a maioria das disciplinas. Entretanto, enquanto o padrão do grupo 10 é o mais prevalente para os pesquisadores com as carreiras mais longas de cinco disciplinas (agronomia, bioquímica, química, medicina e física) e o padrão do grupo 8 é o mais prevalente para os pesquisadores com as carreiras mais longas das três disciplinas restantes (geociências, letras e matemática). Como no caso agregado, o grupo 7 é o menos prevalente entre todos os intervalos de comprimento e disciplinas, exceto por letras que mostra frações significativamente maiores para os três primeiros intervalos de comprimento.

A Figura 3.21 mostra que o comprimento de carreira está diretamente relacionado ao ano de doutoramento de cada pesquisador e funciona como *proxy* para agrupar diferentes gerações de cientistas. De fato, a maioria dos pesquisadores com 10 a 14 anos de carreira concluíram seu doutorado depois do ano 2000, enquanto aqueles com mais de 25 anos de carreira concluíram seu doutorado antes dos anos 1990. Essas duas variáveis são quase perfeitamente correlacionadas. A correlação não é perfeita porque menos de 1% dos pesquisadores não atualizaram o currículo em 2016 (ano anterior à coleta dos dados). Os pesquisadores de nosso estudo representam diferentes gerações que, em cada etapa da carreira, estiveram sujeitos a condições socioeconômicas e culturais específicas, determinado nível base de conhecimento do campo científico e determinado nível de avanço técnico [55, 64]. Assim, a prevalência desigual de tipos de curvas de produtividade entre diferentes gerações pode parcialmente refletir as distintas culturas de pesquisa e publicação em cada período. Em particular, a

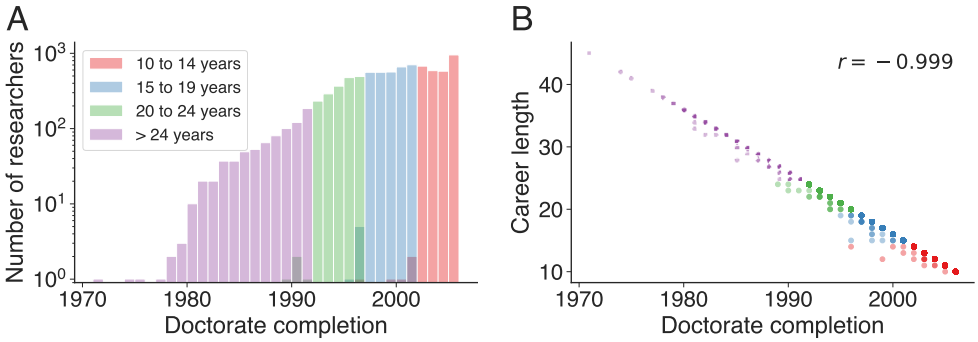

Figura 3.21: Relação do comprimento de carreira com o ano de doutoramento de cada pesquisador. (A) Histograma do ano de doutoramento para todos os pesquisadores do nosso estudo. (B) Associação entre comprimento de carreira e ano de doutoramento.

fração elevada de trajetórias crescentes para a geração mais jovem pode estar associada à pressão exacerbada sobre os cientistas para produzir em grandes quantidades [19, 69, 70], que é excepcionalmente maior sobre cientistas jovens [17]. Ao mesmo tempo, as carreiras de pesquisadores jovens não podem ser consideradas como carreiras completas, pois padrões emergindo depois de 10-14 anos de carreira ainda podem mudar. Por exemplo, parte dos padrões crescentes exibidos por jovens pesquisadores podem eventualmente representar apenas o início de carreiras com aspecto canônico.

A identificação precisa de efeitos geracionais na prevalência de padrões de carreira de produtividade requer um conjunto de dados com carreiras completas de cientistas de diferentes períodos, o que não é o caso de nosso estudo. Entretanto, podemos testar parcialmente essa hipótese analisando os anos iniciais das carreiras dos pesquisadores seniores e comparando a prevalência de seus padrões de produtividade com a prevalência do grupo de pesquisadores mais jovens. Para isso, aplicamos nosso procedimento de agrupamento ao conjunto de dados completo, mas considerando apenas os 14 primeiros anos dos pesquisadores com carreiras mais longas do que 24 anos. As Figuras A.21 e A.22 mostram que a melhor partição do Infomap é novamente formada por dez grupos. Agrupamos os grupos em seis categorias, com os padrões constante e decrescente (grupos 1 e 2) sendo incorporados num grupo único

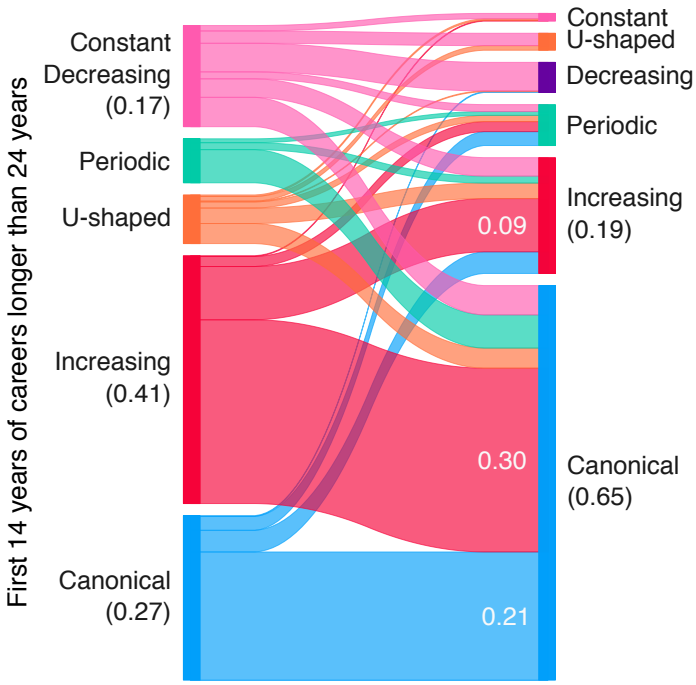

Figura 3.22: Comparação da prevalência dos padrões de produtividade para pesquisadores seniores considerando os anos iniciais de suas carreiras e suas carreiras completas. As barras na esquerda mostram as frações de cada padrão de produtividade considerando apenas os primeiros 14 anos das carreiras de pesquisadores com carreiras maiores do que 24 anos. As barras da direita mostram as frações de cada padrão de produtividade considerando suas carreiras completas. As conexões entre as barras da esquerda e da direita indicam os fluxos de migração entre padrões de produtividade.

(grupo 1 da Figura A.21). A Figura 3.22 mostra a prevalência dos padrões de produtividade associados ao começo da carreira de pesquisadores seniores e as trajetórias para as quais as trajetórias evoluem no final da carreira. Observamos que quase metade das carreiras de pesquisadores seniores classificada como canônicas são classificadas como crescentes em seu início. Apenas 9% dos pesquisadores seniores que exibem carreiras inicialmente crescentes conseguem manter esse padrão em estágios posteriores. Por outro lado, 78% dos pesquisadores seniores com carreiras inicialmente com aspecto canônico mantêm esse padrão em estágios posteriores. Cerca de 21% dos pesquisadores seniores com carreiras com aspecto canônico apresentam padrões inicialmente compatíveis com comportamentos constante/decrescente, periódico e em forma de U. Essas transições mais raras estão geralmente associadas com carreiras localizadas na borda entre duas ou mais comunidades, representando padrões mais complexos de produtividade, como ilustram os dois exemplos selecionados da Figura 3.23. O painel da esquerda ilustra um caso de um pesquisador que apresenta um padrão inicialmente classificado como em forma de U, mas que se torna um padrão com aspecto canônico. O painel da direita representa outra transição atípica em que a carreira é inicialmente classificada como constante/decrescente e depois se torna um padrão com aspecto canônico. A Figura 3.24 mostra as transições considerando os grupos individualmente. Constatamos que a maioria das transições entre padrões crescentes e com aspecto canônico acontece do grupo com padrão sempre crescente (grupo 7) para os grupos 9 e 10. Além disso, as transições entre padrões crescentes são mais frequentes do padrão sempre crescente do grupo 7 para o padrão crescente com taxa decrescente do grupo 6 (lado direito).

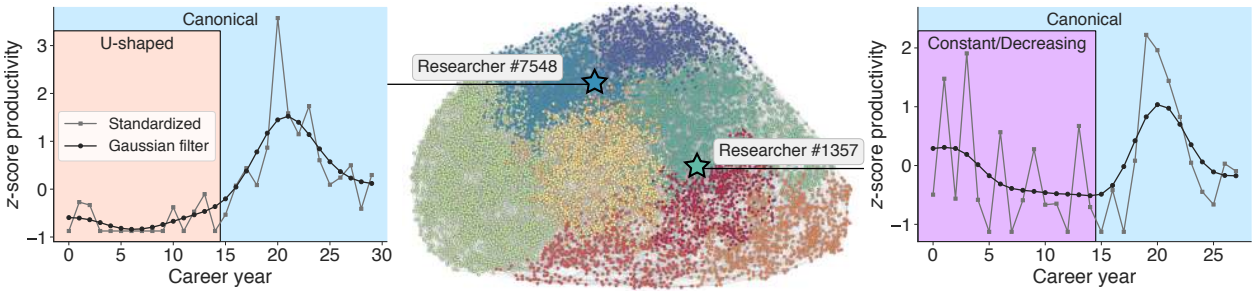

Figura 3.23: Exemplos de transições atípicas de padrões de produtividade encontrados nos primeiros 14 anos de pesquisadores para os padrões encontrados nas carreiras mais longas do que 24 anos e completas. O painel central mostra a representação em rede das similaridades entre os padrões de produtividade, destacando a localização dos pesquisadores escolhidos para ilustrar as transições atípicas. Os painéis laterais mostram dois exemplos de transições atípicas.

Se pesquisadores no início da carreira comportarem-se como os pesquisadores seniores, uma parcela maior de padrões com aspecto canônico deve emergir no futuro. Todavia, não podemos ignorar os potenciais efeitos geracionais na comparação entre os padrões dos cientistas jovens e dos anos iniciais de cientistas experientes. Nossos resultados mostram que padrões crescentes são 10% mais frequentes entre pesquisadores jovens, enquanto pa-

drões periódicos são três vezes mais frequentes nos anos iniciais de pesquisadores seniores, como mostram as Figuras 3.17B e 3.22. Ao mesmo tempo, essas diferenças no começo da carreira são relativamente pequenas, sugerindo que as mudanças estruturais na empreitada científica [6, 18, 19] podem ter um impacto pequeno na trajetória de produtividade de pesquisadores.

Com isso, encerramos a apresentação dos resultados referentes à investigação sobre padrões universais de produtividade em carreiras científicas. No próximo capítulo, discutiremos os resultados apresentados nesta tese.

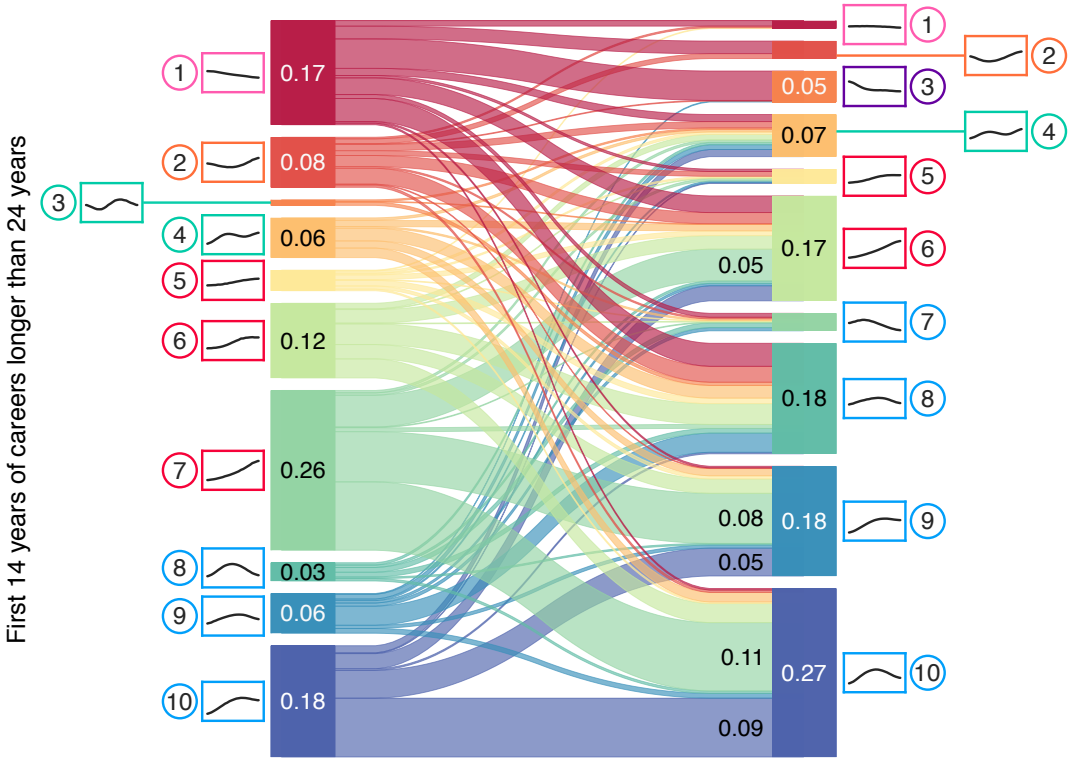

Figura 3.24: Comparação da prevalência dos grupos de padrões de produtividade para pesquisadores seniores considerando os anos iniciais de suas carreiras e suas carreiras completas. As barras na esquerda mostram as frações de cada padrão de produtividade considerando apenas os primeiros 14 anos das carreiras de pesquisadores com carreiras maiores do que 24 anos. As barras da direita mostram as frações de cada padrão de produtividade considerando suas carreiras completas. As conexões entre as barras da esquerda e da direita indicam os fluxos de migração entre padrões de produtividade.

CAPÍTULO 4

Discussão e perspectivas

Nesta tese, investigamos aspectos variados da produção científica de pesquisadores brasileiros. Utilizamos a Plataforma Lattes, que contém os currículos acadêmicos dos pesquisadores brasileiros, como nossa fonte primária de dados. A existência de currículos científicos individuais permitiu a identificação de cada pesquisador de maneira única, resolvendo um dos principais problemas da literatura de ciência da ciência, a desambiguação de nomes [3]. Além disso, destacamos a ampla abrangência da base de dados, que engloba pesquisadores das mais diversas disciplinas, permitindo análises de caráter mais geral.

No Capítulo 2, investigamos a associação entre a produtividade científica anual e o impacto de jornal para mais de seis mil pesquisadores brasileiros bolsistas do CNPq. Nossos resultados exploram essa associação entre disciplinas, estágios da carreira e distinguem pesquisadores com performances *outliers* de não *outliers*. Em contraste com trabalhos anteriores sobre o assunto, nossos resultados levam explicitamente em consideração a inflação temporal dos indicadores bibliométricos, o efeito de escala no prestígio médio de jornal e práticas específicas de cada disciplina por meio de *scores* robustos de padronização. Esse procedimento permitiu a construção do plano prestígio de jornal *versus* produtividade: uma representação direta e coerente das performances dos pesquisadores em impacto de jornal e produtividade. Dessa representação, categorizamos os pesquisadores entre *outliers* e não *outliers* e dividimos os pesquisadores *outliers* em três categorias: hiperprolíficos (*outliers* apenas em produtividade), perfeccionistas (*outliers* apenas em impacto de jornal) e hiperprolífico-perfeccionistas (*outliers* simultaneamente em impacto de jornal e produtividade).

Pesquisadores com performance *outlier* compõem 30% do total de acadêmicos do nosso conjunto de dados, sendo a performance *outlier* em apenas um ano da carreira (47,6% dos casos) o comportamento mais comum. Entre os *outliers*, a vasta maioria dos pesquisa-

dores é exclusivamente hiperprolífica ou exclusivamente perfeccionista. Apesar disso, 16 pesquisadores extremamente hiperprolíficos apresentam anos da carreira apenas no setor hiperprolífico-perfeccionista quando suas performances estão acima de um limiar de produtividade ($P > 27.7$). Apenas 14.4% dos pesquisadores *outliers* conseguem ser hiperprolíficos e perfeccionistas em suas carreiras e somente 6.7% conseguem ser hiperprolífico-perfeccionistas. O grupo de 14.4% de pesquisadores *outliers* (261 indivíduos) não tem um setor *outlier* preferencial, mostra níveis de produtividade maiores que pesquisadores exclusivamente hiperprolíficos ou perfeccionistas e publica em jornais de maior prestígio em comparação com pesquisadores exclusivamente hiperprolíficos ou perfeccionistas. Além disso, encontramos que um aumento no número de anos hiperprolíficos reduz a probabilidade de performar como perfeccionista em nosso conjunto de dados, exceto para engenharia dos materiais em que a relação não é significativa. Essa associação negativa varia entre disciplinas, com a matemática apresentando o maior efeito negativo e a física apresentando o efeito mais brando. Conjuntamente, esses achados corroboram a associação negativa entre produtividade e prestígio de jornal em níveis *outliers* de ambas as quantidades. Em outras palavras, é extremamente difícil para os pesquisadores manterem níveis extremamente altos de produtividade ao mesmo tempo em que publicam em jornais de prestígio elevadíssimo.

Também exploramos os padrões de carreira no curto prazo em relação à produtividade e ao impacto de jornal. Com esse objetivo, estimamos o excesso de transições entre setores do plano prestígio de jornal *versus* produtividade durante anos consecutivos das carreiras de pesquisadores *outliers* e não *outliers*. Identificamos um comportamento persistente em que pesquisadores tendem a permanecer no mesmo setor do plano e assim mostrar performances similares em anos consecutivos. Esse comportamento persistente concorda com resultados obtidos anteriormente na literatura [51, 53, 54, 61]. Transições entre níveis similares de produtividade e prestígio de jornal ocorrem tão frequentemente como ao acaso. Por outro lado, transições entre setores do plano com níveis diferentes de produtividade e impacto de jornal ocorrem muito menos frequentemente que ao acaso, indicando que pesquisadores são aversos a mudanças simultâneas de seus níveis de produtividade e impacto de jornal em anos consecutivos da carreira.

Acreditamos que tanto a aversão a mudanças simultâneas na produtividade e no impacto de jornal quanto a persistência na manutenção de performances similares em anos consecutivos sugerem a adoção de determinadas estratégias de pesquisa em que os pesquisadores optam por táticas focadas em produtividade ou focadas em impacto de jornal [36]. Para manter os níveis de produtividade, os acadêmicos podem adotar estratégias que consistem em expandir colaborações, evitar jornais de alto impacto, dividir seus resultados em vários artigos e selecionar temas de pesquisa mais tradicionais [36]. De outra forma, estratégias focadas em impacto consistem em realizar colaborações somente quando benéfico para a pesquisa, selecionar jornais de alto impacto como primeira opção, publicar os resultados

com maximização do entendimento em mente e escolher temas de pesquisa mais inovadores e arriscados [36]. Além disso, o comportamento persistente indica que essas estratégias de publicação persistem como um hábito e possivelmente refletem características individuais e convenções culturais dos grupos de pesquisa. Porém, investigações mais aprofundadas são necessárias para identificar explicitamente quais são esses hábitos e quais são os mecanismos associados com sua adoção.

Investigamos a tendência média do prestígio de jornal e da produtividade no decorrer de carreiras científicas para todas as disciplinas. Primeiramente, identificamos que o valor médio do prestígio de jornal é ligeiramente maior nos estágios iniciais da carreira com uma tendência decrescente sutil com o passar dos anos para a maioria das disciplinas. Por outro lado, a produtividade média tende a crescer com a progressão da carreira para todas as disciplinas. Estudamos também o efeito do ano da carreira na ocupação dos setores do plano prestígio de jornal *versus* produtividade para cada disciplina. Nossos resultados indicam que cada disciplina apresenta frações de ocupação específicas nesses setores, refletindo as diferentes práticas de publicação vigentes em cada campo do conhecimento. Porém, encontramos que setores de baixa produtividade ($I{-}P{-}$ ou $I{+}P{-}$) são mais povoados durante estágios iniciais das carreiras dos pesquisadores de todas as disciplinas. Também identificamos uma tendência de ocupação crescente de setores de alta produtividade, incluindo o setor hiperprolífico ($P{+}{+}$), em estágios posteriores da carreira para praticamente todas as disciplinas. De modo oposto, os acadêmicos alcançam mais frequentemente performances perfeccionistas ou com prestígio de jornal acima da média em estágios iniciais da carreira. É importante ressaltar que essa tendência no período inicial da carreira pode refletir um efeito de seleção, pois todos os pesquisadores em nosso conjunto de dados pertencem à classe de bolsistas do CNPq. Verificar se essas tendências se manteriam para outros tipos de acadêmicos é uma questão interessante que pesquisas futuras podem abordar. O aumento da produtividade com a progressão da carreira também foi verificada por Sinatra *et al.* [72], podendo refletir uma série de acontecimentos que tendem a ser habituais na progressão de carreiras científicas, tal como maior familiaridade com os temas de pesquisa [25], maior disponibilidade de recursos financeiros [25, 138] e maior quantidade de convites para elaboração de artigos de revisão [25]. Similarmente, a emergência de anos hiperprolíficos em estágios posteriores da carreira pode coincidir com a ocupação de altas posições em centros de pesquisa, o que poderia aumentar as taxas de publicação em grande quantidade, pois existe uma tradição em algumas disciplinas (por exemplo, em ciências médicas e da vida) de incluir a chefia de laboratórios científicos em todas as publicações [139].

Os nossos resultados também mostraram que a relação entre produtividade e impacto de jornal para pesquisadores não *outliers* é similar àquela observada para pesquisadores que alcançam performances *outliers*. Para os pesquisadores não *outliers*, empregamos um modelo bayesiano hierárquico que leva em consideração a heterogeneidade dos comporta-

mentos individuais dos pesquisadores e identifica um padrão emergente para cada disciplina. Encontramos uma associação negativa para a maioria das disciplinas ao considerar apenas pesquisadores não *outliers*. No entanto, a intensidade dessa associação varia entre as disciplinas. A física apresenta a associação mais negativa e a matemática apresenta a associação mais branda da produtividade com prestígio de jornal. Verificamos que, mesmo que o ano da carreira também seja negativamente correlacionado com o impacto de jornal, a associação geral negativa entre impacto de jornal e produtividade não é significativamente afetada por esse fator de confusão. De certa forma, esses resultados contradizem a teoria de Nijstad *et al.* [29] para criatividade denominada "modelo do caminho duplo para criatividade" ("dual pathway to creativity model" [29]), que dita que a criatividade – concebida como ideias inovadoras e adequadas – pode ser alcançada por meio dos caminhos de flexibilidade (uso de uma gama de ideias para gerar novas ideias) e de persistência (exploração exaustiva do mesmo tema). De acordo com essa teoria, os pesquisadores com alta produtividade deveriam estar explorando e associando vários temas e, assim, criando novas ideias criativas pelo caminho da flexibilidade ou trabalhando intensivamente na mesma temática até que ideias criativas emerjam pelo caminho da persistência. Desse modo, como a produtividade não se correlaciona positivamente com o impacto de jornal, os indicadores JIF e SJR podem não ser os indicadores mais adequados para avaliar a criatividade de trabalhos acadêmicos ou, ainda, a teoria de Nijstad *et al.* não se enquadra bem com nossas observações.

No Capítulo 3, realizamos uma análise abrangente das trajetórias de produtividade de mais de oito mil pesquisadores de 56 disciplinas acadêmicas. Diferentemente de investigações passadas que se basearam em disciplinas específicas [60–62, 65], utilizaram dados agregados de pesquisadores [49–55, 57, 58, 60–64] ou assumiram formas particulares de trajetórias de produtividade [55, 56, 65, 66], estimamos as similaridades entre todos os pares de trajetórias levando em consideração a inflação, a diferença em escala e a aleatoriedade nas carreiras de produtividade. Nossa abordagem revela grupos de padrões que são internamente consistentes, mais coesos do que modelos nulos, robustos a diferentes escolhas de métodos de agrupamento, robustos em validações por amostragem e semanticamente congruentes com validação humana. Além disso, nosso procedimento de agrupamento resultou em uma representação de rede em que pesquisadores e grupos com padrões similares de produtividade estão conectados e localizados em regiões próximas. Revelamos uma variedade de padrões de produtividade que vão além da narrativa tradicional e podem ser classificados em seis categorias universais: constante, em forma de U, decrescente, periódica, crescente e com aspecto canônico. Combinados, padrões constantes, em forma de U, decrescentes e periódicos representam aproximadamente um quarto dos pesquisadores, enquanto a maioria dos pesquisadores exibem padrões crescentes ou com aspecto canônico.

Investigamos possíveis efeitos de tamanho de carreira e de geração na prevalência dos diferentes padrões de produtividade. Nossa análise revelou que todos os padrões englobam

carreiras num intervalo amplo de comprimento de carreira. Porém, carreiras crescentes são mais comuns para pesquisadores com carreiras curtas, que também são pesquisadores mais jovens, enquanto carreiras com aspecto canônico são mais frequentes para pesquisadores com carreiras longas, que também são pesquisadores seniores. Hipotetizamos que a maior incidência de padrões crescentes de produtividade entre cientistas jovens pode estar relacionada com mudanças na empreitada científica, como aumento do número de colaborações [67], pressão sobre os acadêmicos para produzir em grandes quantidades [6, 18, 19] (principalmente sobre os mais jovens [17]), bem como a natureza incompleta de suas carreiras. Como a clara identificação de efeitos geracionais na prevalência de padrões de produtividade requer dados de carreiras científicas completas de pesquisadores de várias gerações, pudemos testar apenas parcialmente essa hipótese comparando a prevalência dos padrões nos anos iniciais das carreiras de pesquisadores seniores com a prevalência de pesquisadores jovens. Nossos resultados indicam que quase metade das curvas com aspecto canônico dos pesquisadores seniores são classificadas como carreiras crescentes em sua parte inicial. Além disso, apenas 9% dos cientistas seniores que exibiram carreiras inicialmente crescentes puderam sustentar esse padrão com o passar do tempo. O padrão observado por pesquisadores seniores não necessariamente dita a trajetória de pesquisadores mais jovens. Entretanto, se os pesquisadores jovens seguirem as mesmas tendências dos pesquisadores seniores, a prevalência das curvas com aspecto canônico é provavelmente subestimada.

Mesmo que possivelmente subestimadas, curvas com aspecto canônico – definidas como carreiras com pico único na produtividade – são o padrão de produtividade mais prevalente, representando quase dois quintos dos pesquisadores e quase dois terços quando consideramos apenas pesquisadores seniores. Enquanto esse resultado de certa forma suporta a narrativa canônica de produtividade científica, também observamos que menos de 5% dos pesquisadores em nosso estudo estritamente correspondem à definição de Lehman de "narrativa canônica de produtividade" [49] e exibem curvas de produtividade que "inicialmente crescem rapidamente e, após atingir um pico em seu início, diminuem lentamente" [49] (em tradução livre). Esses pesquisadores pertencem ao grupo 7, que é apenas um dos quatro grupos de padrão com aspecto canônico. Esse grupo tem comprimento de carreira com mediana de 17 anos e apresenta o pico em produtividade aproximadamente 6 anos após o doutoramento. Os outros três grupos (8, 9 e 10) representam quase 90% dos pesquisadores com aspecto canônico. Esses grupos têm carreiras ligeiramente mais longas, mas cujo máximo está localizado aproximadamente 12 anos após o doutoramento. Apesar da subjetividade na definição de máximo no início da carreira na definição de Lehman, nossa pesquisa mostra que o pico em produtividade tem maior probabilidade de ocorrer em estágios intermediários das carreiras científicas. Pontuamos que a ascensão e declínio na produtividade de pesquisadores em nosso estudo é muito mais diversa do que a definição de Lehman.

Ao concentrar nossa atenção nos anos iniciais das carreiras, verificamos que a maioria dos

pesquisadores de nosso estudo exibe padrões crescentes de produtividade. Essa tendência emerge entre os grupos 4 a 10 e representa aproximadamente 80% dos pesquisadores. A alta incidência de trajetórias crescentes em estágios iniciais das carreiras pode ser atribuída à maneira como os processos de financiamento e de contratação são realizados no meio acadêmico. Pesquisas passadas mostraram que a produtividade tem um papel central na colocação profissional [140] e no acesso a recursos financeiros necessários para pesquisa [6,7,20–22,141]. Dessa maneira, é provável que a prevalência de padrões de produtividade inicialmente crescentes reflita os critérios de seleção que regularmente favorecem pesquisadores mais produtivos. Além disso, metade dos pesquisadores em nossa amostra (grupos 7 a 10 e 3) exibe um declínio na produtividade que é mais frequentemente observado após a porção intermediária das carreiras. Várias hipóteses podem explicar esse padrão. Por exemplo, a consolidação do prestígio acadêmico em períodos avançados das carreiras pode reduzir a urgência na manutenção da alta produtividade [142]. A tensão entre o tempo gasto desempenhando pesquisa, que é discutivelmente maior para pesquisadores jovens, e tarefas administrativas, que são, por sua vez, mais comuns para pesquisadores seniores, pode também ser parcialmente responsável pelo declínio na produtividade em períodos posteriores das carreiras [58,143,144]. A maternidade e paternidade também podem contribuir para um declínio na produtividade uma vez que o tempo dedicado à pesquisa naturalmente diminui nessas circunstâncias [145]. Por fim, o quase que inevitável declínio do potencial intelectual com o tempo também pode estar relacionado com a redução da produtividade com o avanço da carreira [56].

De maneira conjunta, as duas investigações apresentadas nessa tese esclarecem aspectos importantes sobre os indicadores prestígio de jornal e produtividade. Esperamos que nossos resultados possam contribuir com a construção de um processo de avaliação na ciência mais justo e mais compreensivo e, ao mesmo tempo, mais efetivo. A utilização desses indicadores de maneira acrítica e sem critérios pode introduzir vieses e reproduzir desigualdades que são inerentes da sociedade em que vivemos. Por isso, esperamos que esses trabalhos também possam inspirar investigações futuras que contribuam e elucidem aspectos importantes da produção científica, incluindo, padrões de colaboração [68,146,147], análises de gênero [146–150], etnia [151,152] e nacionalidade [153].

Referências Bibliográficas

[1] Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y. & Stanley, H. E. The science of science: From the perspective of complex systems. *Physics Reports* **714**, 1–73 (2017).

[2] Fortunato, S. *et al.* Science of science. *Science* **359**, eaao0185 (2018).

[3] Wang, D. & Barabási, A. *The Science of Science* (Cambridge University Press, 2021).

[4] Azoulay, P., Zivin, J. S. G. & Manso, G. Incentives and creativity: Evidence from the academic life sciences. *The RAND Journal of Economics* **42**, 527–554 (2011).

[5] Bromham, L., Dinnage, R. & Hua, X. Interdisciplinary research has consistently lower funding success. *Nature* **534**, 684–687 (2016).

[6] Meirmans, S., Butlin, R. K., Charmantier, A., Engelstädter, J., Groot, A. T., King, K. C., Kokko, H., Reid, J. M. & Neiman, M. Science policies: How should science funding be allocated? An evolutionary biologists' perspective. *Journal of Evolutionary Biology* **32**, 754–768 (2019).

[7] Wilsdon, J. *The Metric Tide: Independent Review of the Role of Metrics in Research Assessment and Management* (Sage, 2016).

[8] Wessely, S. Peer review of grant applications: What do we know? *The Lancet* **352**, 301–305 (1998).

[9] Smith, R. Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine* **99**, 178–182 (2006).

[10] Balietti, S., Goldstone, R. L. & Helbing, D. Peer review and competition in the art exhibition game. *Proceedings of the National Academy of Sciences* **113**, 8414–8419 (2016).

[11] Bornmann, L. & Mutz, R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology* **66**, 2215–2222 (2015).

[12] Ioannidis, J. P., Boyack, K. W. & Klavans, R. Estimates of the continuously publishing core in the scientific workforce. *PloS one* **9**, e101698 (2014).

[13] Traag, V. A. & Waltman, L. Systematic analysis of agreement between metrics and peer review in the UK REF. *Palgrave Communications* **5** (2019).

[14] Cameron, B. D. Trends in the usage of ISI bibliometric data: Uses, abuses, and implications. *portal: Libraries and the Academy* **5**, 105–125 (2005).

[15] Gagolewski, M. Scientific impact assessment cannot be fair. *Journal of Informetrics* **7**, 792–802 (2013).

[16] Siudem, G., Żogała-Siudem, B., Cena, A. & Gagolewski, M. Three dimensions of scientific impact. *Proceedings of the National Academy of Sciences* **117**, 13896–13900 (2020).

[17] Powell, K. Young, talented and fed-up: Scientists tell their stories. *Nature* **538**, 446–449 (2016).

[18] Moher, D., Naudet, F., Cristea, I. A., Miedema, F., Ioannidis, J. P. & Goodman, S. N. Assessing scientists for hiring, promotion, and tenure. *PLoS Biology* **16**, e2004089 (2018).

[19] Schimanski, L. A. & Alperin, J. P. The evaluation of scholarship in academic promotion and tenure processes: Past, present, and future. *F1000Research* **7**, 1–21 (2018).

[20] San Francisco Declaration on Research Assessment. Disponível em: https://sfdora.org/read/. Último acesso em: 21 de agosto de 2023.

[21] Hicks, D., Wouters, P., Waltman, L., Rijcke, S. D. & Rafols, I. Bibliometrics: The Leiden Manifesto for research metrics. *Nature* **520**, 429–431 (2015).

[22] Nuffield Council on Bioethics. The findings of a series of engagement activities exploring the culture of scientific research in the UK. Disponível em: https://www.nuffieldbioethics.org/assets/pdfs/The-culture-of-scientific-research-report.pdf. Último acesso em: 21 de agosto de 2023.

[23] Dennis, W. Productivity among american psychologists. *American Psychologist* **9**, 191–194 (1954).

[24] White, K. G. & White, M. J. On the relation between productivity and impact. *Australian Psychologist* **13**, 369–374 (1978).

[25] Lawani, S. Some bibliometric correlates of quality in scientific research. *Scientometrics* **9**, 13–25 (1986).

[26] Simonton, D. K. *Scientific genius: A psychology of science* (Cambridge University Press, 1988).

[27] Feist, G. J. Quantity, quality, and depth of research as influences on scientific eminence: Is quantity most important? *Creativity Research Journal* **10**, 325–335 (1997).

[28] Haslam, N. & Laham, S. M. Quality, quantity, and impact in academic publication. *European Journal of Social Psychology* **40**, 216–220 (2010).

[29] Nijstad, B. A., Dreu, C. K. D., Rietzschel, E. F. & Baas, M. The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. *European Review of Social Psychology* **21**, 34–77 (2010).

[30] Bosquet, C. & Combes, P.-P. Are academics who publish more also more cited? Individual determinants of publication and citation records. *Scientometrics* **97**, 831–857 (2013).

[31] Abramo, G., Cicero, T. & D'Angelo, C. A. Are the authors of highly cited articles also the most productive ones? *Journal of Informetrics* **8**, 89–97 (2014).

[32] Sandström, U. & van den Besselaar, P. Quantity and/or quality? The importance of publishing many papers. *PloS one* **11**, e0166149 (2016).

[33] Larivière, V. & Costas, R. How many is too many? On the relationship between research productivity and impact. *PloS one* **11**, e0162709 (2016).

[34] Garousi, V. & Fernandes, J. M. Quantity versus impact of software engineering papers: A quantitative study. *Scientometrics* **112**, 963–1006 (2017).

[35] Michalska-Smith, M. J. & Allesina, S. And, not or: Quality, quantity in scientific publishing. *PloS one* **12**, e0178074 (2017).

[36] Kolesnikov, S., Fukumoto, E. & Bozeman, B. Researchers' risk-smoothing publication strategies: Is productivity the enemy of impact? *Scientometrics* **116**, 1995–2017 (2018).

[37] Bornmann, L. & Tekles, A. Productivity does not equal usefulness. *Scientometrics* **118**, 705–707 (2019).

[38] Forthmann, B., Leveling, M., Dong, Y. & Dumas, D. Investigating the quantity–quality relationship in scientific creativity: An empirical examination of expected residual variance and the tilted funnel hypothesis. *Scientometrics* **124**, 2497–2518 (2020).

[39] Larivière, V. & Sugimoto, C. R. *The Journal Impact Factor: A Brief History, Critique, and Discussion of Adverse Effects* (Springer, 2019).

[40] McKiernan, E. C., Schimanski, L. A., Nieves, C. M., Matthias, L., Niles, M. T. & Alperin, J. P. Meta-research: Use of the journal impact factor in academic review, promotion, and tenure evaluations. *eLife* **8**, e47338 (2019).

[41] CAPES – Metodologia do Qualis Referência - Quadriênio 2017-2020. Disponível em: https://www.gov.br/capes/pt-br/acesso-a-informacao/acoes-e-programas/avaliacao/avaliacao-quadrienal/metodologia-do-qualis-referencia-quadrienio-2017-2020. Último acesso em: 21 de agosto de 2023.

[42] RESOLUÇÃO N.º 058/2020-CAD - Universidade Estadual de Maringá. Disponível em: http://www.drh.uem.br/res/ResoluÃğÃčo-058-2020-CAD.pdf. Último acesso em: 21 de agosto de 2023.

[43] Chamada CNPq N° 06/2019 - Bolsas de Produtividade em Pesquisa. Disponível em: http://memoria.cnpq.br/chamadas-publicas?p_p_id=resultadosportlet_WAR_resultadoscnpqportlet_INSTANCE_0ZaM&filtro=encerradas&detalha=chamadaDivulgada&idDivulgacao=8722. Último acesso em: 21 de agosto de 2023.

[44] Bornmann, L. & Leydesdorff, L. Skewness of citation impact data and covariates of citation distributions: A large-scale empirical analysis based on Web of Science data. *Journal of Informetrics* **11**, 164–175 (2017).

[45] Traag, V. A. Inferring the causal effect of journals on citations. *Quantitative Science Studies* **2**, 496–504 (2021).

[46] Kim, L., Portenoy, J. H., West, J. D. & Stovel, K. W. Scientific journals still matter in the era of academic search engines and preprint archives. *Journal of the Association for Information Science and Technology* **71**, 1218–1226 (2020).

[47] Correa, J. C., Laverde-Rojas, H., Tejada, J. & Marmolejo-Ramos, F. The Sci-Hub effect on papers' citations. *Scientometrics* **127**, 99–126 (2021).

[48] Waltman, L. & Traag, V. A. Use of the journal impact factor for assessing individual articles need not be statistically wrong. *F1000Research* **9**, 1–27 (2020).

[49] Lehman, H. C. *Age and Achievement* (Princeton University Press, 1953).

[50] Lehman, H. C. Men's creative production rate at different ages and in different countries. *The Scientific Monthly* **78**, 321–326 (1954).

[51] Dennis, W. Age and productivity among scientists. *Science* **123**, 724–725 (1956).

[52] Behymer, C. E. & Blackburn, R. T. *Environmental and Personal Attributes Related to Faculty Productivity* (ERIC Clearinghouse, 1975).

[53] Cole, S. Age and scientific performance. *American Journal of Sociology* **84**, 958–977 (1979).

[54] Horner, K. L., Rushton, J. P. & Vernon, P. A. Relation between aging and research productivity of academic psychologists. *Psychology and Aging* **1**, 319—324 (1986).

[55] Levin, S. G. & Stephan, P. E. Research productivity over the life cycle: Evidence for academic scientists. *The American Economic Review* **81**, 114–132 (1991).

[56] Simonton, D. K. Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychological Review* **104**, 66–89 (1997).

[57] Gingras, Y., Lariviere, V., Macaluso, B. & Robitaille, J.-P. The effects of aging on researchers' publication and citation patterns. *PloS one* **3**, e4048 (2008).

[58] Rørstad, K. & Aksnes, D. W. Publication rate expressed by age, gender and academic position – a large-scale analysis of Norwegian academic staff. *Journal of Informetrics* **9**, 317–333 (2015).

[59] Sunahara, A. S., Perc, M. & Ribeiro, H. V. Association between productivity and journal impact across disciplines and career age. *Physical Review Research* **3**, 033158 (2021).

[60] Spake, C. S., Zeyl, V. G., Crozier, J. W., Rao, V. & Kalliainen, L. K. An analysis of publication trajectory in plastic surgery across the decades. *Journal of Plastic, Reconstructive & Aesthetic Surgery* **75**, 439–488 (2022).

[61] Over, R. Does research productivity decline with age? *Higher Education* **11**, 511–520 (1982).

[62] Over, R. Is age a good predictor of research productivity? *Australian Psychologist* **17**, 129–139 (1982).

[63] Pelz, D. C. & Andrews, F. M. *Scientists in Organizations: Productive Climates for Research and Development* (John Wiley, 1966).

[64] Bayer, A. E. & Dutton, J. E. Career age and research-professional activities of academic scientists. *The Journal of Higher Education* **48**, 259–282 (1977).

[65] Way, S. F., Morgan, A. C., Clauset, A. & Larremore, D. B. The misleading narrative of the canonical faculty productivity trajectory. *Proceedings of the National Academy of Sciences* **114**, E9216–E9223 (2017).

[66] Rinaldi, S., Cordone, R. & Casagrandi, R. Instabilities in creative professions: A minimal model. *Nonlinear Dynamics, Psychology, and Life Sciences* **4**, 255–273 (2000).

[67] Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).

[68] Araújo, E. B., Moreira, A. A., Furtado, V., Pequeno, T. H. & Andrade, J. S., Jr. Collaboration networks from a large CV database: dynamics, topology and bonus impact. *PloS one* **9**, e90537 (2014).

[69] Miller, A. N., Taylor, S. G. & Bedeian, A. G. Publish or perish: Academic life as management faculty live it. *Career Development International* **16**, 422–445 (2011).

[70] Van Dalen, H. P. & Henkens, K. Intended and unintended consequences of a publish-or-perish culture: A worldwide survey. *Journal of the American Society for Information Science and Technology* **63**, 1282–1293 (2012).

[71] de Solla Price, D. J. *Little Science, Big Science* (Columbia University Press, 1963).

[72] Sinatra, R., Wang, D., Deville, P., Song, C. & Barabási, A.-L. Quantifying the evolution of individual scientific impact. *Science* **354**, aaf5239 (2016).

[73] Sunahara, A. S., Perc, M. & Ribeiro, H. V. Universal productivity patterns in research careers. *Physical Review Research* **5**, 043203 (2023).

[74] Ribeiro, H. V., Sunahara, A. S., Sutton, J., Perc, M. & Hanley, Q. S. City size and the spreading of COVID-19 in Brazil. *PloS one* **15**, e0239699 (2020).

[75] Sunahara, A. S., Pessa, A. A., Perc, M. & Ribeiro, H. V. Complexity of the COVID-19 pandemic in Maringá. *Scientific Reports* **13**, 12695 (2023).

[76] Agresti, A. *Categorical Data Analysis* (Wiley, 2003).

[77] Unpingco, J. *Python for Probability, Statistics, and Machine Learning* (Springer, 2016).

[78] Kutner, M. H. *Applied Linear Statistical Models* (McGraw-Hill Irwin, 2005).

[79] Hosmer, D. W. & Lemeshow, S. *Applied Logistic Regression* (Wiley, 2004).

[80] Myung, I. J. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology* **47**, 90–100 (2003).

[81] Seabold, S. & Perktold, J. statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference* (2010).

[82] Rencher, A. C. & Schaalje, G. B. *Linear Models in Statistics* (Wiley, 2008).

[83] Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E. D., Robinson, B. S., Hodgson, D. J. & Inger, R. A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* **6**, e4794 (2018).

[84] Laird, N. M. & Ware, J. H. Random-effects models for longitudinal data. *Biometrics* **38**, 963–974 (1982).

[85] Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**, 1–48 (2015).

[86] Downey, A. *Think Bayes: Bayesian Statistics in Python* (O'Reilly Media, 2013).

[87] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. *Bayesian Data Analysis* (Taylor & Francis, 2013).

[88] Lambert, B. *A Student's Guide to Bayesian Statistics* (SAGE, 2018).

[89] Murphy, K. P. *Machine Learning: A Probabilistic Perspective* (MIT Press, 2012).

[90] Robert, C. & Casella, G. A short history of Markov Chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science* **26**, 102–115 (2011).

[91] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**, 1087–1092 (1953).

[92] Geman, S. & Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6**, 721–741 (1984).

[93] Betancourt, M. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv: 1701.02434* (2017).

[94] Duane, S., Kennedy, A. D., Pendleton, B. J. & Roweth, D. Hybrid Monte Carlo. *Physics Letters B* **195**, 216–222 (1987).

[95] Neal, R. M. MCMC using Hamiltonian dynamics. *arXiv: 1206.1901* (2012).

[96] Monnahan, C. C., Thorson, J. T. & Branch, T. A. Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution* **8**, 339–348 (2017).

[97] Eastwood, J. W. & Hockney, R. W. *Computer Simulation Using Particles* (A. Hilger, 1988).

[98] Homan, M. D. & Gelman, A. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15**, 1593–1623 (2014).

[99] Andrieu, C. & Thoms, J. A tutorial on adaptive MCMC. *Statistics and Computing* **18**, 343–373 (2008).

[100] Nesterov, Y. Primal-dual subgradient methods for convex problems. *Mathematical Programming* **120**, 221–259 (2009).

[101] Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–472 (1992).

[102] PyMC3 – Plots. Disponível em: https://pymc3-testing.readthedocs.io/en/rtd-docs/api/plots.html. Último acesso em: 21 de agosto de 2023.

[103] Rousseeuw, P. J. & Croux, C. Alternatives to the median absolute deviation. *Journal of the American Statistical Association* **88**, 1273–1283 (1993).

[104] Huber, P. J. *Robust Statistics* (Wiley, 2004).

[105] Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* (Springer, 2003).

[106] Huber, P. J. Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35**, 73–101 (1964).

[107] Staudte, R. G. & Sheather, S. J. *Robust Estimation and Testing* (Wiley, 1990).

[108] Süli, E. & Mayers, D. F. *An Introduction to Numerical Analysis* (Cambridge University Press, 2003).

[109] McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv: 1802.03426* (2020).

[110] Ghojogh, B., Crowley, M., Karray, F. & Ghodsi, A. *Elements of dimensionality reduction and manifold learning* (Springer Nature, 2023).

[111] Lee, E. K., Balasubramanian, H., Tsolias, A., Anakwe, S. U., Medalla, M., Shenoy, K. V. & Chandrasekaran, C. Non-linear dimensionality reduction on extracellular waveforms reveals cell type diversity in premotor cortex. *Elife* **10**, e67490 (2021).

[112] Belkin, M. & Niyogi, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems* **14**, 1–7 (2001).

[113] Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* **105**, 1118–1123 (2008).

[114] Rosvall, M., Axelsson, D. & Bergstrom, C. T. The map equation. *The European Physical Journal Special Topics* **178**, 13–23 (2009).

[115] Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423 (1948).

[116] Brin, S. & Page, L. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems* **30**, 107–117 (1998).

[117] Rosvall, M. & Bergstrom, C. T. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one* **6**, e18209 (2011).

[118] Plataforma Lattes. Disponível em: http://lattes.cnpq.br. Último acesso em: 21 de agosto de 2023.

[119] Guerrero-Bote, V. P. & Moya-Anegón, F. A further step forward in measuring journals' scientific prestige: The SJR2 indicator. *Journal of Informetrics* **6**, 674–688 (2012).

[120] Bordons, M., Fernández, M. & Gómez, I. Advantages and limitations in the use of impact factor measures for the assessment of research performance. *Scientometrics* **53**, 195–206 (2002).

[121] Foster, J. G., Rzhetsky, A. & Evans, J. A. Tradition and innovation in scientists' research strategies. *American Sociological Review* **80**, 875–908 (2015).

[122] Antonoyiannakis, M. Impact factors and the Central Limit Theorem: Why citation averages are scale dependent. *Journal of Informetrics* **12**, 1072–1088 (2018).

[123] Antonoyiannakis, M. Impact factor volatility due to a single paper: A comprehensive analysis. *Quantitative Science Studies* **1**, 639–663 (2020).

[124] Gelman, A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 515–534 (2006).

[125] Dundar, H. & Lewis, D. R. Determinants of research productivity in higher education. *Research in Higher Education* **39**, 607–631 (1998).

[126] Petersen, A. M., Pan, R. K., Pammolli, F. & Fortunato, S. Methods to account for citation inflation in research evaluation. *Research Policy* **48**, 1855–1865 (2019).

[127] Björk, B.-C. & Solomon, D. The publishing delay in scholarly peer-reviewed journals. *Journal of Informetrics* **7**, 914–923 (2013).

[128] Virtanen, P. *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020).

[129] Sakoe, H. & Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **26**, 43–49 (1978).

[130] Meert, W., Hendrickx, K. & Van Craenendonck, T. dtaidistance. Disponível em: https://pypi.org/project/dtaidistance/. Último acesso em: 21 de novembro de 2023.

[131] McInnes, L., Healy, J., Saul, N. & Grossberger, L. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software* **3**, 861–862 (2018).

[132] Lancichinetti, A. & Fortunato, S. Community detection algorithms: A comparative analysis. *Physical Review E* **80**, 056117 (2009).

[133] Fortunato, S. Community detection in graphs. *Physics Reports* **486**, 75–174 (2010).

[134] Fortunato, S. & Hric, D. Community detection in networks: A user guide. *Physics Reports* **659**, 1–44 (2016).

[135] Infomap. Disponível em: https://www.mapequation.org/infomap/. Último acesso em: 21 de agosto de 2023.

[136] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008).

[137] Traag, V. A., Waltman, L. & Van Eck, N. J. From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports* **9**, 1–12 (2019).

[138] Hicks, D. Performance-based university research funding systems. *Research Policy* **41**, 251–261 (2012).

[139] Price, M. Some scientists publish more than 70 papers a year. Here's how – and why – they do it. Disponível em: https://doi.org/10.1126/science.aav4004 (2018). Último acesso em: 21 de agosto de 2023.

[140] Clauset, A., Arbesman, S. & Larremore, D. B. Systematic inequality and hierarchy in faculty hiring networks. *Science Advances* **1**, e1400005 (2015).

[141] Stephan, P. *How Economics Shapes Science* (Harvard University Press, 2015).

[142] Bourdieu, P. The specificity of the scientific field and the social conditions of the progress of reason. *Social Science Information* **14**, 19–47 (1975).

[143] Bourdieu, P. The peculiar history of scientific reason. *Sociological Forum* **6**, 3–26 (1991).

[144] Bourdieu, P. *Science of Science and Reflexivity* (Polity Press, 2004).

[145] Morgan, A. C., Way, S. F., Hoefer, M. J., Larremore, D. B., Galesic, M. & Clauset, A. The unequal impact of parenthood in academia. *Science Advances* **7**, eabd1996 (2021).

[146] Zeng, X. H. T., Duch, J., Sales-Pardo, M., Moreira, J. A., Radicchi, F., Ribeiro, H. V., Woodruff, T. K. & Amaral, L. A. N. Differences in collaboration patterns across discipline, career stage, and gender. *PLoS Biology* **14**, e1002573 (2016).

[147] Araújo, E. B., Araújo, N. A., Moreira, A. A., Herrmann, H. J. & Andrade Jr, J. S. Gender differences in scientific collaborations: Women are more egalitarian than men. *PloS one* **12**, e0176791 (2017).

[148] Duch, J., Zeng, X. H. T., Sales-Pardo, M., Radicchi, F., Otis, S., Woodruff, T. K. & Nunes Amaral, L. A. The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PloS one* **7**, e51332 (2012).

[149] Huang, J., Gates, A. J., Sinatra, R. & Barabási, A.-L. Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences* **117**, 4609–4616 (2020).

[150] Spoon, K., LaBerge, N., Wapman, K. H., Zhang, S., Morgan, A. C., Galesic, M., Fosdick, B. K., Larremore, D. B. & Clauset, A. Gender and retention patterns among us faculty. *Science Advances* **9**, eadi2205 (2023).

[151] AlShebli, B. K., Rahwan, T. & Woon, W. L. The preeminence of ethnic diversity in scientific collaboration. *Nature communications* **9**, 5163 (2018).

[152] Peng, H., Lakhani, K. & Teplitskiy, M. Acceptance in top journals shows large disparities across name-inferred ethnicities. *SocArXiv* (2021).

[153] Arnett, J. J. The neglected 95%: Why American psychology needs to become less American. *American Psychologist* **63**, 602–614 (2016).

APÊNDICE A

Material suplementar

Neste apêndice, apresentamos todo material suplementar mencionado no texto principal.

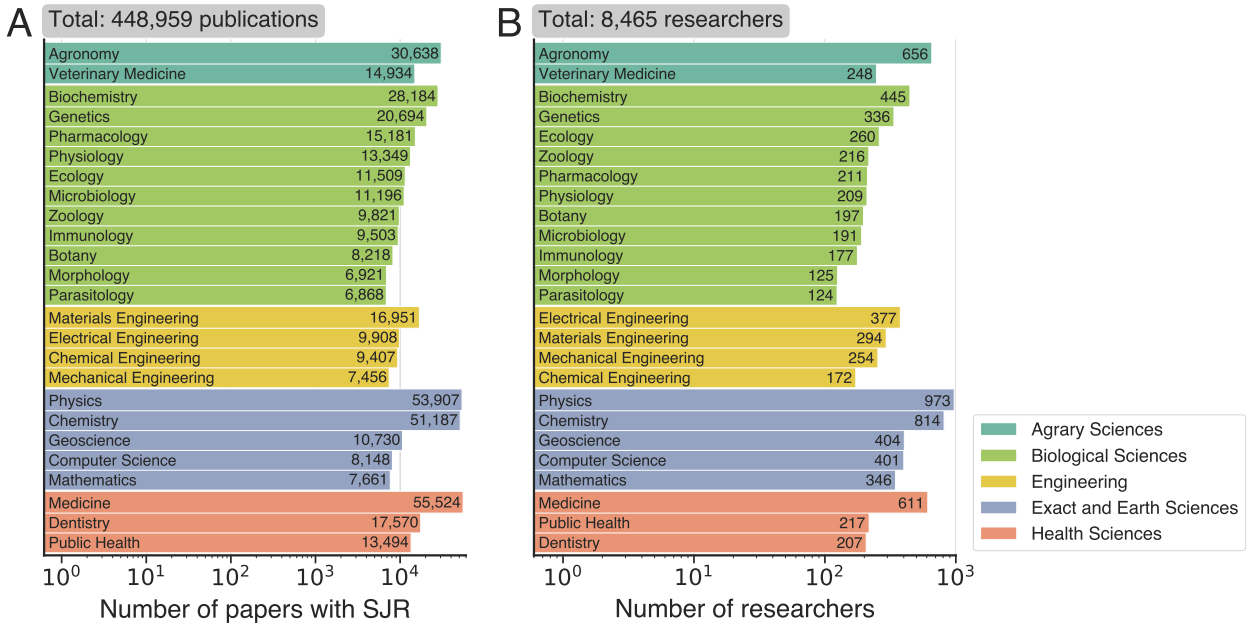

Figura A.1: Número de publicações e pesquisadores no conjunto de dados SJR. O painel (A) mostra o número total de artigos e o painel (B) mostra o número total de pesquisadores para cada disciplina no conjunto de dados SJR. As cores das barras representam os diferentes campos da ciência em nosso conjunto de dados.

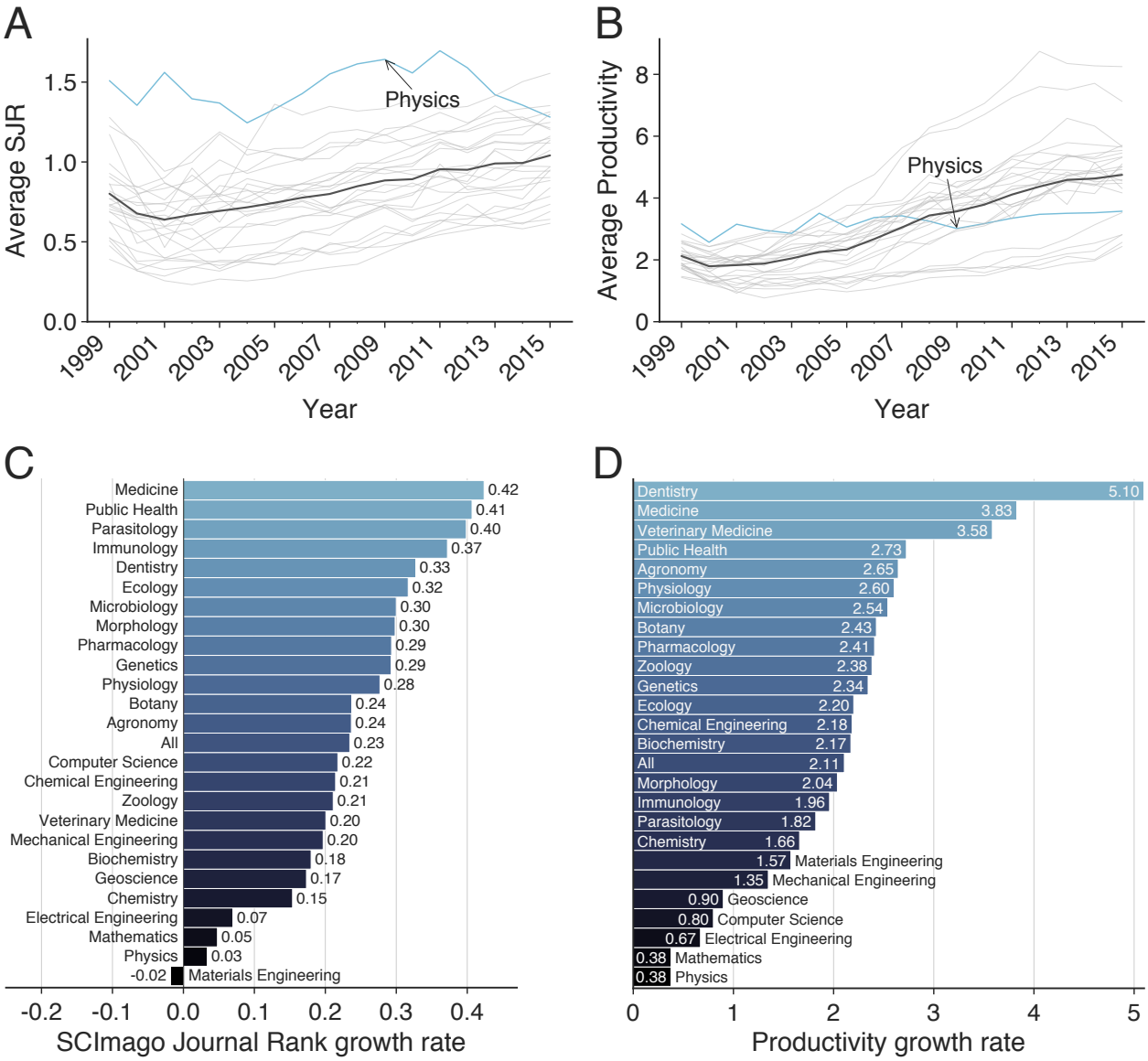

Figura A.2: Evolução temporal do prestígio médio de jornal e da produtividade. Os painéis (A) e (B) mostram a evolução temporal dos valores médios do prestígio médio de jornal e da produtividade, respectivamente, para o conjunto de dados SJR. As curvas em cinza mostram o comportamento médio das disciplinas separadamente, as curvas em preto representam o comportamento médio agregado de todas as disciplinas e as curvas em azul ilustram o comportamento médio da disciplina de física. Os valores médios foram estimados utilizando o estimador de localização Huber. Os painéis (C) e (D) mostram as taxas de crescimento por década do prestígio médio de jornal e da produtividade, respectivamente, estimadas a partir do conjunto de dados SJR. Estimamos as taxas de crescimento ajustando um modelo linear à evolução temporal reportada nos painéis (A) e (B) para cada disciplina. Além disso, estimamos a taxa de crescimento agregando os dados de todas as disciplinas (indicado por *All* nos gráficos de barra).

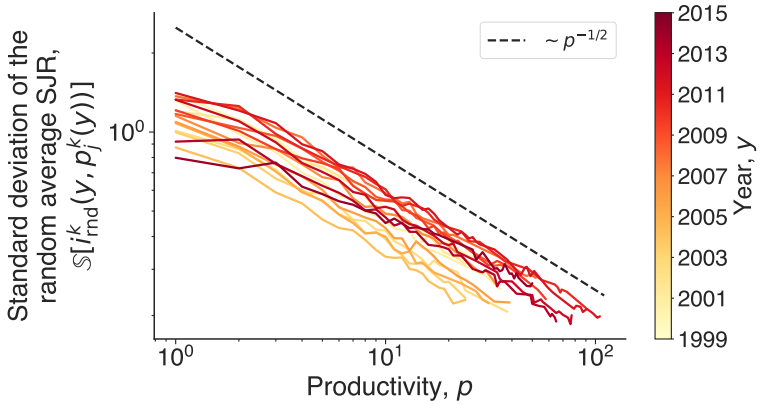

Figura A.3: Efeito do tamanho da produtividade na dispersão do prestígio médio de jornal. Desvio padrão $\mathbb{S}[i^k_{\mathrm{rnd}}(y, p^k_j(y))]$ do valor médio do ranque de jornais SCImago (SJR) para 1000 amostras aleatórias de p publicações de pesquisadores da física como uma função de p em todos os anos disponíveis no conjunto de dados SJR. O código de cor refere-se a cada ano do conjunto de dados e a linha tracejada representa o comportamento esperado pelo Teorema Central do Limite.

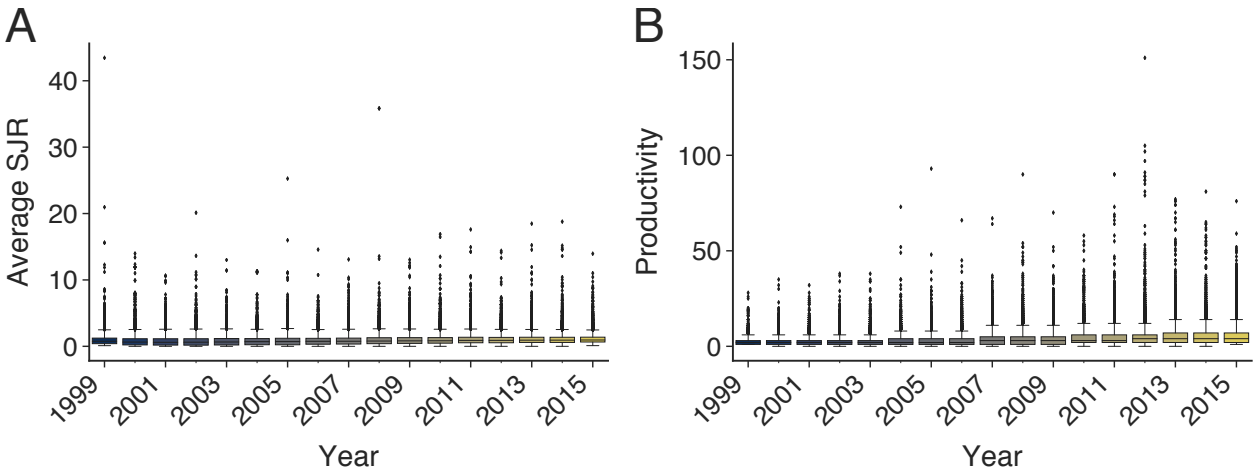

Figura A.4: Valores *outliers* do prestígio médio de jornal e da produtividade. Os diagramas de caixa retratam o grau de dispersão do (A) prestígio médio de jornal (SJR) e da (B) produtividade dos pesquisadores no conjunto de dados SJR em cada ano. Existem observações extremas em todos os anos, que estão representados por marcadores pretos além dos bigodes (aqui definidos como 1.5 vezes o intervalo interquartil).

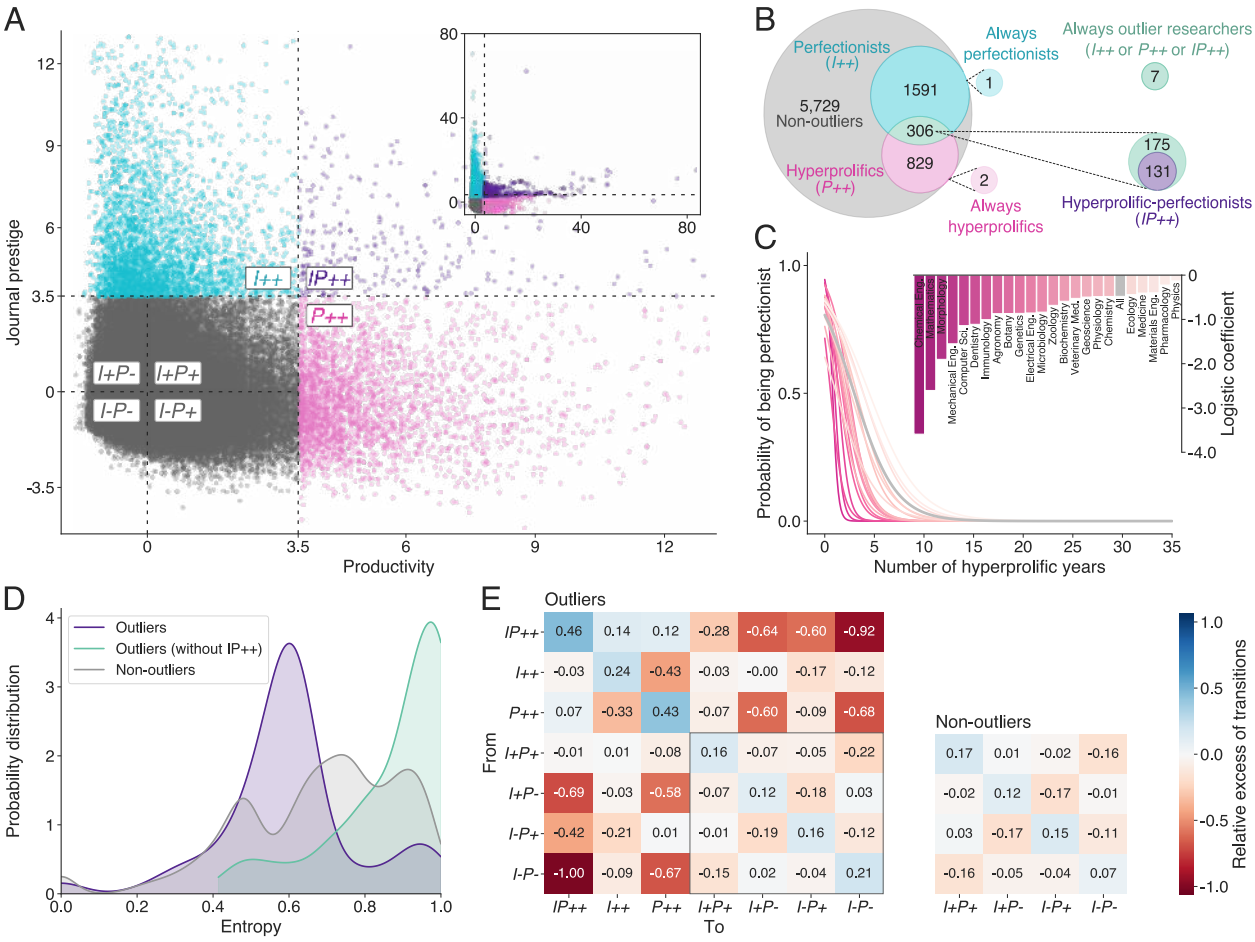

Figura A.5: Prestígio de jornal *versus* produtividade considerando o conjunto de dados SJR. (A) Plano prestígio de jornal *versus* produtividade em unidades padronizadas. A inserção mostra o intervalo completo do plano. Os marcadores representam anos de carreira de pesquisadores de 25 disciplinas em nosso estudo. (B) Diagrama de Venn mostrando o conjunto de relações entre as quatro categorias de pesquisadores (não *outliers*, perfeccionistas, hiperprolíficos e simultaneamente perfeccionistas e hiperprolíficos). (C) Probabilidade de ser um pesquisador perfeccionista tendo um determinado número de anos da carreira no setor hiperprolífico ($P{+}{+}$) estimada via regressão logística (veja a seção 1.1). A inserção mostra os coeficientes logísticos. As curvas e barras coloridas referem-se a diferentes disciplinas, enquanto a curva e a barra em cinza representam o resultado ao agregar todas as disciplinas. As disciplinas de parasitologia e saúde pública (omitidas nesse painel) são as únicas disciplinas que não apresentam uma associação significativa. (D) Distribuição de probabilidade da entropia normalizada de Shannon associada à ocupação dos setores do plano para as carreiras individuais dos pesquisadores. A curva em roxo mostra a entropia associada à ocupação de setores *outliers* por pesquisadores *outliers*, enquanto a curva em verde representa o mesmo mas ignorando o setor $IP{+}{+}$. A curva em cinza mostra a distribuição da entropia para pesquisadores não *outliers*. (E) Matriz de transição entre setores do plano prestígio de jornal *versus* produtividade para pesquisadores *outliers* (esquerda) e não *outliers* (direita). Cada célula representa o excesso relativo de transições entre dois setores comparado com o modelo nulo. O modelo nulo fornece os valores médios de excesso para versões embaralhadas das carreiras dos pesquisadores considerando 10 000 realizações.

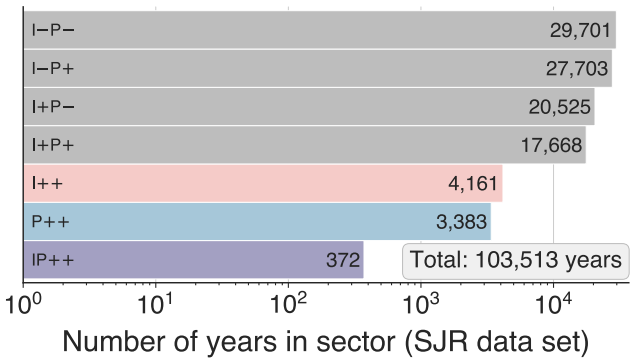

Figura A.6: Demografia do plano prestígio de jornal *versus* produtividade para o conjunto de dados SJR. As barras mostram o número de anos de carreira em cada setor do plano prestígio de jornal *versus* produtividade.

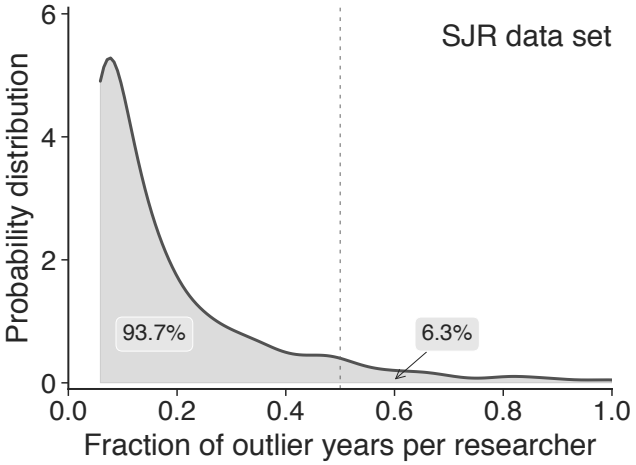

Figura A.7: Distribuição de probabilidade da fração de anos *outliers* na carreira de pesquisadores para o conjunto de dados SJR.

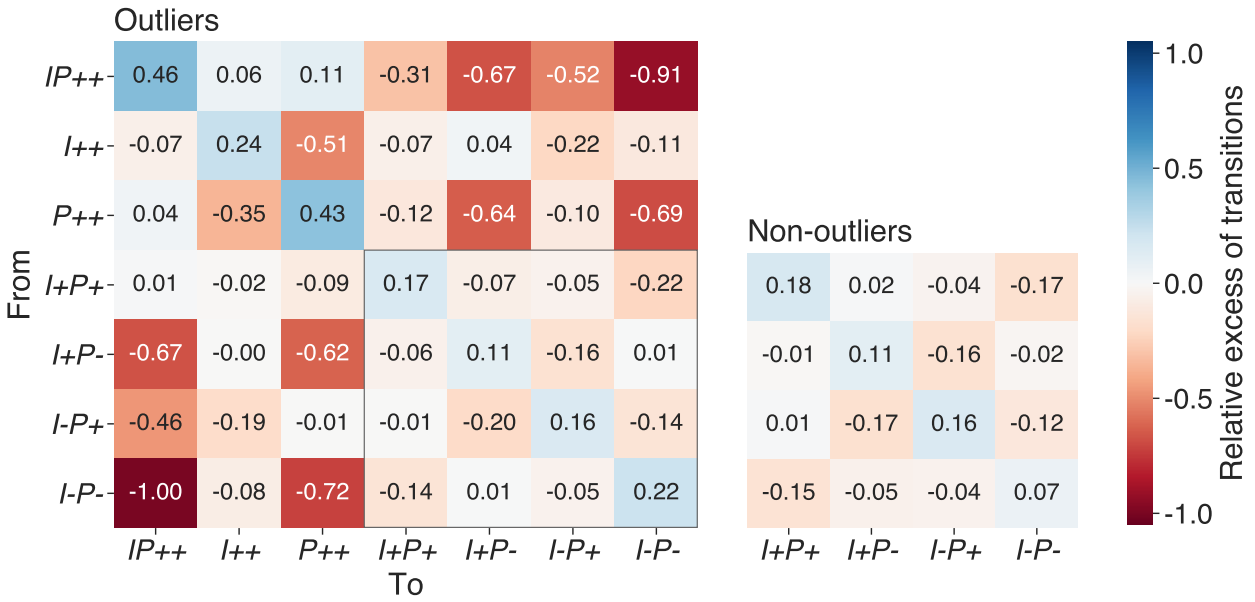

Figura A.8: Matriz de transição entre setores do plano prestígio de jornal *versus* produtividade para o conjunto de dados SJR considerando apenas o conjunto de disciplinas presentes no conjunto de dados JIF. Cada célula representa o excesso relativo de transições entre dois setores comparado com o modelo nulo. O modelo nulo fornece os valores médios de excesso para versões embaralhadas das carreiras dos pesquisadores considerando 10 000 realizações. Notamos que os padrões de transições mostrados nesta figura são muito similares àquelas reportadas na Figura A.5E.

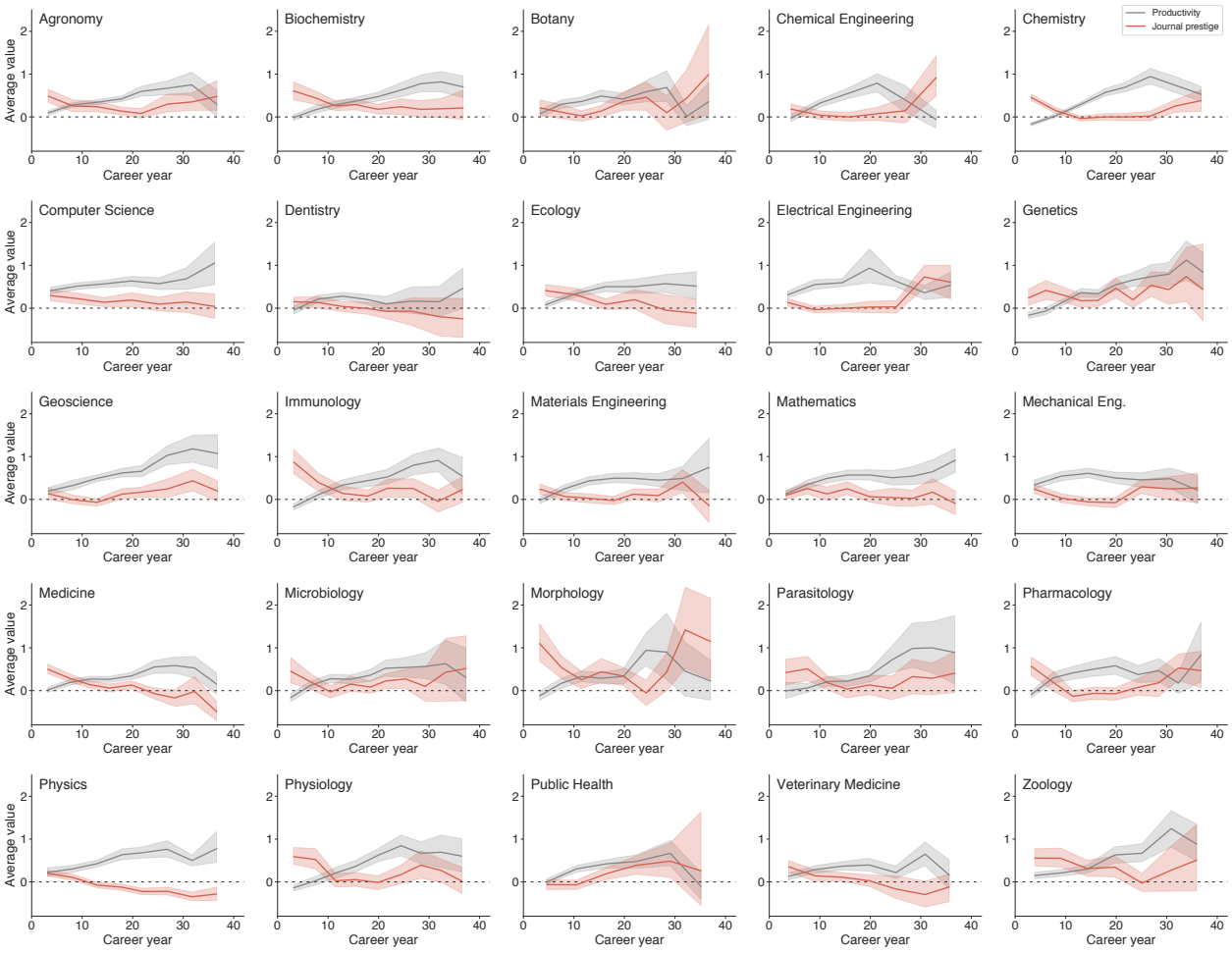

Figura A.9: Valores médios da produtividade e do impacto de jornal ao longo da carreira dos pesquisadores para diferentes disciplinas considerando o conjunto de dados SJR. Essas visualizações mostram os valores médios da produtividade (curva em cinza) e do prestígio de jornal (curva em vermelho) calculados a partir de médias móveis de 5 anos ao longo dos anos da carreira para cada disciplina do conjunto de dados SJR. As regiões sombreadas correspondem a intervalos de confiança de 95% obtidos pelo método de *bootstrap*.

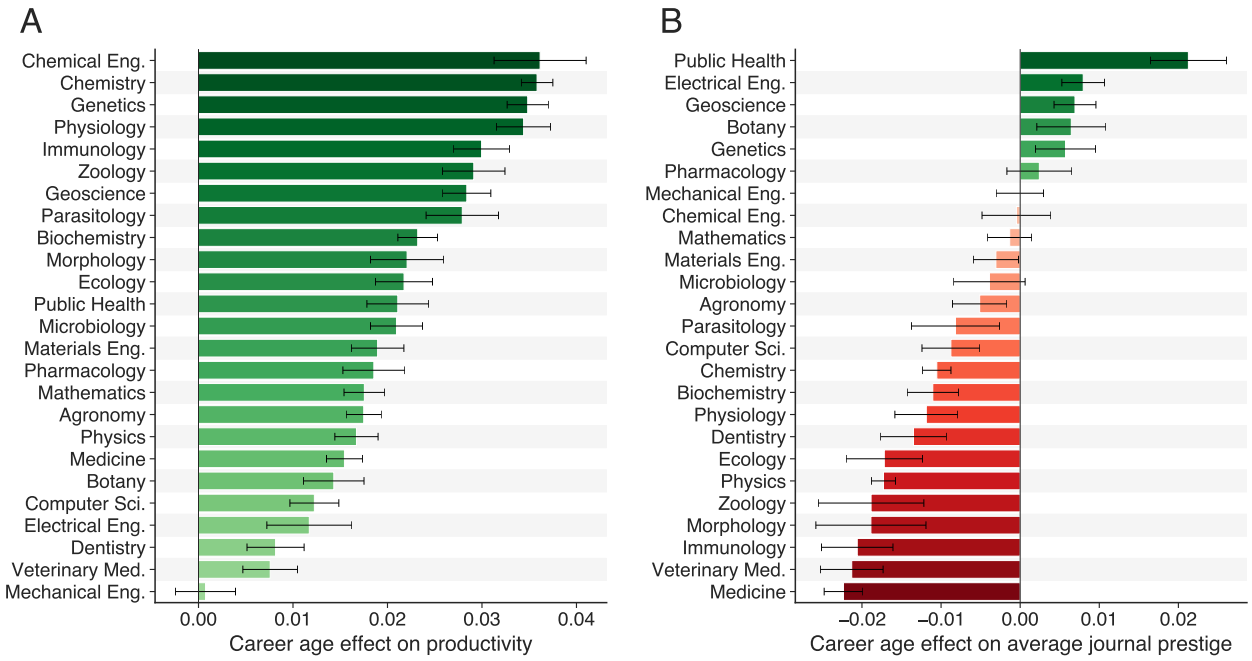

Figura A.10: Efeito do ano da carreira na produtividade e no prestígio de jornal para diferentes disciplinas considerando o conjunto de dados SJR. Os gráficos de barra mostram o efeito do ano da carreira na (A) produtividade e no (B) prestígio de jornal para cada disciplina no conjunto de dados SJR. Estimamos os valores por meio de um modelo linear da associação média entre idade da carreira e produtividade entre idade da carreira e prestígio de jornal (Figura A.9) para cada disciplina. As barras de erro indicam o erro padrão dos coeficientes lineares.

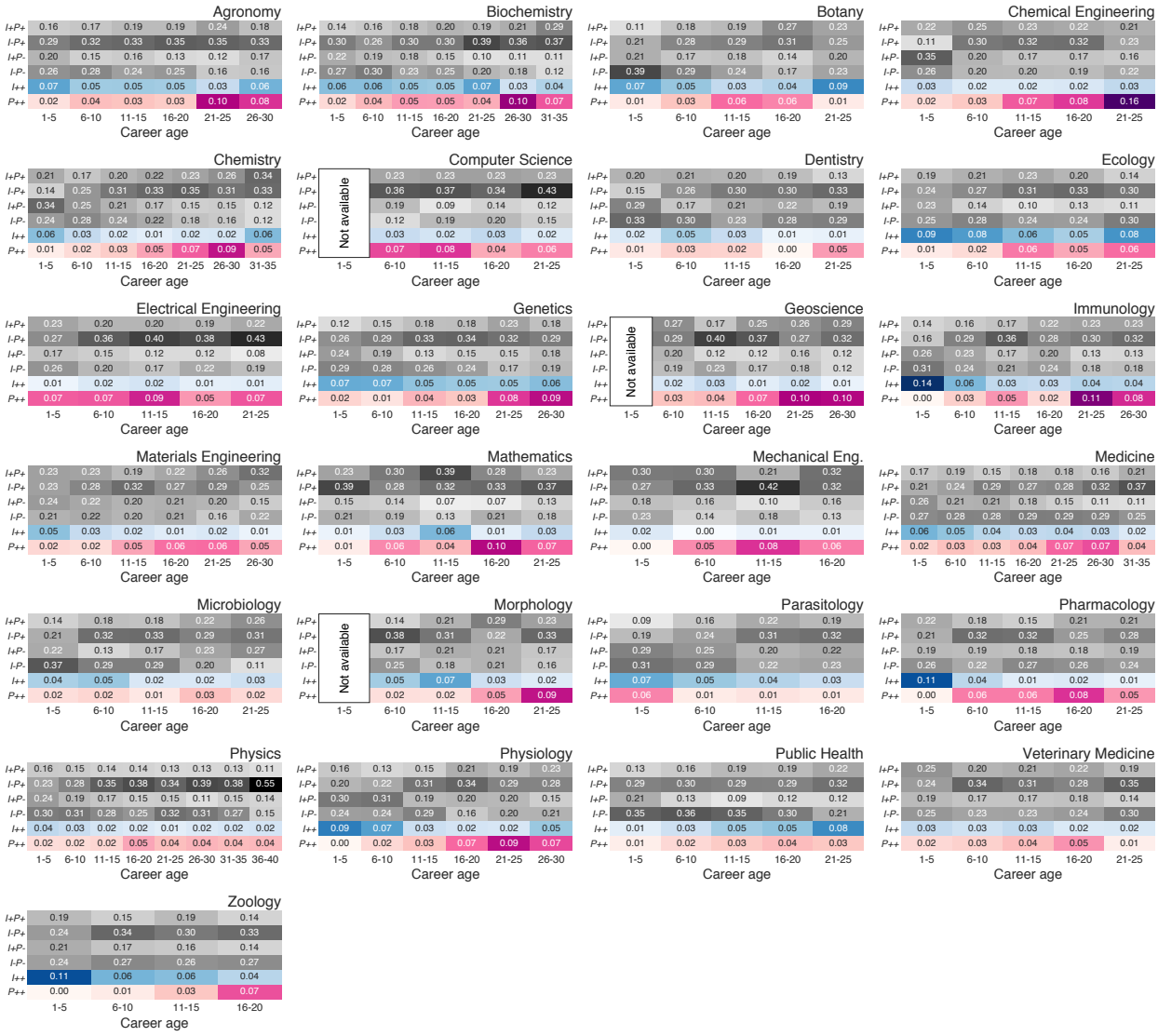

Figura A.11: Tendências de ocupação do plano prestígio de jornal *versus* produtividade ao longo das carreiras dos pesquisadores considerando o conjunto de dados SJR. Os painéis mostram a fração dos anos das carreiras em cada setor não *outlier* e nos setores *outliers* $I{+}{+}$ e $P{+}{+}$ como uma função do ano da carreira dos pesquisadores de 25 disciplinas no conjunto de dados SJR. As colunas indicam intervalos de 5 anos e as linhas representam os diferentes setores. O código de cor indica as frações para setores não *outliers* (tons de cinza) e setores *outliers* para os setores $I{+}{+}$ (tons de azul) e $P{+}{+}$ (tons de rosa). O setor $IP{+}{+}$ foi omitido uma vez que anos de carreira nesse setor são muito raros. Apenas intervalos de 5 anos com pelo menos 20 pesquisadores são mostrados nessas visualizações.

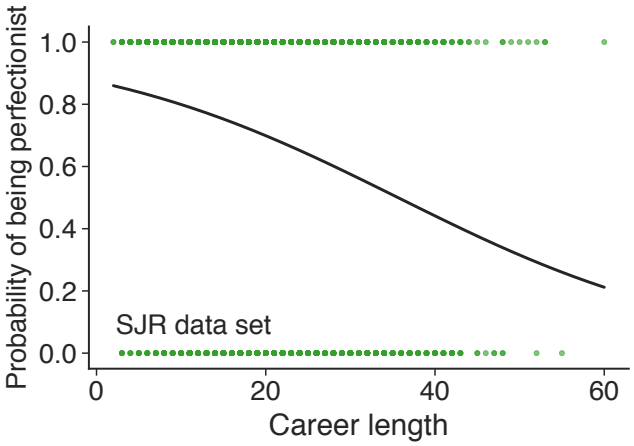

Figura A.12: Efeito do comprimento da carreira na probabilidade de ser perfeccionista para o conjunto de dados SJR. Estimamos a probabilidade de ser perfeccionista como uma função do comprimento da carreira do pesquisador via modelo logístico (veja a seção 1.1).

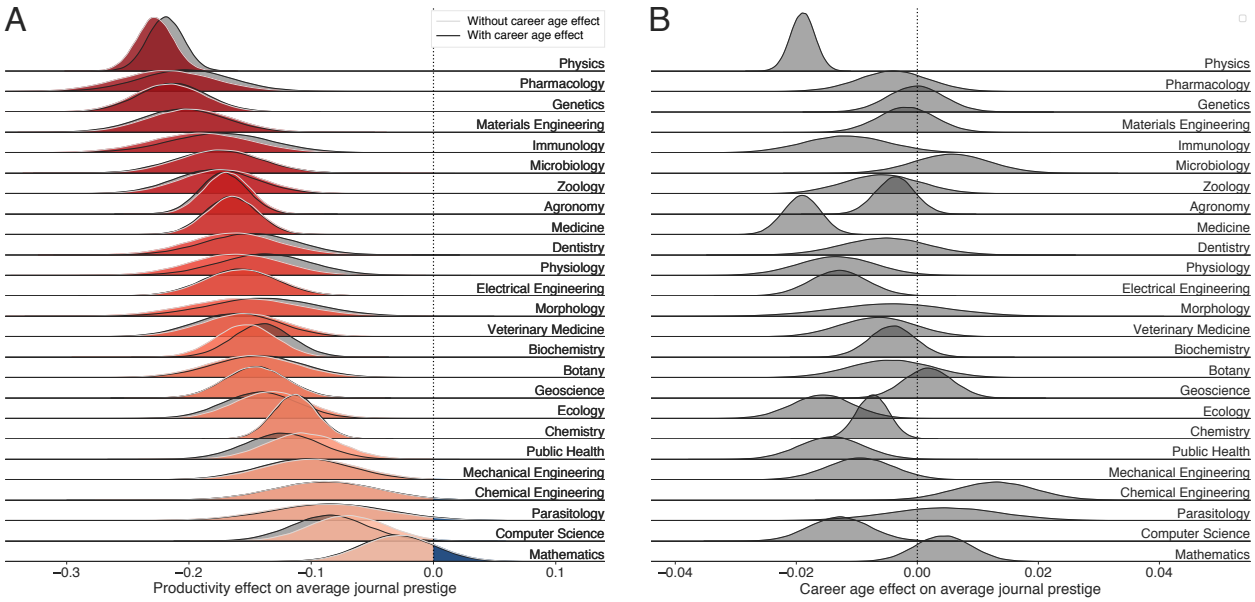

Figura A.13: Efeito da produtividade no prestígio de jornal para pesquisadores não *outliers* considerando o conjunto de dados SJR. (A) Distribuições de probabilidade a *posteriori* do valor médio do coeficiente linear (μ_P) ao considerar a associação entre produtividade e impacto de jornal para pesquisadores não *outliers* de cada disciplina. As curvas coloridas preenchidas representam os resultados sem levar em consideração os efeitos do ano da carreira, enquanto as curvas preenchidas em cinza mostram as distribuições de μ_P após incluir o ano da carreira como fator de confusão no modelo bayesiano hierárquico. (B) Distribuições de probabilidade a *posteriori* do valor médio do coeficiente linear (μ_A) relacionado ao efeito do ano da carreira no impacto de jornal para pesquisadores não *outliers* de cada disciplina.

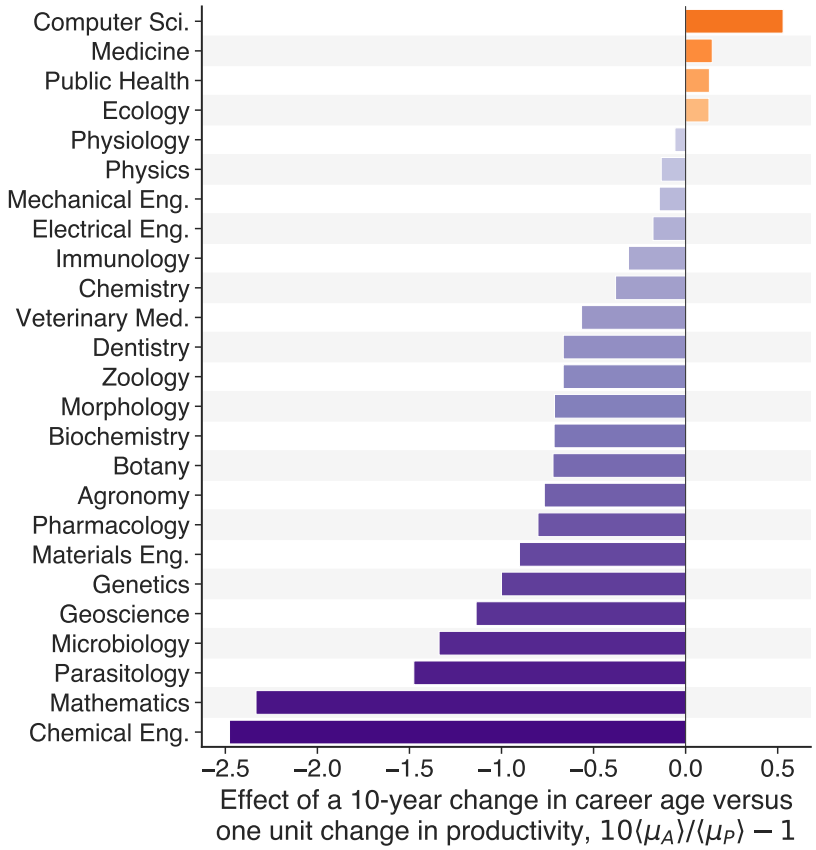

Figura A.14: Comparação entre os efeitos do ano da carreira e produtividade no prestígio de jornal considerando o conjunto de dados SJR. As barras comparam o efeito de uma progressão de 10 anos na carreira com o efeito de aumentar uma unidade da produtividade (z-score) para um pesquisador típico de cada disciplina no conjunto de dados SJR. Esses valores representam a fração de quão maior ou menor é o efeito do ano da carreira comparado com o efeito da produtividade (isto é, $10\langle\mu_A\rangle/\langle\mu_P\rangle - 1$, em que $\langle\mu_A\rangle$ e $\langle\mu_P\rangle$ são os valores médios, respectivamente, de μ_A e μ_P para cada disciplina). Frações ao redor de zero indicam que um aumento de 10 anos na idade da carreira afeta o impacto de jornal de maneira similar ao aumento de uma unidade na produtividade. Valores positivos indicam que uma mudança de 10 anos na idade da carreira afeta mais o impacto de jornal do que o aumento de uma unidade de produtividade, enquanto valores negativos indicam que produtividade tem maior impacto no prestígio de jornal.

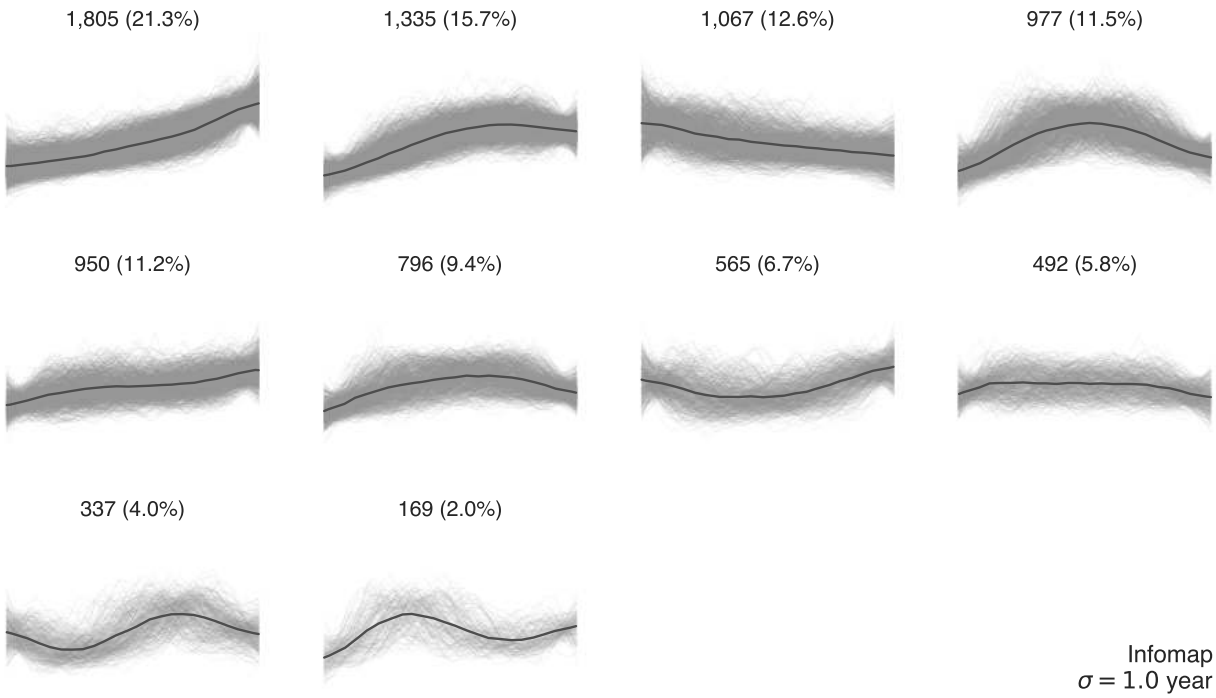

Figura A.15: Agrupamento das curvas de produtividade usando nosso procedimento com séries temporais suavizadas e um filtro gaussiano com desvio padrão de $\sigma = 1.0$ ano. Esse valor é menor do que o valor escolhido para os resultados mostrados no texto principal ($\sigma = 2.0$ anos). Os painéis mostram as curvas de produtividade em cada comunidade identificada. As curvas pretas representam o comportamento médio de cada grupo. Os comprimentos das carreiras de cada grupo foram reescaladas para o intervalo unitário e as frações de pesquisadores em cada grupo são mostradas em cada painel. Os padrões de agrupamento obtidos usando $\sigma = 1.0$ ano são similares aos padrões obtidos para $\sigma \in \{1.5, 2.0, 2.5\}$ anos.

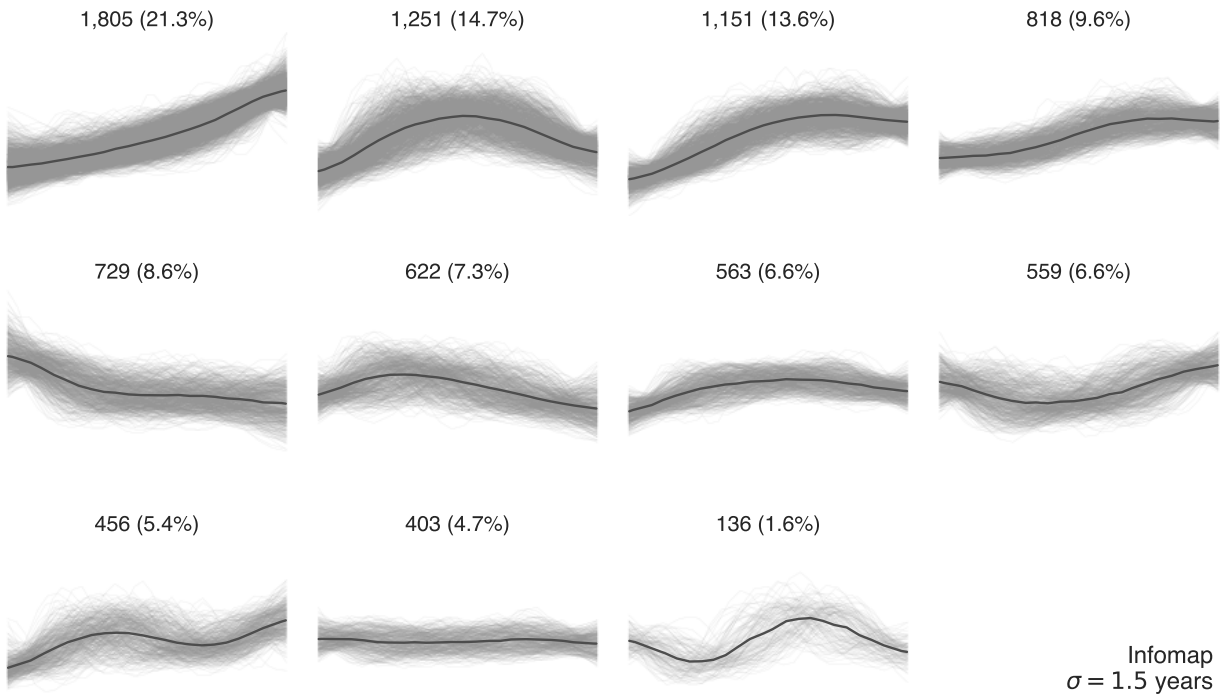

Figura A.16: Agrupamento das curvas de produtividade usando nosso procedimento com séries temporais suavizadas e um filtro gaussiano com desvio padrão de $\sigma = 1.5$ anos. Esse valor é menor do que o valor escolhido para os resultados mostrados no texto principal ($\sigma = 2.0$ anos). Os painéis mostram as curvas de produtividade em cada comunidade identificada. As curvas pretas representam o comportamento médio de cada grupo. Os comprimentos das carreiras de cada grupo foram reescaladas para o intervalo unitário e as frações de pesquisadores em cada grupo são mostradas em cada painel. Os padrões de agrupamento obtidos usando $\sigma = 1.5$ anos são similares aos padrões obtidos para $\sigma \in \{1.0, 2.0, 2.5\}$ anos.

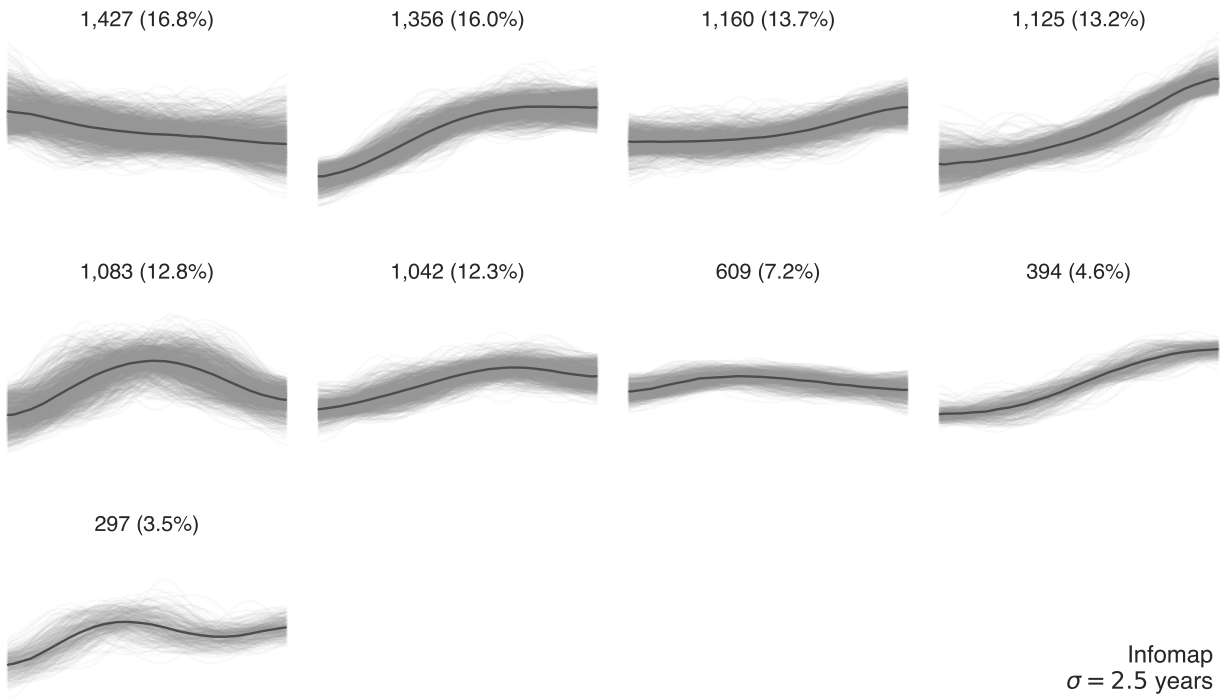

Figura A.17: Agrupamento das curvas de produtividade usando nosso procedimento com séries temporais suavizadas e um filtro gaussiano com desvio padrão de $\sigma = 2.5$ anos. Esse valor é maior do que o valor escolhido para os resultados mostrados no texto principal ($\sigma = 2.0$ anos). Os painéis mostram as curvas de produtividade em cada comunidade identificada. As curvas pretas representam o comportamento médio de cada grupo. Os comprimentos das carreiras de cada grupo foram reescaladas para o intervalo unitário e as frações de pesquisadores em cada grupo são mostradas em cada painel. Os padrões de agrupamento obtidos usando $\sigma = 2.5$ anos são similares aos padrões obtidos para $\sigma \in \{1.0, 1.5, 2.0\}$ anos.

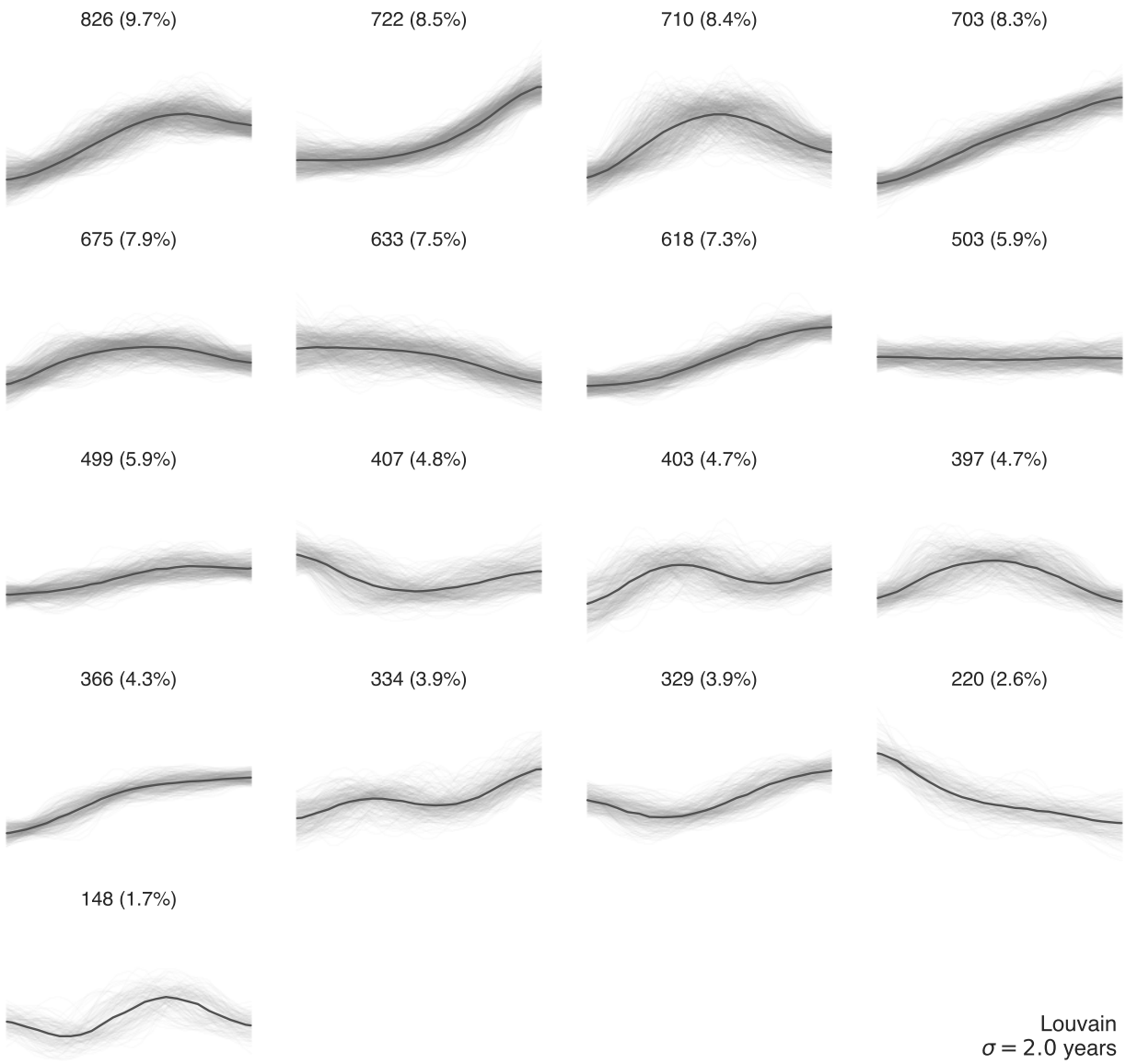

Figura A.18: Agrupamento das curvas de produtividade usando o algoritmo de detecção de comunidades Louvain. Os painéis mostram as curvas de produtividade em cada comunidade identificada. As curvas pretas representam o comportamento médio de cada grupo. Os comprimentos das carreiras de cada grupo foram reescaladas para o intervalo unitário e as frações de pesquisadores em cada grupo são mostradas em cada painel. Os padrões de agrupamento são similares aos obtidos pelo algoritmo Infomap.

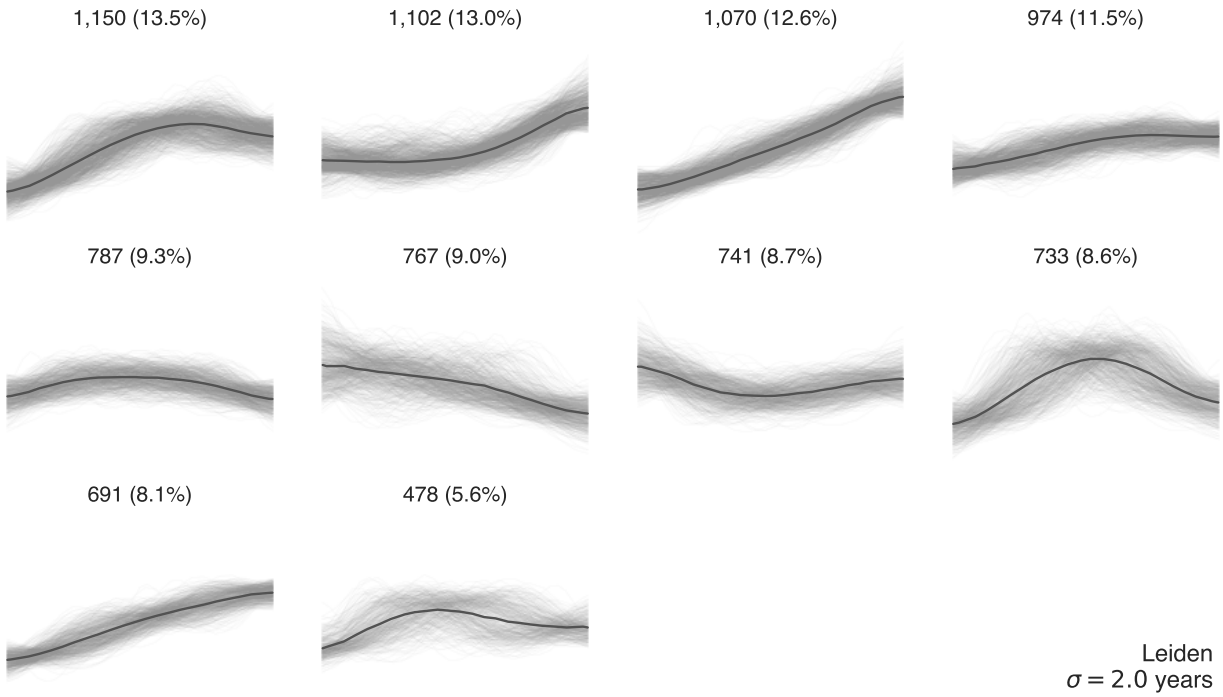

Figura A.19: Agrupamento das curvas de produtividade usando o algoritmo de detecção de comunidades Leiden. Os painéis mostram as curvas de produtividade em cada comunidade identificada. As curvas pretas representam o comportamento médio de cada grupo. Os comprimentos das carreiras de cada grupo foram reescaladas para o intervalo unitário e as frações de pesquisadores em cada grupo são mostradas em cada painel. Os padrões de agrupamento são similares aos obtidos pelo algoritmo Infomap.

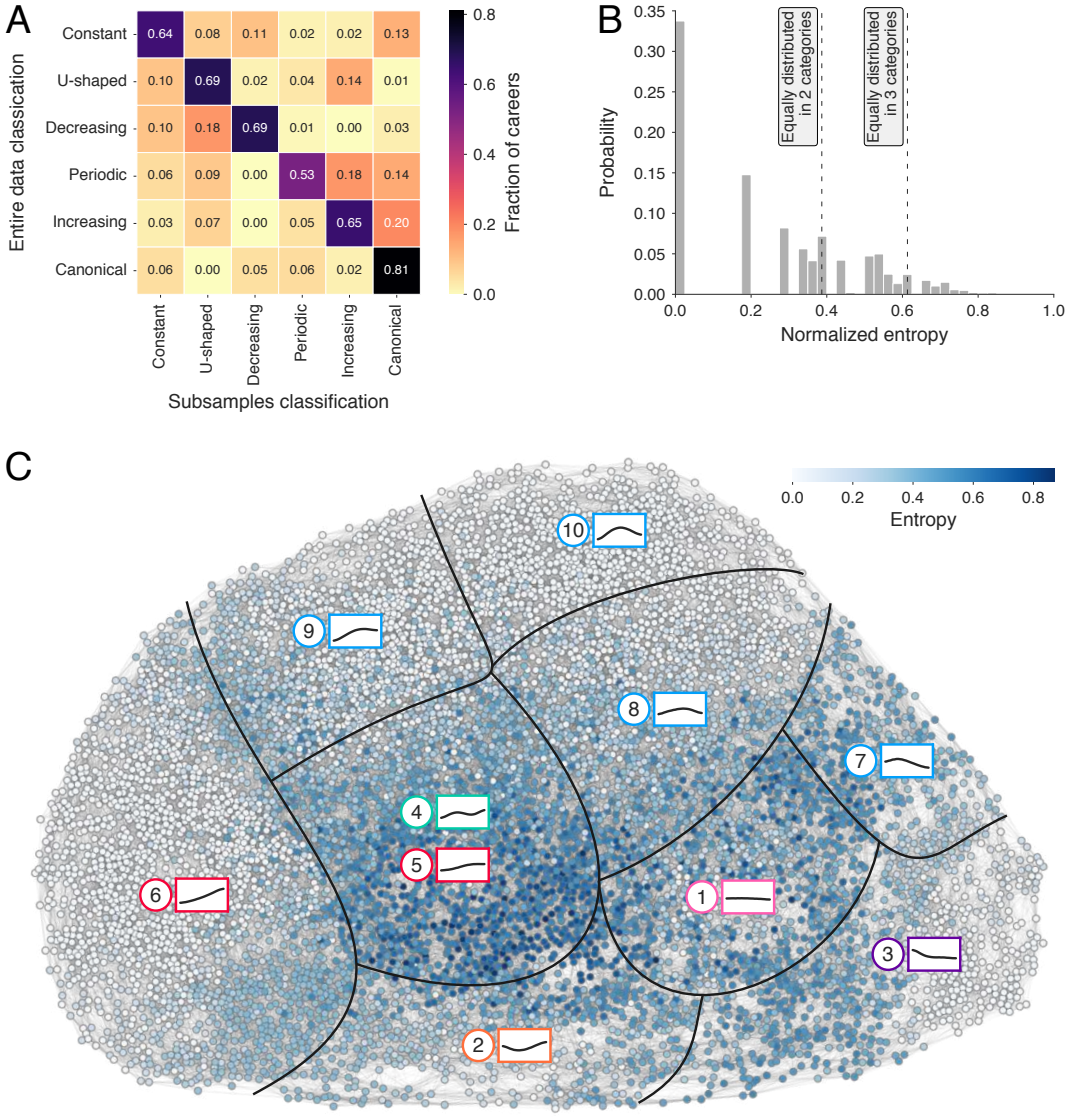

Figura A.20: Validação da robustez das seis categorias de padrões de produtividade após amostrar os dados aleatoriamente em duas partes iguais. (A) Matriz de confusão média associada à classificação dos dados completos (linhas) e à classificação por amostragem (colunas) calculada usando vinte amostras. (B) Histograma da entropia normalizada associada com as probabilidades de pertencimento às seis categorias de trajetória de produtividade. (C) Representação em rede em que vértices representam pesquisadores e arestas pesadas conectam pesquisadores com curvas de produtividade similares. As linhas em preto delimitam aproximadamente os dez grupos de curvas de produtividade (indicados no painel por seus números e padrões), enquanto os tons de azul correspondem aos valores normalizados de entropia.

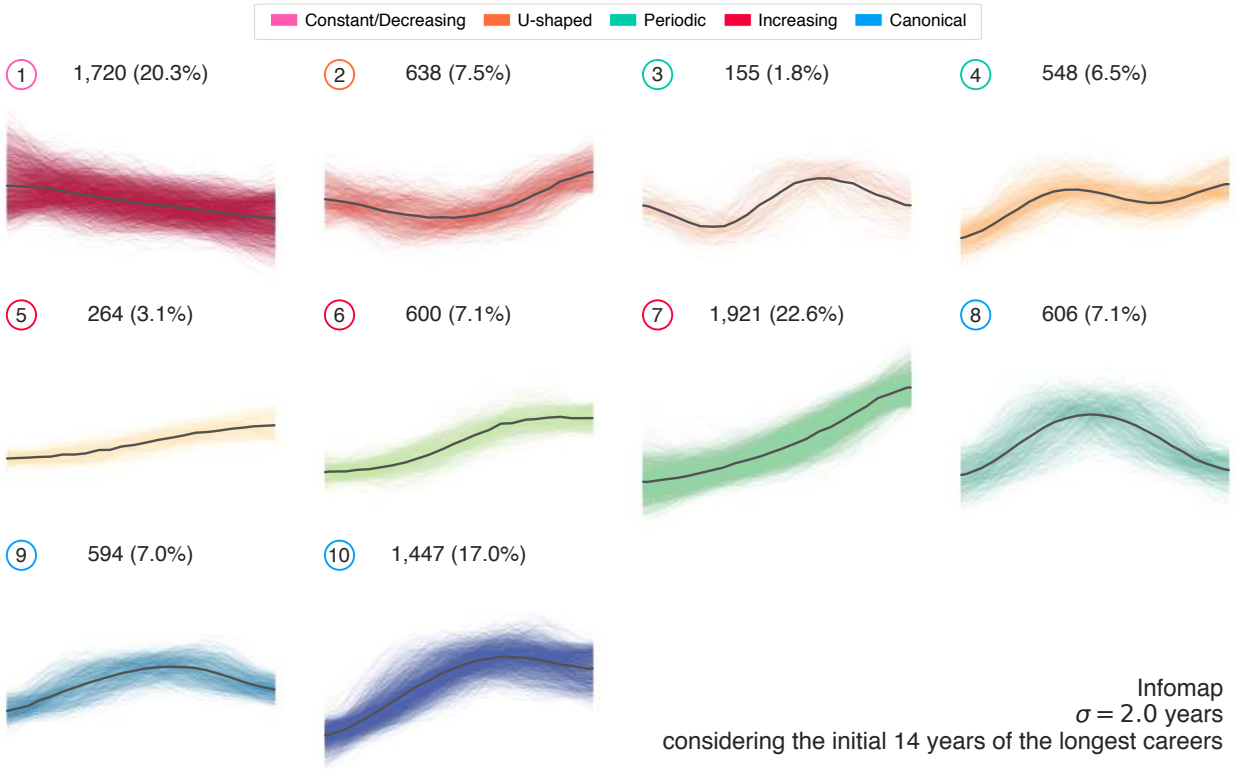

Figura A.21: Agrupamento das curvas de produtividade aplicando nosso procedimento ao conjunto de dados completo, mas considerando apenas os primeiros 14 anos das carreiras de pesquisadores com carreiras mais longas do que 24 anos. Os painéis mostram as curvas de produtividade em cada comunidade identificada. As curvas pretas representam o comportamento médio de cada grupo. Os comprimentos das carreiras de cada grupo foram reescaladas para o intervalo unitário e as frações de pesquisadores em cada grupo são mostradas em cada painel. Analisando as derivadas de cada um dos grupos por meio da Figura A.22, os dez grupos são ainda classificados em seis categorias: constante/decrescente (grupo 1), em forma de U (grupo 2), periódica (grupos 3 e 4), crescente (grupos 5 a 7) e com aspecto canônico (grupos 8 a 10). Os padrões de agrupamento são similares aos obtidos considerando as carreiras completas dos pesquisadores seniores. Apenas os padrões constante e decrescente (grupos 1 e 2 da Figura 3.9 do texto principal) foram combinados num único grupo (grupo 1) e as curvas periódicas (grupo 4 da Figura 3.9 do texto principal) emergiram como dois padrões distintos (grupos 3 e 4).

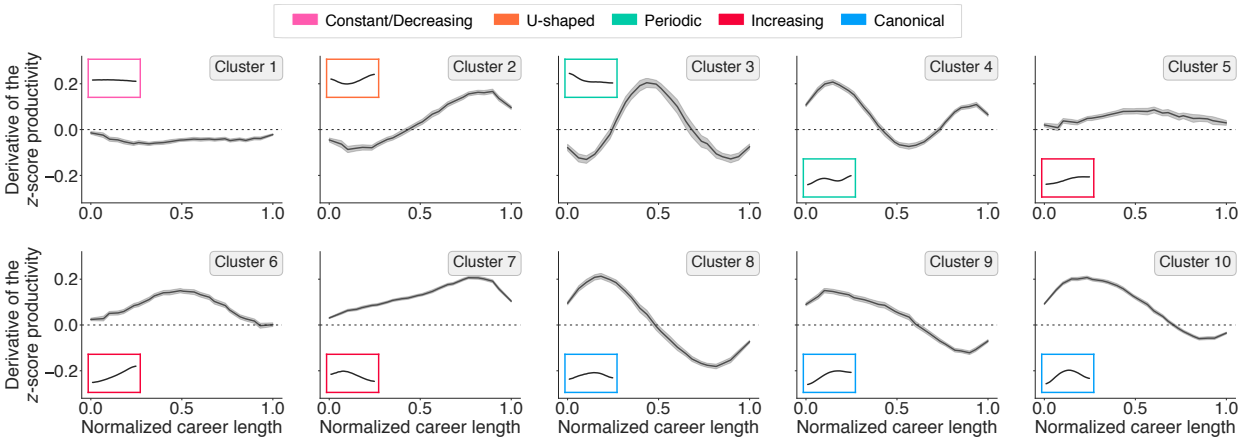

Figura A.22: Derivadas das trajetórias de produtividade de cada grupo ao considerar apenas os primeiros 14 anos das carreiras de pesquisadores com carreiras mais longas do que 24 anos (Figura A.21). As curvas em cada painel mostram a média móvel das trajetórias de produtividade diferenciadas, com as regiões sombreadas representando os intervalos de confiança de 95%. Os comprimentos das carreiras dos pesquisadores em cada grupo foram reescalados para o intervalo unitário antes da estimativa das médias. Valores positivos indicam taxas crescentes de produtividade, valores negativos indicam taxas decrescentes de produtividade e valores próximos de zero indicam produtividade constante em anos consecutivos da carreira. Com base nessas curvas e nos padrões médios de produtividade em cada grupo, definimos seis categorias de narrativas de produtividade: constante/decrescente (grupo 1), em forma de U (grupo 2), periódica (grupos 3 e 4), crescente (grupos 5 a 7) e com aspecto canônico (grupos 8 a 10).

Tabela A.1: Descrição do conjunto de dados SJR usado na análise bayesiana hierárquica. Número de pesquisadores e de observações para cada disciplina no conjunto de dados SJR após filtrar pesquisadores com carreiras mais curtas do que cinco anos.

Disciplina	Número de pesquisadores	Número de observações
Agronomia	408	4 391
Bioquímica	239	3 123
Botânica	124	1 359
Ciência da Computação	230	2 036
Ecologia	160	1 821
Engenharia Elétrica	239	2 297
Engenharia Mecânica	187	1 921
Engenharia Química	124	1 536
Engenharia dos Materiais	204	2 535
Farmacologia	142	1 878
Fisiologia	133	1 672
Física	670	8 474
Genética	188	2 409
Geociências	273	2 725
Imunologia	102	1 299
Matemática	215	2 147
Medicina	361	4 983
Medicina Veterinária	178	2 138
Microbiologia	131	1 698
Morfologia	71	956
Odontologia	151	1 937
Parasitologia	72	956
Química	566	7 314
Saúde Coletiva	144	1 734
Zoologia	126	1 418