UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE FÍSICA

Andre Seiji Sunahara

Impactos do tamanho em sistemas complexos

Maringá, junho de 2022.

UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE FÍSICA

Andre Seiji Sunahara

Impactos do tamanho em sistemas complexos

Monografia para qualificação de doutorado apresentada ao Programa de Pós-Graduação em Física da Universidade Estadual de Maringá como requisito parcial para obtenção do título de Doutor em Física.

Orientador: Prof. Dr. Haroldo Valentin Ribeiro

Maringá, 23 de setembro de 2022

Resumo

Neste trabalho, estudamos sistemas complexos empregando ferramentas da física estatística e da ciência de dados. Em particular, restringimos nossa análise a dois temas distintos. Primeiramente, estudamos a dinâmica de espalhamento da COVID-19 no Brasil. Investigamos como o número de casos e mortes por COVID-19 escalam com a população das cidades brasileiras. Os resultados indicam que as cidades menores são proporcionalmente mais afetadas durante o período inicial de espalhamento da doença. Por outro lado, em períodos posteriores, as cidades grandes têm maior incidência de casos e mortes por COVID-19. Argumentamos que essa vantagem urbana pode ter como causa a melhor infraestrutura de saúde e a menor proporção de idosos nos grandes centros urbanos. Entretanto, observamos que as taxas de crescimento têm comportamento oposto à incidência de casos e mortes por COVID-19, sendo inicialmente maiores e posteriormente menores para cidades grandes. Em segundo lugar, investigamos a associação entre produtividade e impacto dos pesquisadores bolsistas do CNPq. Observamos que a associação é dependente da disciplina e do estágio da carreira e, além disso, é similar entre pesquisadores com comportamento *outlier* ou não *outlier*. Pesquisadores *outliers* tendem a ter performances além do normal apenas em produtividade ou impacto, mas raramente nas duas categorias. Pesquisadores não *outliers* apresentam uma associação negativa entre essas duas grandezas com intensidade dependente da disciplina. Em anos consecutivos da carreira, a tendência é de manutenção dos níveis de produtividade e impacto. Por fim, os pesquisadores, em média, apresentam produtividade crescente e impacto decrescente ao longo da carreira.

Palavras-chave: Sistemas Complexos. Ciência da Ciência. COVID-19. Análise de Dados. Física Estatística.

Abstract

In this work, we study complex systems by using statistical physics and data science methods. In the first place, we study the spreading dynamics of the COVID-19 in Brazil. We investigate how the number of cases and deaths by COVID-19 scale with city size. Our results indicate that small cities are proportionally more affected during the initial stages of propagation. On the other hand, large cities present a higher incidence of cases and deaths by COVID-19 in posterior stages. We argue this urban advantage could arise from the improved health infrastructure and the proportionally smaller elderly population in larger cities. However, we observe growth rates to display the opposite behavior, that is, they are initially larger but eventually decrease in posterior stages for large cities. In the second place, we investigate the association between productivity and impact of Brazilian researchers. We observe the association to be discipline specific, career stage dependent, and similar among researchers with outlier and nonoutlier performances. Outlier researchers tend to outperform in productivity or impact, but rarely outperform in both categories. Similarly, nonoutlier researchers present a negative association between productivity and impact with discipline-dependent intensity. Overall, researchers tend to maintain productivity and impact levels over consecutive career years. On average, researchers also display increasing productivity and decreasing impact over career years.

Keywords: Complex Systems. Science of Science. COVID-19. Data Science. Statistical Physics.

Sumário

Introdução

Wang e Barabási, na introdução de seu livro "*The Science of Science*" [1], afirmam que "revoluções científicas são geralmente impulsionadas pela invenção de novos instrumentos – o microscópio, o telescópio, o sequenciamento genético – cada qual mudando radicalmente nossa habilidade de perceber, mensurar e raciocinar o mundo". Em verdade, o advento de novos instrumentos muda não apenas a maneira como praticamos ciência, mas também nossa própria vivência cotidiana. No decorrer das últimas décadas, a utilização de aparelhos eletrônicos (como celulares, computadores, GPS, e *smartwatches*) incorporou-se ao dia a dia da maioria das pessoas. De 2000 até 2016, o número de usuários da internet cresceu em oito vezes [2] (Figura 1A). O número de usuários em redes sociais apresentou taxas de crescimento parecidas num período mais recente [2] (Figura 1B). Em 2019, 82,7% dos domicílios brasileiros possuíam acesso à internet [3]. Dessa forma, podemos perceber que a tecnologia tem se incorporado cada vez mais à vida humana nas últimas décadas. Retornando ao escopo acadêmico, observamos avanços tecnológicos principalmente na capacidade de processamento, na precisão de sensores e na expansão do armazenamento de dados. Por exemplo, os sensores do colisor de partículas LHC produzem por volta de 90 *petabytes*[1] de dados por ano [4]. Esse feito era inimaginável décadas atrás, seja pela precisão dos sensores ou pela imensa quantidade de dados armazenados. É desse generalizado, facilitado e desenvolvido uso de dispositivos eletrônicos que surge uma das mais destacadas ferramentas científicas dos últimos tempos: a ampla disponibilidade de dados digitais. Hoje, presenciamos uma variedade enorme de conjuntos de dados sobre os mais diversos assuntos, de sistemas biológicos, físicos e naturais até sociais[2].

É nesse contexto de massiva disponibilidade de dados sobre os mais diversos temas que surge a possibilidade de estudar os sistemas denominados *complexos*, o tema de investiga-

[1] 1 *petabyte* = 1.000.000.000.000.000 *bytes*.

[2] Acesse o catálogo "Awesome Public Datasets" para uma extensa lista de conjuntos de dados públicos.

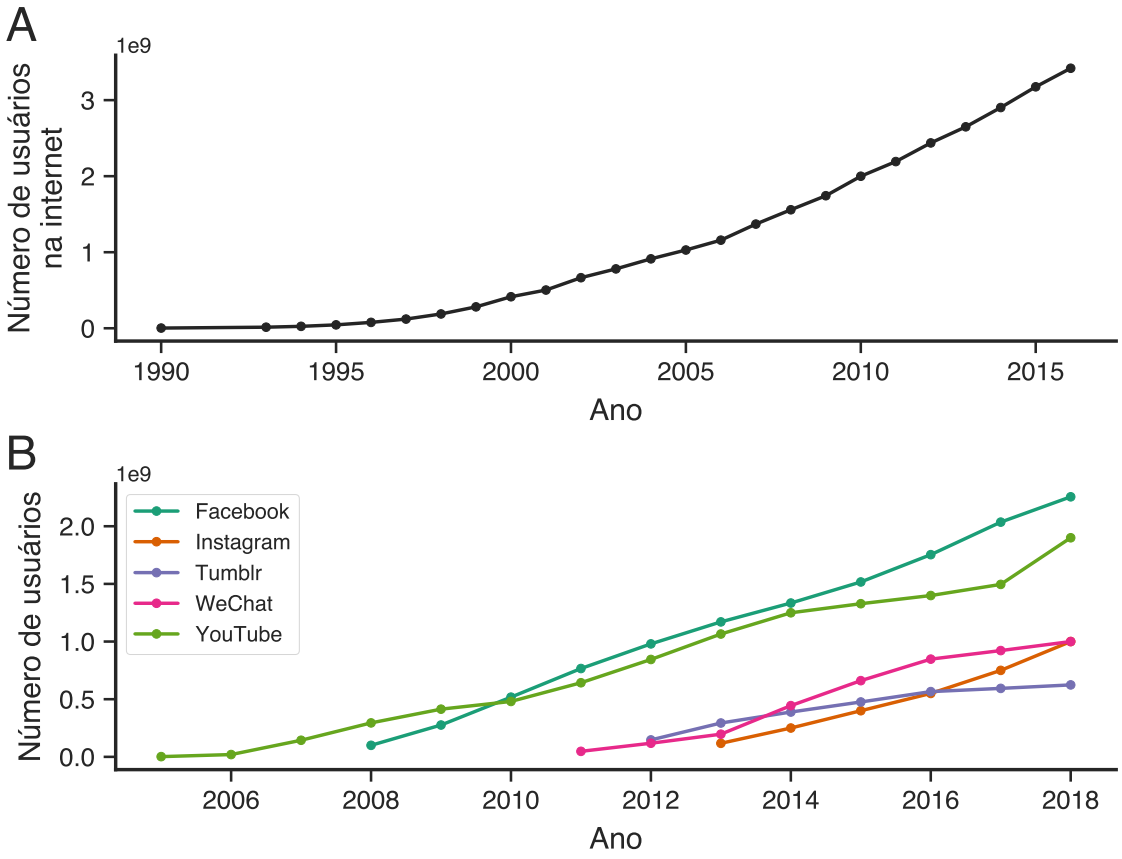

Figura 1: Crescimento no uso da internet. **(A)** Evolução temporal do número de usuários da internet. **(B)** Evolução temporal do número de usuários de diversas redes sociais.

ção de nosso laboratório de pesquisa ComplexLab. Os sistemas complexos são geralmente caracterizados, porém, não exclusivamente, por determinados atributos [5–7]: comportamento *emergente* (as partes não explicam o todo); grande número de agentes interagentes; ausência de comando central; dinâmica não linear ou fora do equilíbrio; comportamento auto-organizado; comportamento persistente no tempo; paralelismo nas ações; adaptação, aprendizado e evolução. No estudo desses sistemas, utilizamos um arsenal de ferramentas da matemática, da estatística e da ciência de dados a fim de extrair suas tendências, associações e padrões.

Como exemplo, podemos citar investigações do nosso grupo de pesquisa sobre redes de corrupção [8, 9], cristais líquidos [10–12], leis de escala urbana [13–17], mercado financeiro [18, 19], arte [20], linguagem natural [21], ciência [22], política [23, 24], plantas aquáticas [25] e esportes [26]. Também vale mencionar que a aplicação de métodos estatísticos para o estudo de sistemas complexos extrapola o mundo acadêmico. Recentemente, um novo poço de petróleo no estado do Rio de Janeiro foi descoberto pela Petrobras via ferramentas de inteligência artificial [27]. Além disso, na recente pandemia da COVID-19, nosso grupo de pesquisa manteve o portal Observatório COVID-19 Maringá para acompanhamento dos

números da doença na cidade [28]. Nesse caso, podemos também enxergar o espalhamento da COVID-19 como um sistema complexo. Dentre os indicadores que acompanhamos está o número de reprodução R [29,30]. Esse indicador é uma medida epidemiológica que indica o número médio de infecções decorrentes de um indivíduo que se tornou infeccioso no tempo t. Valores acima do limiar $R = 1$ indicam que a doença está num ritmo acelerado de espalhamento, enquanto valores abaixo de $R = 1$ indicam uma desaceleração no espalhamento. A Figura 2 mostra a progressão do número de reprodução durante os seis primeiros meses desde o primeiro caso de COVID-19 em Maringá. No período inicial da pandemia, na ausência de vacinas contra a COVID-19, as medidas de contenção da doença restringiam-se ao isolamento social, ao uso de máscaras e à higiene pessoal. Diante disso, o acompanhamento de indicadores de espalhamento da doença (como o R) para a adoção de medidas de restrição ou flexibilização do isolamento social foi de suma importância para evitar o colapso do sistema de saúde da cidade.

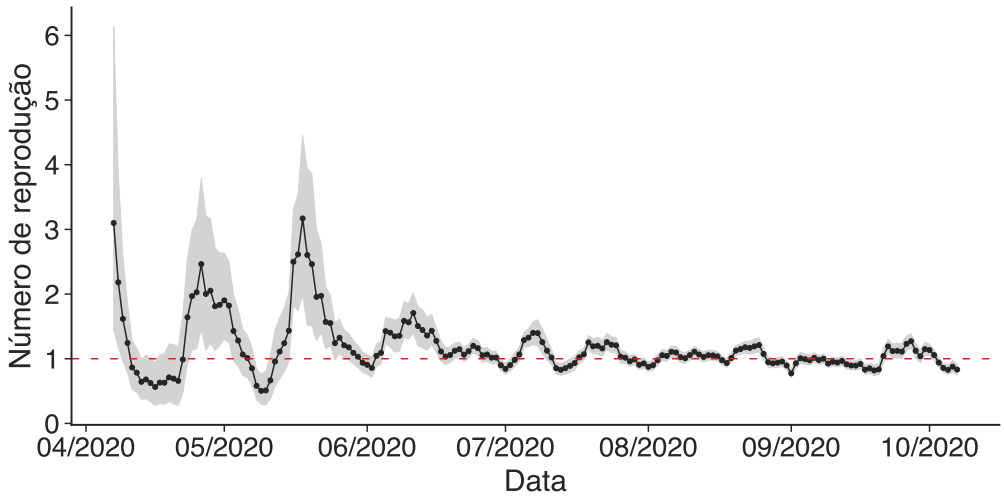

Figura 2: Número de reprodução da COVID-19 em Maringá. A ilustração mostra a evolução do número de reprodução da COVID-19 nos primeiros seis meses desde o primeiro caso da doença em Maringá. Valores acima de um indicam o espalhamento acelerado da doença, enquanto valores abaixo de um indicam a diminuição no ritmo de transmissão da doença. A região sombreada denota o intervalo de confiança de 90%.

Como podemos ver, o campo de estudo de sistemas complexos é muito abrangente. Neste trabalho, entretanto, focamos nosso estudo na dinâmica de dois sistemas complexos específicos e bastante distintos. No Capítulo 2, investigamos a dinâmica de espalhamento da COVID-19 no Brasil [31]. Mais especificamente, investigamos como o número de casos e de mortes por COVID-19 escalam com a população das cidades brasileiras. No Capítulo 3, caracterizamos o comportamento (em relação à produtividade e impacto) dos bolsistas produtividade em pesquisa do CNPq considerando a disciplina, etapa da carreira e performance do pesquisador [32].

CAPÍTULO 1

Métodos estatísticos para análise de dados

Neste capítulo, fundamentamos os métodos estatísticos empregados nas análises descritas no decorrer do texto. Sugerimos a leitura a partir do Capítulo 2 para o leitor com mais familiaridade com os tópicos a seguir.

1.1 Regressão Linear

Neste trabalho, utilizamos o modelo linear para estudar as leis de escala da COVID-19 no Brasil (Capítulo 2). Além disso, empregamos uma variante dessa regressão, o modelo linear misto, para estimar a associação entre a produtividade e o impacto das revistas científicas que publicam trabalhos de pesquisadores brasileiros (Capítulo 3). Este último modelo será detalhado em uma seção posterior a este capítulo.

Para definir o modelo linear simples, considere os vetores $\boldsymbol{x} = (x_1, \ldots, x_N)^\top$ e $\boldsymbol{y} = (y_1, \ldots, y_N)^\top$ representando o conjunto de dados das variáveis, respectivamente, independente e dependente de um conjunto de dados de tamanho N. Supondo que a relação estatística entre as variáveis x e y apresenta dispersão uniforme, podemos escrever a regressão linear simples como [33]

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ , \tag{1.1}$$

em que ε_i representa os efeitos aleatórios e o índice i refere-se à i-ésima observação do conjunto de dados. Podemos também escrever essa relação estatística em notação matricial.

Para isso, começamos definindo o vetor de variáveis dependentes

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} ,$$

e a matriz de regressores

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} ,$$

que também é conhecida como matriz *design* [33]. A matriz *design* anterior possui duas colunas: a primeira é unitária representando a hipótese de intercepto constante; a segunda contém os elementos da variável independente x representando a influência de x em y a qual queremos inferir. A regressão tornar-se-ia multivariada caso houvesse colunas adicionais com outras variáveis independentes [34]. Os parâmetros da Eq. (1.1) podem ser representados por um vetor de parâmetros

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} .$$

Em se tratando de uma regressão linear multivariada, a derivada parcial de y em relação a uma variável independente x particular é equivalente ao coeficiente linear β correspondente. Dessa maneira, o coeficiente β equivale à variação em y devido ao acréscimo unitário em x mantendo as demais variáveis constantes. Por esse motivo, os coeficientes $\boldsymbol{\beta}$ são chamados de coeficientes parciais de regressão [34]. Com base no exposto, podemos escrever o modelo linear em notação matricial

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} ,$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \beta_0 + x_1\beta_1 \\ \beta_0 + x_2\beta_1 \\ \vdots \\ \beta_0 + x_N\beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix} ,$$

em que $\boldsymbol{\varepsilon}$ é o vetor de termos aleatórios. O valor esperado de nosso modelo linear é dado por

$$\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta} .$$

A variância de $\boldsymbol{Y}$ e o valor esperado de $\boldsymbol{\varepsilon}$ são

$$Var(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I}$$
$$\text{e} \quad \mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \ ,$$

em que $\mathbf{0}$ é o vetor nulo de dimensão N.

Em seguida, precisamos definir a distribuição do termo aleatório da Eq. (1.1). Aqui, escolhemos uma distribuição normal, assumindo que o comportamento da variável dependente depende de vários fatores aleatórios não inclusos no modelo. Considerando ainda que esses fatores adicionais são independentes, o Teorema Central do Limite garante que a distribuição do erro é gaussiana no limite em que o número de variáveis não inclusas é suficientemente grande [33]. Dessa maneira, podemos escrever

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

em que $\boldsymbol{I}$ é a matriz identidade de dimensão N e σ^2 é a variância. A partir dessa definição, podemos escrever o vetor de variáveis dependentes $\boldsymbol{Y}$ distribuído como

$$\boldsymbol{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}) \ . \tag{1.2}$$

Assim, cada y_i é uma variável aleatória com sua média e variância correspondentes; $\boldsymbol{\beta}$ é o vetor de coeficientes em que β_0 representa o intercepto e β_1 representa a inclinação da relação linear; e a estrutura do erro é bem definida – uma distribuição normal de média nula e variância σ^2 (Figura 1.1).

Para estimar os parâmetros do modelo (σ^2, β_0 e β_1), empregamos o método de máxima verossimilhança [35]. Para cada y_i, temos a verossimilhança definida como

$$f(y_i; \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{[y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2} \right\} , \tag{1.3}$$

em que os parâmetros β_0, β_1 e σ^2 são destacados após o ponto e vírgula pois ainda não foram estimados.

A verossimilhança do modelo é a distribuição conjunta para todos os y_i. Considerando a independência das variáveis, a densidade conjunta é o produto das densidades individuais

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left[-\frac{1}{2\sigma^2} |\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}|^2 \right] . \tag{1.4}$$

Os parâmetros, então, podem ser estimados maximizando a verossimilhança, isto é, a distribuição de probabilidade conjunta do conjunto de dados em função dos parâmetros do modelo. Para facilitar os cálculos, podemos aplicar o logaritmo na Eq. (1.4). Essa transfor-

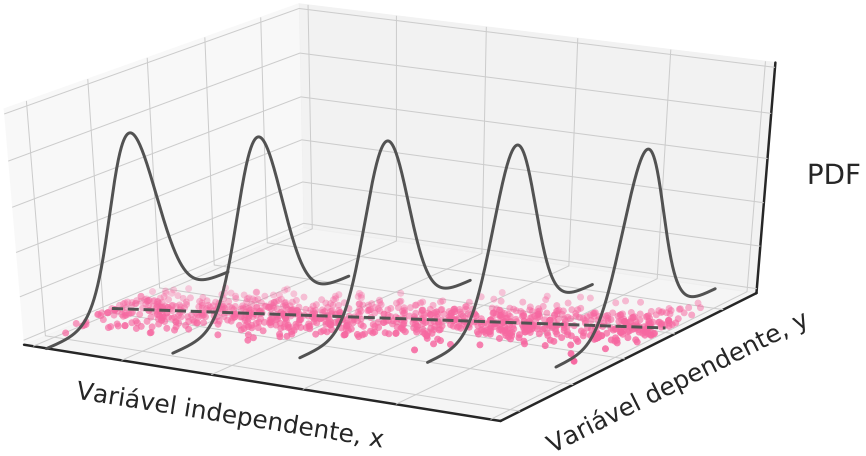

Figura 1.1: Exemplo de regressão linear simples. A reta pontilhada ilustra um modelo linear e as curvas representam a suposição do erro distribuído normalmente.

mação lineariza a equação preservando a localização de seu máximo. Os valores de $\boldsymbol{\beta}$ e σ^2 podem ser estimados por

$$\frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta}} = \frac{\partial |\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}|^2}{\partial \boldsymbol{\beta}} = \mathbf{0} \ ,$$

conduzindo a

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} \ .$$

De maneira similar, temos

$$\hat{\sigma}^2 = \frac{|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}|^2}{N - 2} \ ,$$

em que os símbolos munidos de "chapéu" representam uma estimativa do parâmetro. Para a variância, o denominador foi alterado a fim de remover o viés do tamanho da amostra [33].

1.2 Regressão Logística

A regressão logística é utilizada no estudo de variáveis binárias, isto é, uma variável dependente y_i que representa a realização ($y_i = 1$) ou não realização ($y_i = 0$) de um evento [36]. A Figura 1.2A mostra um gráfico de dispersão de uma variável dependente binária y como função de uma variável independente contínua x. Podemos entender cada y_i como um ensaio de Bernoulli [37], isto é,

$$P(y) = \pi(x)^y [1 - \pi(x)]^{1-y} \ , \tag{1.5}$$

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, isto é, 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

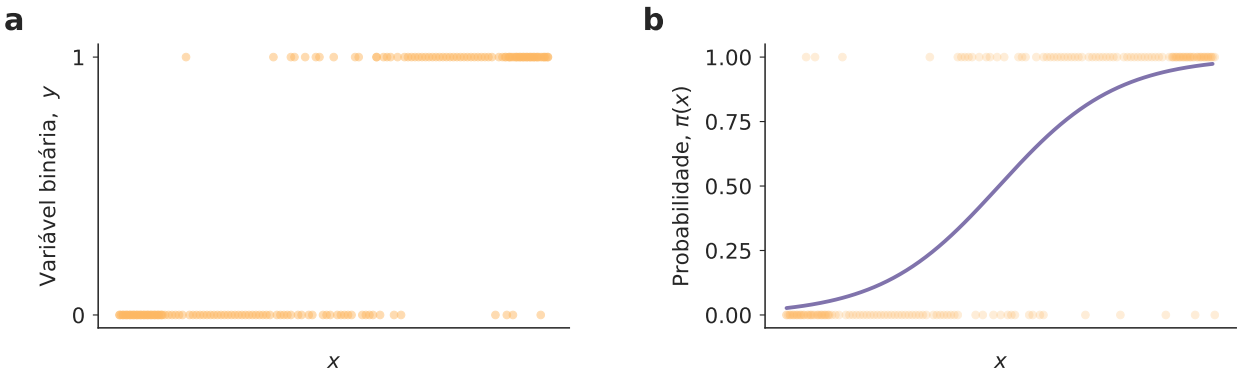

Figura 1.2: Exemplo de regressão logística. **(A)** Diagrama de dispersão de uma variável binária y em relação a uma variável contínua x. **(B)** Exemplo de ajuste de uma curva logística.

mapear a variável x para o intervalo $[0, 1]$ [37]

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

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

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x \ , \tag{1.6}$$

em que definimos a função *logit* (logística) [37] de $\pi(x)$ na forma do modelo linear simples como descrito na Eq. (1.2). A Figura 1.2B ilustra a curva logística ajustada a um conjunto de dados binários gerado artificialmente em função de uma variável contínua x. Na forma logística, 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, ou seja,

$$\text{chance} := \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x \ .$$

Da mesma maneira que definimos o modelo na seção 1.1, podemos escrever a relação para cada x_i uma vez que $\pi(x)$ depende de x, ou seja,

$$\sum_{i=1}^{k} y_i = \mathbb{E}[y|x] + \varepsilon_i = k\pi(x) + \varepsilon_i \ , \tag{1.7}$$

em que k é o número total de eventos para um valor específico de x e ε_i é seu erro corres-

pondente. Como estamos tratando de eventos de Bernoulli, o erro ε_i pode ser caracterizado como uma distribuição binomial com valor esperado dado por [38]

$$\begin{aligned}\mathbb{E}[\varepsilon_i] &= \mathbb{E}\left[\sum_{i=1}^{k} y_i\right] - \mathbb{E}[k\pi(x)] \\ &= k\pi(x) - k\pi(x) \\ &= 0 \ ,\end{aligned} \tag{1.8}$$

e variância

$$\begin{aligned}Var[\varepsilon_i] &= Var\left[\sum_{i=1}^{k} y_i\right] - Var[k\pi(x)] \\ &= k\pi(x)[1-\pi(x)] - 0 \\ &= k\pi(x)[1-\pi(x)] \ .\end{aligned} \tag{1.9}$$

A estimação dos parâmetros pode ser realizada, novamente, por meio do método da máxima verossimilhança. 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

$$\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \ , \tag{1.10}$$

com y_i podendo assumir os valores 0 (fracasso) ou 1 (sucesso). Supondo independência entre as observações, a verossimilhança assume a forma do produto das contribuições individuais, ou seja,

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \ . \tag{1.11}$$

Aplicamos o logaritmo à verossimilhança para linearizar a equação, temos

$$\log \mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \{y_i \log \pi(x_i) + (1-y_i)\log[1-\pi(x_i)]\} \ . \tag{1.12}$$

Para encontrar as estimativas dos parâmetros β_0 e β_1, diferenciamos a Eq. (1.12) em relação a cada um deles e igualamos a zero, obtendo

$$\begin{aligned}&\sum_{i=1}^{N}[y_i - \pi(x_i)] = 0 \quad \text{e} \\ &\sum_{i=1}^{N} x_i[y_i - \pi(x_i)] = 0 \ .\end{aligned} \tag{1.13}$$

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* [39] do *Python*.

1.3 Regressão Linear Mista

A regressão linear simples estudada na seção 1.1 apresenta determinadas limitações em função das suposições do modelo: relação linear; normalidade do erro; homoscedasticidade (variância constante); independência estatística; distribuição idêntica dos dados. Em dados do mundo real, algumas dessas suposições podem ser violadas e, de fato, o são com frequência.

Por exemplo, considere um estudo longitudinal de biologia em que coletamos dados temporais do peso de galinhas na tentativa de inferir a taxa de crescimento média da espécie [40]. Notamos que cada animal apresenta um peso particular para um tempo t desde seu nascimento. Esse fato decorre de inúmeras causas – como fatores genéticos e ambientais – e tem como efeito prático o não colapso das curvas de crescimento mesmo que, porventura, exista uma taxa de crescimento única da espécie (Figura 1.3A). Nesse caso, há violação das hipóteses de independência estatística (há correlação entre os dados de cada animal) e de distribuição idêntica das variáveis (as observações provêm de distribuições com médias distintas) da regressão linear simples. Além disto, determinadas galinhas podem apresentar mais observações. Nessa situação, o conjunto de dados é denominado desbalanceado. Os animais com maior quantidade de observações serão mais influentes no resultado da regressão, ocultando a relação verdadeira entre as variáveis estudadas. Para solucionar esse problema, recorremos à regressão linear mista [41], que considera a estrutura hierárquica dos dados.

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" [40]. 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 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

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

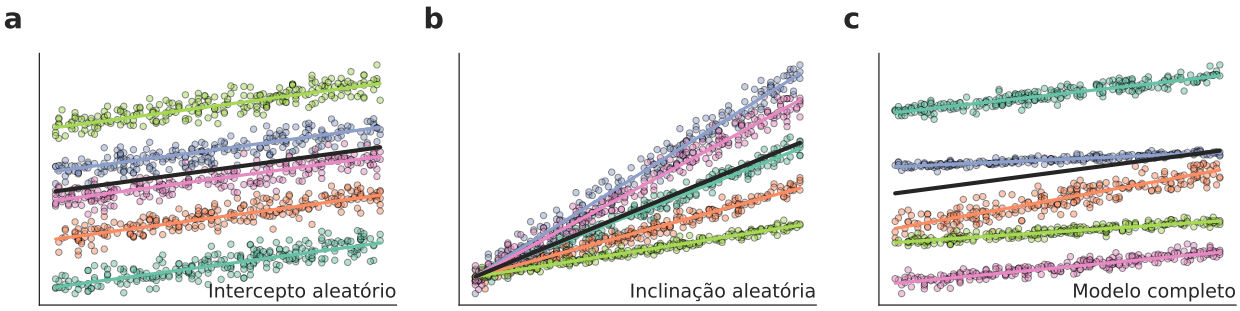

Figura 1.3: Ilustração de modelos lineares mistos. (A) Intercepto aleatório. **(B)** Inclinação aleatória. **(C)** Modelo completo com inclinação e intercepto aleatórios. As linhas contínuas em preto representam os valores médios do modelo

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 1.3A 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 1.3B 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 1.3C 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 1.3, fica evidente que a regressão linear simples não é capaz de capturar a estrutura hierárquica desse tipo de dados por conta de suas limitações já 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.

1.3.1 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* da linguagem R [42]. 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

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

em que $\boldsymbol{\mu}_{Y|\mathcal{U}=\mathbf{u}}$ é o vetor de preditores lineares, $\mathbf{Zb}$ 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 a matriz de covariância positiva e semi-definida $\boldsymbol{\Sigma}$,

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

Com intuito de permitir a singularidade em $\boldsymbol{\Sigma}$ [42], definimos $\boldsymbol{\Sigma}$ em termos de um fator relativo de covariância $\boldsymbol{\Lambda_\theta}$, cujos parâmetros $\boldsymbol{\theta}$ correspondem aos pesos dos elementos da matriz covariância dos efeitos aleatórios, isto é,

$$\boldsymbol{\Sigma_\theta} = \sigma^2\boldsymbol{\Lambda_\theta}\boldsymbol{\Lambda_\theta}^\top \ .$$

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

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

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

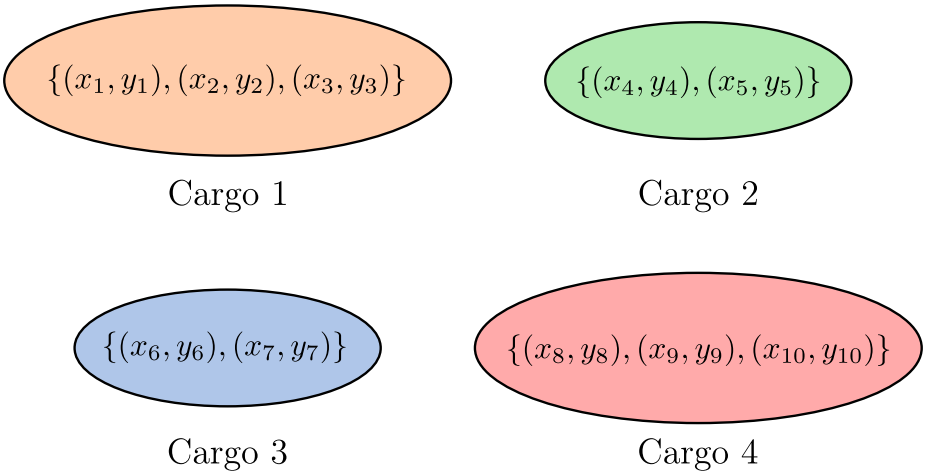

Figura 1.4: Disposição hipotética dos dados de salário por cargo em uma empresa. Nesse caso, o conjunto de dados consiste de $N = 10$ observações (funcionários) que estão distribuídas entre $l = 4$ grupos (cargos).

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

$$\mathcal{B} = \boldsymbol{\Lambda_\theta} \mathcal{U} \ .$$

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

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

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

$$\underbrace{\boldsymbol{Y}}_{N\times 1} = \underbrace{\overbrace{\boldsymbol{X}}^{N\times p}\ \overbrace{\boldsymbol{\beta}}^{p\times 1}}_{N\times 1} + \underbrace{\overbrace{\boldsymbol{Z}}^{N\times q}\ \overbrace{\boldsymbol{\Lambda_\theta}}^{q\times q}\ \overbrace{\boldsymbol{u}}^{q\times 1}}_{N\times 1} + \underbrace{\boldsymbol{\varepsilon}}_{N\times 1} \ , \tag{1.14}$$

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 $\boldsymbol{\Lambda_\theta}$, retomemos o exemplo da progressão salarial em um empresa. Suponha que tenhamos um conjunto de dados estruturados como mostra a Fi-

gura 1.4 e que queremos modelar a taxa de crescimento salarial como um efeito aleatório. A 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

$$\boldsymbol{Z} = \begin{pmatrix} x_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ x_2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ x_3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & x_4 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & x_5 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & x_6 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & x_7 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & x_8 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & x_9 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & x_{10} & 1 \\ \underbrace{}_{\text{Cargo 1}} & & \underbrace{}_{\text{Cargo 2}} & & \underbrace{}_{\text{Cargo 3}} & & \underbrace{}_{\text{Cargo 4}} & \end{pmatrix}. \tag{1.15}$$

A matriz de covariância relativa correspondente é dada por

$$\boldsymbol{\Lambda_\theta} = \begin{pmatrix} a & . & . & . & . & . & . & . \\ c & b & . & . & . & . & . & . \\ . & . & a & . & . & . & . & . \\ . & . & c & b & . & . & . & . \\ . & . & . & . & a & . & . & . \\ . & . & . & . & c & b & . & . \\ . & . & . & . & . & . & a & . \\ . & . & . & . & . & . & c & b \end{pmatrix} ,$$

$$\underbrace{\quad}_{\text{Cargo 1}} \quad \underbrace{\quad}_{\text{Cargo 2}} \quad \underbrace{\quad}_{\text{Cargo 3}} \quad \underbrace{\quad}_{\text{Cargo 4}}$$

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

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

1.4 Modelos hierárquicos bayesianos

Diferentemente do tratamento dado às regressões linear simples e logística, utilizamos uma abordagem bayesiana para estimar os parâmetros do modelo linear misto. Naquele contexto, empregamos o método frequentista de estimação dos parâmetros por máxima verossimilhança. Neste contexto, porém, usamos 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 [43]

$$P(\boldsymbol{\theta}|D) = \frac{P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathbf{D})} \ , \tag{1.16}$$

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), $P(\mathbf{D})$ 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).

Assim, precisamos definir a distribuição de probabilidade a *posteriori* do modelo hierárquico a partir da Eq. (1.16), a qual será amostrada para encontrar as distribuições marginais dos parâmetros e, portanto, as próprias estimativas dos parâmetros (veja a Seção 1.5 para detalhes sobre o processo de amostragem). 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 a 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

$$\boldsymbol{Y}_{ij}|\boldsymbol{\theta}_j \sim P(y_{ij}|\boldsymbol{\theta}_j) \ . \tag{1.17}$$

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 [44], isto é,

$$P(\boldsymbol{y}_j|\boldsymbol{\theta}_j) = \prod_{i=1}^{n_j} P(y_{ij}|\boldsymbol{\theta}_j) \ . \tag{1.18}$$

Além disso, sendo os parâmetros específicos de cada j-ésimo grupo independentes, podemos escrever a distribuição a *priori* como [44]

$$P(\boldsymbol{\theta}|\boldsymbol{\phi}) = \prod_{j=1}^{l} P(\boldsymbol{\theta}_j|\boldsymbol{\phi}) \ . \tag{1.19}$$

Em termos bayesianos, os "parâmetros dos parâmetros" são denominados hiperparâmetros e suas distribuições a *priori* correspondentes são chamadas distribuições a *hiperpriori* [45]. De modo particular, 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 [46].

Escolhemos as distribuições a *priori* e *hiperpriori* adequadas e, a partir do Teorema de Bayes, conseguimos definir a distribuição hierárquica para o modelo hierárquico de dois níveis como [44]

$$\begin{aligned} P(\boldsymbol{\theta}, \boldsymbol{\phi}|D) &\propto P(D|\boldsymbol{\theta})P(\boldsymbol{\theta}|\boldsymbol{\phi})P(\boldsymbol{\phi}) \\ &\propto p(\boldsymbol{\phi}) \prod_{j=1}^{l} P(\boldsymbol{\theta}_j|\boldsymbol{\phi})P(\boldsymbol{y}_j|\boldsymbol{\theta}_j) \ . \end{aligned} \tag{1.20}$$

1.5 Amostrador No-U-Turn

Em problemas do mundo real, o cálculo da distribuição a *posteriori* dificilmente é realizado de forma analítica por meio da Eq. (1.16). A razão da dificuldade está no cálculo do

denominador, isto é, no cálculo da integral

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

Sua solução analítica é trabalhosa e frequentemente inviável, sendo possível apenas nos casos mais simples. A solução numérica, por sua vez, funciona apenas para um número reduzido de dimensões em tempo computacional razoável. Esse não é o caso para maior parte dos problemas reais. Para exemplificar isso, vamos considerar um modelo linear misto com intercepto β_0 e inclinação β_1 aleatórios considerando a estrutura da Figura 1.4. 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

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

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

$$\mathbb{E}[\boldsymbol{\theta}|D] = \int_{\text{todo } \boldsymbol{\theta}} \boldsymbol{\theta}P(\boldsymbol{\theta}|D)d\boldsymbol{\theta} \ , \tag{1.21}$$

e da variância

$$\begin{aligned} Var[\boldsymbol{\theta}|D] &= \mathbb{E}[\boldsymbol{\theta}^2|D] - (\mathbb{E}[\boldsymbol{\theta}|D])^2 \\ &= \int \boldsymbol{\theta}^2 P(\boldsymbol{\theta}|D)d\boldsymbol{\theta} - \left[\int \boldsymbol{\theta}P(\boldsymbol{\theta}|D)d\boldsymbol{\theta}\right]^2 \ , \end{aligned} \tag{1.22}$$

também envolvem 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, isto é, precisamos estimar, no caso mais simples,

$$\begin{aligned} P(\boldsymbol{\theta}|D) &= \frac{P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathbf{D})} \\ &\propto P(D|\boldsymbol{\theta})P(\boldsymbol{\theta}) \ , \end{aligned} \tag{1.23}$$

ou, num modelo hierárquico, por meio da relação expressa na Eq. (1.20). Uma possível solução é, portanto, 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) [47]. A vantagem dos métodos MCMC é que não precisamos saber antecipadamente a forma da distribuição a *posteriori*. De modo detalhado, 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 [48] e o algoritmo de Gibbs [49]. Entretanto, esses algoritmos apresentam dificuldades ao tratar de modelos multidimensionais. A principal dificuldade decorre do fenômeno conhecido como "concentração de medida" [50]. 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. Dessa forma, o caminhante aleatório não consegue explorar todo o espaço dos parâmetros. Por esse motivo, neste trabalho, optamos por utilizar 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 da mecânica clássica [51, 52]. 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

$$\pi(\boldsymbol{q}, \boldsymbol{p}) = \pi(\boldsymbol{p}|\boldsymbol{q})\pi(\boldsymbol{q}) \ , \tag{1.24}$$

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

$$\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}) = -\log \pi(\boldsymbol{q}, \boldsymbol{p}) \ , \tag{1.25}$$

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

$$\begin{aligned} P &\propto e^{-\mathcal{H}(\boldsymbol{q},\boldsymbol{p})} \\ P &\propto \pi(\boldsymbol{q}, \boldsymbol{p}) \ . \end{aligned} \tag{1.26}$$

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

$$\mathcal{H}(\boldsymbol{q}, \boldsymbol{p}) = K(\boldsymbol{p}, \boldsymbol{q}) + V(\boldsymbol{q}) \ , \tag{1.27}$$

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. Nessa formulação, a energia potencial é definida como o logaritmo da distribuição a *posteriori*. A energia cinética, por sua vez, pode ser escolhida de maneira mais conveniente para cada modelo estudado [50]. Se escolhemos um sistema na ausência de atrito, a energia total é

constante e as trajetórias ficam confinadas a um nível energético determinado, ou seja,

$$\mathcal{H}^{-1}(\mathbf{E}) = \{\boldsymbol{q}, \boldsymbol{p} | \mathcal{H}(\boldsymbol{q}, \boldsymbol{p}) = E\} \ , \tag{1.28}$$

hipersuperfícies em $(2D-1)$ dimensões do espaço de parâmetros com dimensão D. Ainda, podemos decompor a distribuição canônica em termos microcanônicos:

$$\pi(\boldsymbol{q}, \boldsymbol{p}) = \pi(q_E|E)\pi(\mathbf{E}), \tag{1.29}$$

em que $\pi(q_E|E)$ é a distribuição microcanônica e $\pi(\mathbf{E})$ 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, ou seja, considerando o sistema sem atrito da Eq. (1.28); e *ii)* a exploração estocástica dos níveis de energia, representados pela distribuição marginal de energias na Eq. (1.29), 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 roxo 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 $\pi(\mathbf{E})$ da Eq. (1.29), sendo esta representativa de todas as energias relevantes ao sistema. Neste caso ideal, a amostragem é realizada de maneira independente [50]. Comumente, a escolha da forma da energia cinética restringe-se a dois tipos: Gaussiana-Euclidiana e Gaussiana-Riemanniana [50].

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

$$K(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{M}^{-1}\boldsymbol{p} + \log|\boldsymbol{M}| + \textit{constante} \ , \tag{1.30}$$

em que $\boldsymbol{M}$ é a matriz de massa responsável por realizar transformações lineares na *posteriori* [53]. Os elementos de variância esticam ou comprimem os parâmetros da *posteriori* para que eles apresentem a mesma escala, enquanto os elementos de covariância rotacionam a *posteriori* a fim de 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

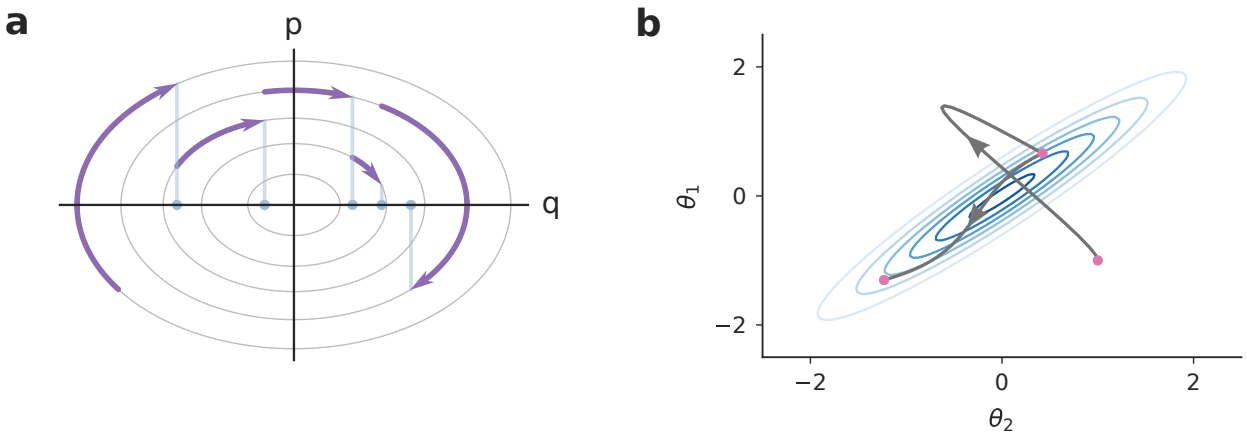

Figura 1.5: Amostrador de Monte Carlo Hamiltoniano. (A) Espaço de fases e transições energéticas do algoritmo HMC. As elipses concêntricas ilustram os níveis energéticos e as coordenadas permitidas. As curvas roxas representam as trajetórias no espaço de fase. As linhas verticais indicam as posições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

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,

$$K(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{2}\boldsymbol{p}^\top [\boldsymbol{\Sigma}(\boldsymbol{q})]^{-1}\boldsymbol{p} + \log|\boldsymbol{\Sigma}(\boldsymbol{q})| + \text{constante} \ , \tag{1.31}$$

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 [50], 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

$$K(\boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^\top \boldsymbol{p} + \text{constante} \ , \tag{1.32}$$

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

$$\begin{aligned} \frac{d\boldsymbol{q}}{dt} &= \frac{\partial \mathcal{H}(\boldsymbol{p}, \boldsymbol{q})}{\partial \boldsymbol{p}} = \frac{\partial K(\boldsymbol{p})}{\partial \boldsymbol{p}} + \frac{\partial V(\boldsymbol{q})}{\partial \boldsymbol{p}} \\ \frac{d\boldsymbol{p}}{dt} &= -\frac{\partial \mathcal{H}(\boldsymbol{p}, \boldsymbol{q})}{\partial \boldsymbol{q}} = -\frac{\partial K(\boldsymbol{p})}{\partial \boldsymbol{q}} + -\frac{\partial V(\boldsymbol{q})}{\partial \boldsymbol{q}} \ . \end{aligned} \tag{1.33}$$

Após realizar as simplificações, temos que

$$\begin{aligned} \frac{d\boldsymbol{q}}{dt} &= \boldsymbol{p} \\ \frac{d\boldsymbol{p}}{dt} &= -\frac{\partial V(\boldsymbol{q})}{\partial \boldsymbol{q}} \ . \end{aligned} \tag{1.34}$$

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

- Amostragem do momento $\boldsymbol{p} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$;
- Simulação das trajetórias no espaço de parâmetros, $\boldsymbol{q}(t)$ e $\boldsymbol{p}(t)$, por meio das equações de Hamilton em T passos temporais;
- 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 [50]. Aqui, recorremos a métodos numéricos a fim de 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* [54] descrito por

$$\begin{aligned} \boldsymbol{p}_{t+\epsilon/2} &= \boldsymbol{p}_t + (\epsilon/2)\nabla_{\boldsymbol{q}} V(\boldsymbol{q}_t) \\ \boldsymbol{q}_{t+\epsilon} &= \boldsymbol{q}_t + \epsilon \boldsymbol{p}_{t+\epsilon/2} \\ \boldsymbol{p}_{t+\epsilon} &= \boldsymbol{p}_{t+\epsilon/2} + (\epsilon/2)\nabla_{\boldsymbol{q}} V(\boldsymbol{q}_{t+\epsilon}) \ , \end{aligned} \tag{1.35}$$

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

$$\nabla_{\boldsymbol{q}} V(\boldsymbol{q}_t) \to \frac{\partial V(q_{t,i})}{\partial q_i} \ , \tag{1.36}$$

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 preservem o hipervolume e respeitem o princípio de balanço detalhado na cadeia de Markov [55]. Após realizar a trajetória por L passos, o passo é realizado com probabilidade

$$\alpha = \min\left(1, \frac{\exp V(\tilde{\boldsymbol{q}}) - \frac{1}{2}\tilde{\boldsymbol{p}} \cdot \tilde{\boldsymbol{p}}}{\exp V(\boldsymbol{q}_0) - \frac{1}{2}\boldsymbol{p}_0 \cdot \boldsymbol{p}_0}\right) \ , \tag{1.37}$$

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 Eq. (1.37) indica a energia perdida entre 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 [50]. Se pudéssemos calcular exatamente a trajetória, obteríamos sempre que $\alpha = 1$ e as proposições seriam sempre aceitas. O algoritmo 1 descreve

o código com uma implementação do HMC. Para que haja reversibilidade temporal e respeito do balanço detalhado, trajetórias aceitas têm coordenadas de momento com sinal trocado armazenadas [50].

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 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 $\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}}$

Em amostradores mais simples, como o amostrador de Metropolis e Gibbs, as proposições dos passos são aleatórias ao redor da posição atual e, em decorrência disso, passos subsequentes são altamente correlacionados. Esse fato pode gerar dificuldades para o caminhante alcançar regiões distantes da posição atual dependendo da geometria da distribuição a *posteriori* [53]. Para o amostrador HMC, a autocorrelação em seus passos ainda existe, mas num grau muito menor se comparado aos mencionados anteriormente, uma vez que o método utiliza informações sobre a geometria da distribuição a *posteriori* na construção dos 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 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, uma vez que esse algoritmo é muito sensível à variação desses parâmetros [55]. No caso de trajetórias curtas, o comportamento é o mesmo de um caminhante aleatório, que avança apenas para regiões circundantes à posição atual. No caso de trajetórias longas, por outro lado, o caminhante, num comportamento cíclico, pode acabar visitando as mesmas regiões desnecessariamente. Por isso, podemos incorporar ao amostrador HMC um procedimento que cessa a progressão das trajetórias no sinal 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*) [55].

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. Nos passos subsequentes, uma direção aleatória é escolhida por meio do sorteio $u \sim \mathcal{U}(\{-1, 1\})$, sendo o tamanho do deslocamento é sempre o dobro do tamanho na iteração anterior[1], pois, assim, podemos construir 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 respectivas coordenadas de momento. O critério de parada do algoritmo, para uma métrica euclidiana, pode ser definido como [55]

$$\begin{aligned} & \boldsymbol{p}_+(t)^\top \cdot [\boldsymbol{q}_+(t) - \boldsymbol{q}_-(t)] < 0 \\ \text{e} \quad & \boldsymbol{p}_-(t)^\top \cdot [\boldsymbol{q}_-(t) - \boldsymbol{q}_+(t)] < 0 \end{aligned} , \tag{1.38}$$

em que se define que os momentos das extremidades estão alinhados de maneira contrária à linha que une as posições [50]. 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. Após a parada, a amostra do algoritmo NUTS é sorteada a partir de todas as posições por que o caminhante passou até a última iteração. Como não há probabilidade de aceitação no NUTS, empregamos a taxa de aceitação média do algoritmo hamiltoniano tradicional na última dobra [55].

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 ($\mathcal{H} \to \infty$). Essa situação denomina-se *transição divergente* e acarreta baixas taxas de aceitação [53]. 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

$$G_t = \delta - \alpha_t \ , \tag{1.39}$$

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

$$\mathbb{E}_t[G_t|\epsilon] = g(\epsilon) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[G_t|\epsilon] \ . \tag{1.40}$$

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

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

$$\begin{aligned} \epsilon_{t+1} &\leftarrow \mu - \frac{\sqrt{t}}{\gamma}\frac{1}{t+t_0}\sum_{i=0}^{t} G_i \\ \bar{\epsilon}_{t+1} &\leftarrow \nu_t \epsilon_{t+1} + \bar{\epsilon}_t - \nu_t \bar{\epsilon}_t \end{aligned}, \tag{1.41}$$

em que μ é o valor de convergência escolhido para ϵ_t, $\gamma > 0$ define a intensidade de concentração para μ, t_0 é um valor que estabiliza as primeiras iterações, a somatória refere-se ao valor médio de G_t até o tempo t, ν_t é o tamanho do passo em cada iteração e definimos $\bar{\epsilon}_1 = \epsilon_1$. Nesse caso, 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 $\sum_t \nu_t^2 < \infty$ [55]. 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, é possível consultar as referências [55–57].

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 proposições (Figura 1.6A). O algoritmo de Gibbs, por sua vez, não apresenta o problema de alta taxa de rejeição, porém, tem dificuldades em explorar uniformemente as regiões da distribuição alvo (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 ao certo qual sua verdadeira forma e quantas iterações são necessárias para representá-la bem. Por isso, pre-

[2]Para uma comparação entre as diversas técnicas de amostragem, acesse o aplicativo do link https://chi-feng.github.io/mcmc-demo/app.html. [Último acesso em 5 de maio de 2022]

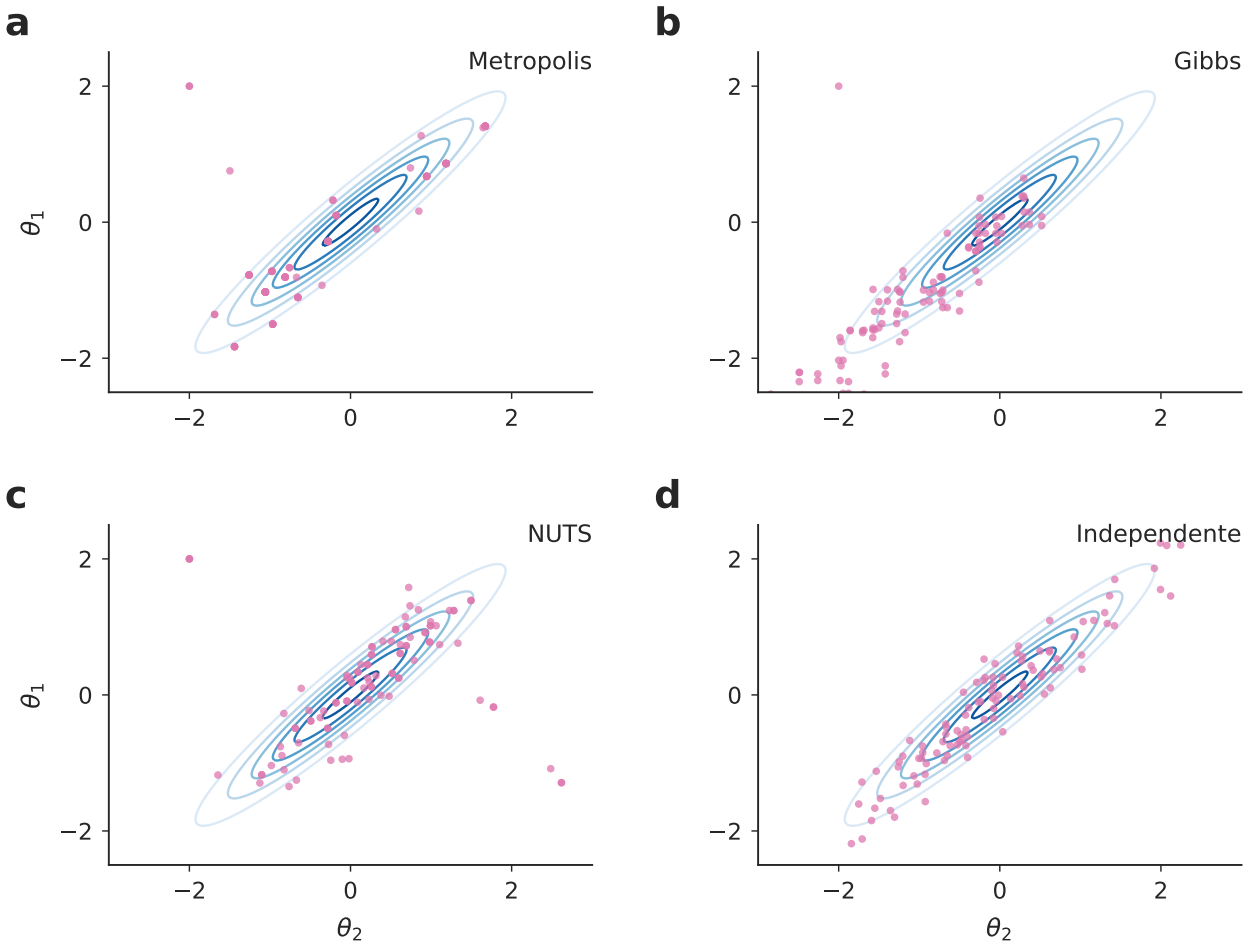

Figura 1.6: Tipos de amostradores e sua performance. (A) Amostragem de Metropolis. **(B)** Amostragem de Gibbs. **(C)** Amostragem NUTS. **(D)** Amostragem independente. Os marcadores em rosa representam os pontos amostrados.

cisamos de métricas para quantificar a convergência das múltiplas cadeias de Markov, pois a convergência é um bom indício de que a distribuição amostrada representa a distribuição verdadeira [45]. 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

$$W = \frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} (\theta_{ij} - \overline{\theta}_j)^2 \ , \tag{1.42}$$

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 da 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 [58]

$$\begin{aligned} var(\hat{\theta}|D) &= \frac{n-1}{n}W + \frac{1}{n}B \\ &= W + \frac{1}{n}(B - W) \ , \end{aligned} \tag{1.43}$$

uma variância superestimada, mas não enviesada no estado estacionário das cadeias [58]. 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 [45, 58]

$$\hat{R} = \sqrt{\frac{W + \frac{1}{n}(B - W)}{W}} \ , \tag{1.44}$$

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 ($B \gg W$) o valor de R chapéu é grande ($\hat{R} \gg 1$). 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 R chapéu converge para um ($\hat{R} \to 1$). Na prática, no entanto, é comum considerar o valor $\hat{R} \approx 1.1$ como indício de boa mistura das cadeias [45].

Outra ferramenta útil para avaliar a mistura das múltiplas cadeias de Markov é a visualização denominada *trace plot* [45, 59]. O *trace plot* ilustra a evolução temporal (isto é, a cada iteração) 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 [44]

$$n_{ef} = \frac{mT}{1 + 2\sum_{\tau=1}^{\infty} \rho_\tau} \ , \tag{1.45}$$

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

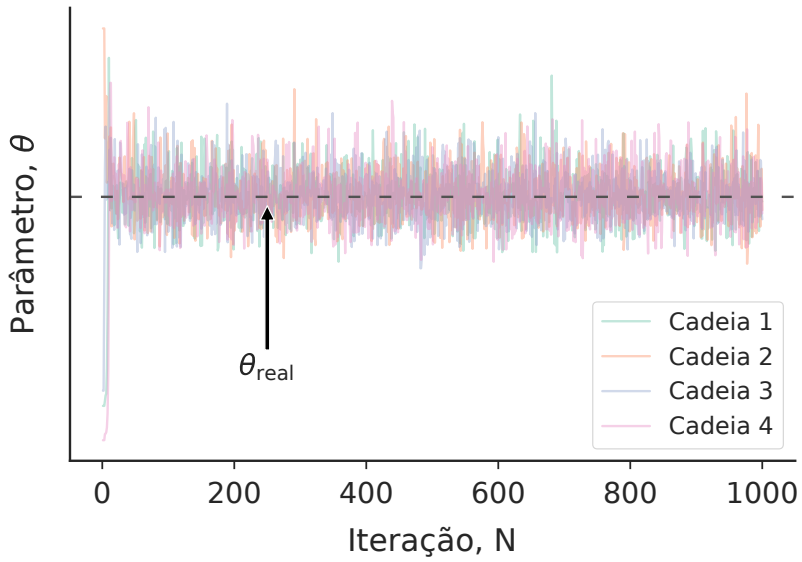

Figura 1.7: Mistura das cadeias de Markov. O *trace plot* de um processo ilustrativo de amostragem usando quatro cadeias de Markov.

quantidade de passos realizados de maneira independente para um amostrador que utiliza cadeias de Markov.

1.6 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

$$\mu = \frac{1}{N-1}\sum_{i=1}^{N} y_i \ , \tag{1.46}$$

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

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2 \ . \tag{1.47}$$

Porém, essas medidas deixam de ser representativas na presença de *outliers*[3], isto é, na presença de um único *outlier* divergente $y_k \to \infty$, as medidas de média e variância também divergem.

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

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 habitual. Ela é definida como o valor central que divide a amostra 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,

$$\text{MAD} = k\ \text{mediana}(|y_i - \text{mediana}(y)|)\ , \tag{1.48}$$

em que estabelecemos a constante $k = 1.4826$ para que o estimador MAD seja consistente com o desvio padrão [60]. Entretanto, apesar da simplicidade dessas medidas, intuitivamente elas deixam de apresentar a interpretação de valor médio e variância, representando, respectivamente, o ponto médio e o desvio desse ponto médio.

Daqui em diante, 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 distribuição 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

$$\mathcal{L}(\mu, \sigma) = \sum_i \sigma^{-1} f\left(\frac{y_i - \mu}{\sigma}\right)\ , \tag{1.49}$$

em que a somatória abrange todo o conjunto de dados e a distribuição de probabilidade é normalizada e centrada na origem. Assim como realizamos nas seções 1.1 e 1.2, aplicamos uma transformação logarítmica que preserva as características do máximo da verossimilhança e tomamos seu negativo, isto é,

$$\rho = -\log \mathcal{L}(\mu, \sigma) = \sum_i \left[\log \sigma - \log f\left(\frac{y_i - \mu}{\sigma}\right)\right]\ . \tag{1.50}$$

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, isto é, o "M" dos estimadores-M decorre da "M"aximização da verossimilhança. Dessa forma, um estimador-M pode ser considerado qualquer variável que maximiza a expressão

$$\sum_i \psi(y_i; \theta) = 0\ , \tag{1.51}$$

em que $\psi(y_i; \theta)$ é a derivada de ρ em relação ao parâmetro θ [61]. As formas da equação de

[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 [61].

maximização para os parâmetros de localização e escala definidos na Eq. (1.50) são dadas, respectivamente, por

$$\begin{aligned} &\sum_i \psi\left(\frac{y_i-\mu}{\sigma}\right) = 0 \\ &\sum_i \left[\left(\frac{y_i-\mu}{\sigma}\right)\psi\left(\frac{y_i-\mu}{\sigma}\right) - 1\right] = 0 \ . \end{aligned} \tag{1.52}$$

O último passo para obtenção dos estimadores-M é a escolha de uma função ψ adequada [62]. A função ψ representa a função geradora da amostra sobre a qual queremos inferir as medidas de localização e escala. Vamos listar algumas possíveis escolhas para a função ψ. Primeiramente, temos a função

$$\psi(y) = \begin{cases} y & \text{se } |y| < c \\ 0 & \text{caso contrário} \end{cases} , \tag{1.53}$$

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

$$\psi(y) = \begin{cases} -c & \text{se } y < -c \\ y & \text{se } |y| < c \\ c & \text{se } y > c \end{cases} , \tag{1.54}$$

em que o peso é não nulo de valor c. Essa é a função geradora originalmente proposta por Huber [63]. 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, no caso da proposta de Huber, as caudas são distribuições exponenciais duplas, ou seja,

$$\rho_H(y) = \begin{cases} y^2 & \text{se } |y| < c \\ c(2|y| - c) & \text{caso contrário} \end{cases} , \tag{1.55}$$

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

$$\psi(y) = y\left[1 - \left(\frac{y}{R}\right)^2\right]_+^2 , \tag{1.56}$$

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

$$\psi(y) = sign(x) \begin{cases} |y| & \text{se } 0 < |y| < a \\ a & \text{se } a < |y| < b \\ a(c - |y|)/(c - b) & \text{se } b < |y| < c \\ 0 & \text{se } c < |y| \end{cases} , \tag{1.57}$$

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 realizar uma pequena correção na equação de maximização da verossimilhança do parâmetro de escala a fim de que a estimativa não seja enviesada para a distribuição normal [62,64]. A expressão torna-se, então,

$$\begin{aligned} \sum_i \left[\left(\frac{y_i - \mu}{\sigma} \right) \psi \left(\frac{y_i - \mu}{\sigma} \right) \right] &= (n-1)a(c) \\ \sum_i \psi \left(\frac{y_i - \mu}{\sigma} \right) &= 0 \end{aligned} , \tag{1.58}$$

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

Computacionalmente, a solução das Eqs. (1.58), isto é, a estimativa dos parâmetros de localização e de escala, raízes da equação, pode ser realizada por meio do método numérico de Newton [65]. Partindo de valores próximos ao valor verdadeiro do parâmetro, o método consiste em aproximar a função como a reta tangente a ela mesma para estimar uma raiz aproximada e repete 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

$$y = h'(x_n)(x - x_n) + h(x_n) \ . \tag{1.59}$$

O valor da raiz aproximada é obtida igualando a Eq. (1.59) a zero, isto é,

$$x_{n+1} = x_n - \frac{h'(x_n)}{h(x_n)} \ . \tag{1.60}$$

A Figura 1.9 ilustra duas iterações do método de Newton para uma função arbitrária $h(x)$. Em nosso trabalho, utilizamos a mediana e o MAD como estimativas iniciais para calcular o valor dos parâmetros, respectivamente, de localização e de escala. As equações de

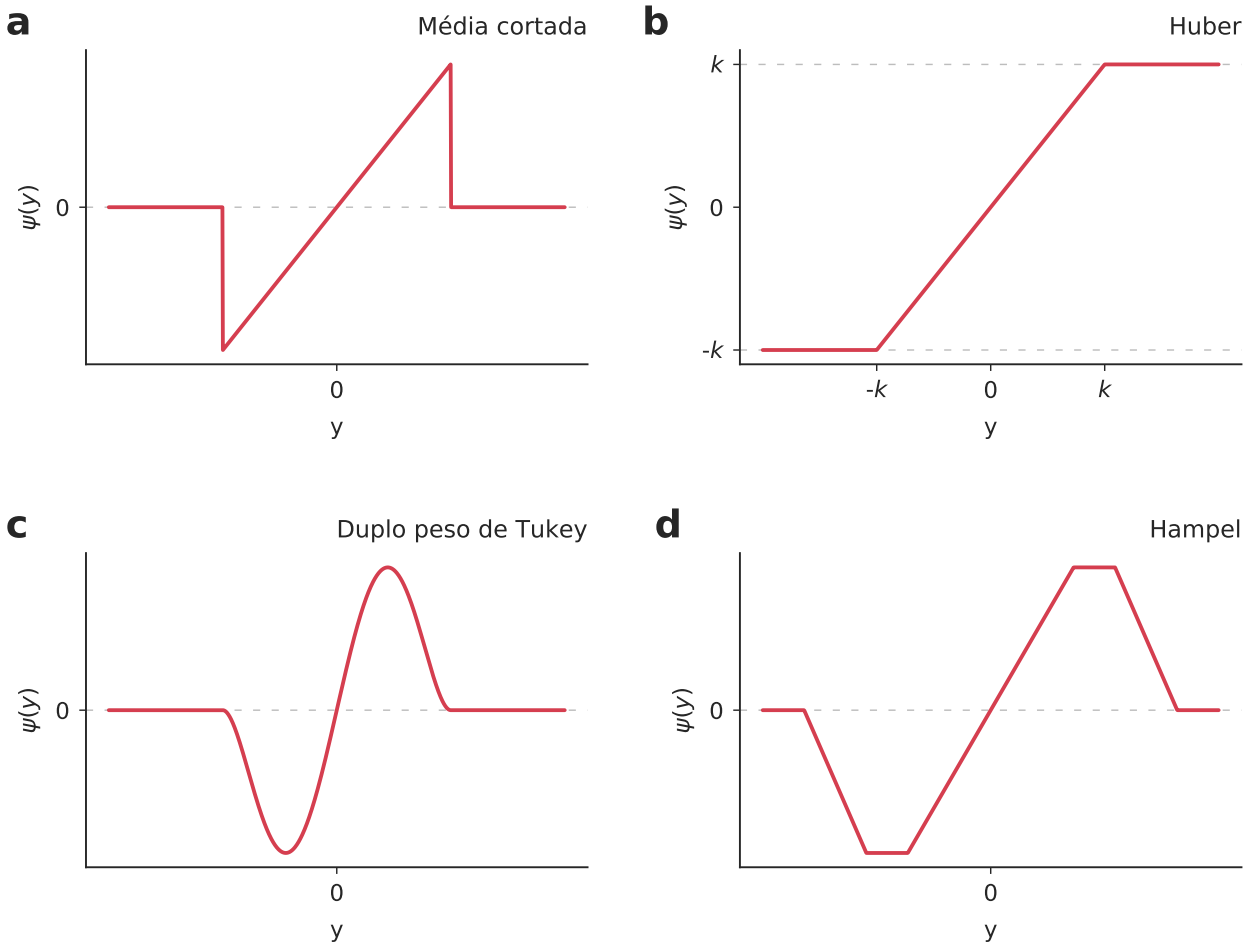

Figura 1.8: Diferentes funções $\psi(y)$ para determinação do estimador-M. (A) Média cortada. **(B)** Função de Huber. **(C)** Função duplo peso de Tukey. **(D)** Função de Hampel.

atualização nos valores dos parâmetros são dadas por

$$
\begin{aligned}
[\sigma_{n+1}]^2 &= \frac{1}{(n-1)a(c)} \sum_i \psi^2(y_{n,i})[\sigma_n]^2 \\
\mu_{n+1} &= \mu_n + \frac{\sum_i \psi(y_{n,i})\sigma_n}{\psi'(y_{n,i})} \ ,
\end{aligned}
\tag{1.61}
$$

em que ψ é o mesmo da Eq.(1.54) 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* [39] do *Python* para o cálculo dessas medidas. Por padrão, a constante $|c|$ é igual a 1.5 neste pacote.

[5]Para mais detalhes veja Staudte e Sheather [64].

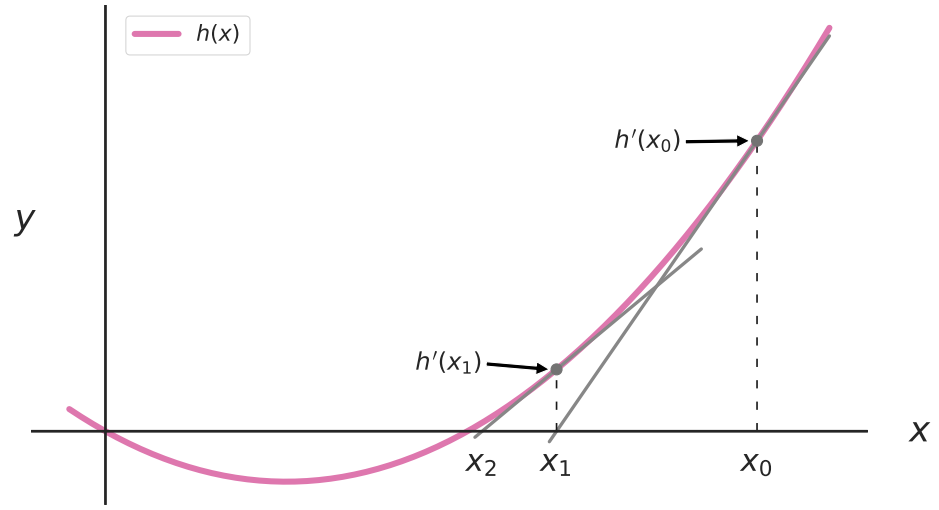

Figura 1.9: Ilustração do método de Newton para uma função arbitrária $h(x)$.

1.7 Coeficiente de correlação de Pearson

O coeficiente de correlação de Pearson representa a intensidade da associação linear entre duas variáveis x e y. Em sua forma amostral, podemos definir o coeficiente como [66]

$$r_{xy} = \frac{\sum_i (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_i (x_i - \mu_x)^2}\sqrt{\sum_i (y_i - \mu_y)^2}} \ , \tag{1.62}$$

em que x_i é a i-ésima observação de x, y_i é a i-ésima observação de y, μ_x é a média de x e μ_y é a média de y. Para interpretar o coeficiente de Pearson como uma espécie de covariância normalizada, podemos reescrever a Eq. (1.62) como

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\mathbb{E}[\sum_i (x_i - \mu_x)(y_i - \mu_y)]}{\sqrt{\mathbb{E}[\sum_i (x_i - \mu_x)^2]}\sqrt{\mathbb{E}[\sum_i (y_i - \mu_y)^2]}} \ , \tag{1.63}$$

em que σ_{xy} é a covariância entre x e y, σ_x é o desvio padrão de x e σ_y é o desvio padrão de y. A fim de definir o intervalo do coeficiente de Pearson, utilizamos a desigualdade de Cauchy-Schwarz

$$\begin{aligned} |\sigma_{xy}|^2 &\leq |\sigma_x||\sigma_y| \\ |\sigma_{xy}| &\leq \sqrt{|\sigma_x||\sigma_y|} \\ |r_{xy}| &\leq 1 \\ -1 \leq r_{xy} &\leq 1 \ . \end{aligned} \tag{1.64}$$

Assim, o valor do coeficiente r_{xy} está contido no intervalo $[-1, 1]$. No extremo positivo quando $r_{xy} = 1$, temos uma correlação linear positiva perfeita, isto é, uma variação em x acarreta

uma variação proporcional em y para qualquer observação da amostra. No outro extremo quando $r_{xy} = -1$, temos uma correlação linear negativa perfeita, ou seja, uma variação em x acarreta uma variação proporcional de sinal oposto em y para qualquer observação da amostra. No valor intermediário de $r_{xy} = 0$, temos a ausência de uma correlação linear entre as duas variáveis.

CAPÍTULO 2

Tamanho das cidades e o espalhamento da COVID-19 no Brasil

As atividades humanas têm se tornado cada vez mais concentradas em áreas urbanas. Desde o ano de 2007, uma proporção maior de pessoas vive em cidades [67], com projeções indicando que a população urbana global pode chegar a mais de 90% do total no fim deste século [68]. Além de ser cada vez mais urbanizado, atualmente o mundo apresenta condições excepcionais de mobilidade e conectividade: o número de passageiros de avião excedeu 4 bilhões de pessoas em 2018 [69]. De um lado, uma sociedade altamente urbanizada e conectada proporcionou altos níveis de inovação, de crescimento econômico, de acesso à educação e de acesso ao sistema de saúde; porém, por outro lado, também acarretou poluição, degradação ambiental, problemas de privacidade, sub-condições de moradia e condições favoráveis à disseminação de doenças infecciosas em nível global. Em particular, a emergência de surtos de doenças infecciosas tem crescido consideravelmente com o tempo. A maioria desses eventos é causada por patógenos originados em animais selvagens [70], o que, por sua vez, tem sido associado com mudanças climáticas, mudanças no uso da terra e nas práticas agrícolas e o crescimento de áreas com grande população humana [71].

O novo coronavírus (SARS-CoV-2) parece encaixar-se bem no contexto anteriormente citado, pois foi identificado pela primeira vez na cidade de Wuhan – uma influente cidade chinesa com mais de 11 milhões de habitantes – em dezembro de 2019 possivelmente de uma recombinação de vírus do tipo corona do morcego e do pangolim [72]. A "*corona virus disease 2019*" (COVID-19) inicialmente se espalhou pelo continente chinês, mas rapidamente gerou surtos em outros países, fazendo com que a Organização Mundial da Saúde (OMS) declarasse uma "emergência de saúde pública de escala internacional" em janeiro de 2020, sendo posteriormente reclassificada como uma pandemia em março de 2020. Em 16 de agosto de 2020, mais de 21.2 milhões de casos de COVID-19 foram confirmados em quase

todos os países e o número de mortes excede 761 mil pessoas [73]. A pandemia do novo coronavírus apresenta-se como uma ameaça sem precedentes para a saúde e a economia na nossa sociedade e, ao entender seus padrões de espalhamento, podemos encontrar fatores importantes para contribuir com sua mitigação e com o controle de seu avanço.

Trabalhos têm focado em modelar o espalhamento inicial da COVID-19 [74], em modelar as curvas de fatalidade [75], em projetar o pico do surto e a ocupação hospitalar [76], em entender os efeitos da mobilidade [77], da demografia [78], das restrições de viagem [79], das mudanças comportamentais humanas nos padrões de transmissão do vírus [80], das estratégias de mitigação [81], das intervenções não farmacêuticas [82], das estratégias de distanciamento social baseadas em métodos de rede [83], entre outros. Apesar do enorme volume de investigações científicas sobre o assunto, pouca atenção havia sido dada ao entendimento dos efeitos do tamanho das cidades nos padrões de espalhamento de casos e mortes por COVID-19 em áreas urbanas. A ideia de que tamanho (medido pela população) afeta diferentes indicadores urbanos tem sido extensivamente estudada e pode ser resumida pela hipótese de escala urbana [84–87]. Essa teoria dita que indicadores urbanos estão associados não linearmente com a população de cidades de tal maneira que indicadores socioeconômicos geralmente mostram retornos crescentes de escala [84, 88, 89], indicadores de infraestrutura geralmente apresentam economia de escala [84,85] e quantidades relacionadas a necessidades individuais geralmente escalam linearmente com a população de cidades [84, 85].

Estudos de escala urbana de variáveis associadas à saúde têm mostrado que a incidência e a mortalidade de doenças são relacionadas não linearmente com a população urbana [90–93]. Apesar da existência de várias exceções [93], doenças não infecciosas (como a diabetes) são comumente menos prevalentes em cidades grandes, enquanto doenças infecciosas (como a AIDS) são comumente mais prevalentes em cidades grandes. Essa diferença provavelmente reflete o fato de que pessoas vivendo em cidades grandes tendem a ter proporcionalmente mais contatos e um maior grau de interação social do que aquelas vivendo em cidades pequenas [85, 94]. Nesse contexto, o recente trabalho de Stier, Berman e Bettencourt [95] indica que cidades grandes dos Estados Unidos vivenciaram um aumento mais pronunciado nas taxas de crescimento de casos de COVID-19 durante as primeiras semanas após a introdução da doença no país. Similarmente, Cardoso e Gonçalves [96] descobriram que as taxas de contato per capita da COVID-19 aumentam com o tamanho e a densidade das cidades dos Estados Unidos, Brasil e Alemanha. Essas descobertas manifestam sérias consequências para a evolução da COVID-19 e sugerem que grandes metrópoles tendem a se tornar eixos de transmissão com picos potencialmente mais altos e mais prematuros de pessoas infectadas. Portanto, investigar se esse comportamento generaliza-se para outros lugares e como diferentes quantidades – tais como número de casos e mortes – escalam com o tamanho da cidade são elementos importantes para um melhor entendimento do espalhamento da COVID-19 em áreas urbanas.

Neste trabalho, investigamos como o tamanho populacional está associado com número de casos e mortes por COVID-19 nas cidades brasileiras. O Brasil é o sexto país mais populoso do mundo, com mais de 211 milhões de pessoas, das quais mais de 85% vivem em cidades. Apesar de ser muito provável que o novo coronavírus já estivesse circulando pelo Brasil no começo de fevereiro de 2020 [97], o primeiro caso confirmado no país data do dia 26 de fevereiro de 2020 na cidade de São Paulo. Entre o primeiro caso até 12 de agosto de 2020, o Brasil acumulou 3.088.670 casos confirmados de COVID-19 (o segundo maior número) espalhados por mais de 98.9% das 5.570 cidades brasileiras. A doença causou 102.817 mortes (o segundo maior número) com 3.892 cidades reportando pelo menos uma fatalidade em 12 de agosto de 2020.

2.1 Métodos

2.1.1 Dados

O conjunto de dados primário usado neste trabalho foi coletado da API `brasil.io` [98]. Essa API recupera dados dos boletins diários de COVID-19 publicados pelas secretárias estaduais de cada uma das 27 federações brasileiras (26 estados e o Distrito Federal), tornando os dados disponíveis gratuitamente. Esse conjunto de dados engloba informações sobre o número cumulativo de casos e mortes por COVID-19 de 25 de fevereiro de 2020 (data do primeiro caso no Brasil) até o dia 12 de agosto de 2020 (data da nossa última atualização) para todas as cidades brasileiras que reportaram ao menos um caso de COVID-19. A API `brasil.io` também fornece dados populacionais das cidades brasileiras, que, por sua vez, se baseiam em estimativas populacionais de 2019 do Instituto Brasileiro de Geografia e Estatística (IBGE). Havia um total de 5.507 cidades brasileiras com pelo menos um caso reportado de COVID-19 em 12 de agosto de 2020, correspondendo a 98,9% do número total de cidades do país. Além disso, 3.892 cidades reportaram mortes decorrentes dessa doença, representando 69,9% do total. Para garantir que nossas estimativas embasam-se em pelo menos 50 cidades, consideramos um limite superior adequado para o tamanho da série temporal (Figura A.1). Os dados demográficos de idade são referentes ao último censo brasileiro que aconteceu em 2010, enquanto dados sobre o número de leitos de UTI são de abril de 2020. Esses dois conjuntos de dados são mantidos e disponibilizados gratuitamente pelo Departamento de Informática do Sistema Único de Saúde (DATASUS) [99].

2.1.2 Ajustando leis de escala urbana

De modo geral, entende-se por escala urbana ou escalonamento urbano [84] a associação lei de potência entre uma propriedade urbana Y e a população P da cidade expressa por

$$Y = Y_0 P^{\beta} \,, \tag{2.1}$$

em que Y_0 é uma constante e β é o expoente de escala urbana. A equação (2.1) pode ser linearizada tomando o logaritmo em ambos os lados, isto é,

$$\log Y = \log Y_0 + \beta \log P \,, \tag{2.2}$$

em que $\log Y$ é a variável dependente e $\log P$ é a variável independente da relação linear entre $\log Y$ e $\log P$. Estimamos os expoentes da lei de potência da Eq. (2.1) usando a abordagem probabilística de Leitão *et al.* [100]: um modelo em que as flutuações lognormais em $\log Y$ são independentes de P. Neste trabalho, assumimos esse tipo de flutuação em todos os procedimentos de ajuste para estimar os valores de β via maximização da verossimilhança. O procedimento é análogo ao método dos mínimos quadrados (veja a Seção 1.1) com a diferença de que as variáveis são transformadas em seu logaritmo ($\log Y$ *versus* $\log P$).

2.1.3 Taxa de crescimento logarítmica de casos e mortes

Considere que x_t $(t = 1, \ldots, n)$ represente o número cumulativo de casos (Y_c) ou o número cumulativo de mortes (Y_d) por COVID-19 em uma dada cidade em determinado tempo t (número de dias desde o primeiro caso t_c ou morte t_d). A taxa de crescimento logarítmica r_t no tempo t é definida como

$$r_t = \log(x_t / x_{t-\tau})/\tau \quad (t = \tau, \tau + 1, \ldots, n) \tag{2.3}$$

em que τ é o atraso no tempo. Se assumirmos que os números de casos ou mortes inicialmente crescem exponencialmente ($x_t \sim e^{rt}$, em que r é a taxa de crescimento exponencial), r_t representa uma estimativa para a taxa de crescimento desse comportamento inicialmente exponencial. Estimamos r_t para o número de casos (r_c) e mortes (r_d) para valores até t_c e t_d, garantindo que o tamanho amostral fosse de, no mínimo, 50 cidades para estimar as relações alométricas entre as taxas de crescimento e as populações urbanas (Figura A.1). No texto principal, todos os resultados mostrados correspondem à escolha de $\tau = 14$. Porém, os resultados são robustos para τ entre 9 e 21 dias (Figuras A.2-A.15).

2.2 Resultados

Nossa investigação baseia-se em dados de boletins diários publicados pelas secretarias de saúde de cada uma dos 27 entes federativos brasileiros. Esses boletins atualizam o número de casos confirmados (Y_c) e o número de mortes (Y_d) por COVID-19 em todas as cidades brasileiras de 25 de fevereiro de 2020 (data do primeiro caso no Brasil) até 12 de agosto de 2020 (data da nossa última atualização). A partir desses dados, criamos as séries temporais do número de casos $Y_c(t_c)$ para cada cidade, em que t_c refere-se ao número de dias desde os primeiros dois casos diários reportados em cada cidade. Similarmente, criamos as séries temporais do número de mortes $Y_d(t_d)$, em que t_d refere-se ao número de dias desde as primeiras duas mortes diárias reportadas em cada cidade. Ao realizar esse procedimento, agrupamos todas as cidades de acordo com o estágio de propagação da doença (medido por t_c ou t_d) para investigar a evolução das relações alométricas entre o número total de casos ou mortes e a população da cidade. Também consideramos diferentes números iniciais diários de casos ou mortes como ponto de referência (de 1 a 7 casos ou mortes diárias) e nossos resultados permanecem robustos (Figuras A.16-A.29).

A Figura 2.1A mostra a relação entre casos de COVID-19 e população das cidades em escala logarítmica ($\log Y_c$ *versus* $\log P$) para diferentes números de dias desde os primeiros dois casos diários ($t_c = 15, 58, 101$ e 141 dias). O comportamento aproximadamente linear na escala logarítmica indica que o número de casos é bem descrito por uma função lei de potência da população das cidades

$$Y_c \sim P^{\beta_c}\,, \tag{2.4}$$

em que β_c é o chamado expoente de escala urbana [84], como já mencionado. Similarmente, a Figura 2.1B mostra a associação entre o número de mortes e a população das cidades em escala logarítmica ($\log Y_d$ *versus* $\log P$) para diferentes números de dias desde as primeiras duas mortes diárias ($t_d = 15, 50, 85$ e 120 dias). Novamente, os resultados indicam que o número de mortes é aproximado por uma função de lei de potência da população de cidade

$$Y_d \sim P^{\beta_d}\,, \tag{2.5}$$

em que β_d representa o expoente de escala urbana do número de mortes.

Os resultados da Figura 2.1 também mostram as relações alométricas ajustadas (linhas tracejadas) e os expoentes de escala β_c e β_d. Esses expoentes exibem uma tendência crescente com o tempo, excedendo o valor um após certo número de dias desde os primeiros dois casos ou duas mortes diárias. Esse comportamento dinâmico é melhor visualizado na Figura 2.2, em que ilustramos β_c e β_d como uma função do número de dias desde os primeiros dois casos diários (t_c) ou primeiras duas mortes diárias (t_d). O expoente de escala do número de casos β_c é sublinear ($\beta_c < 1$) durante os primeiros quatro meses e parece se aproximar de um platô

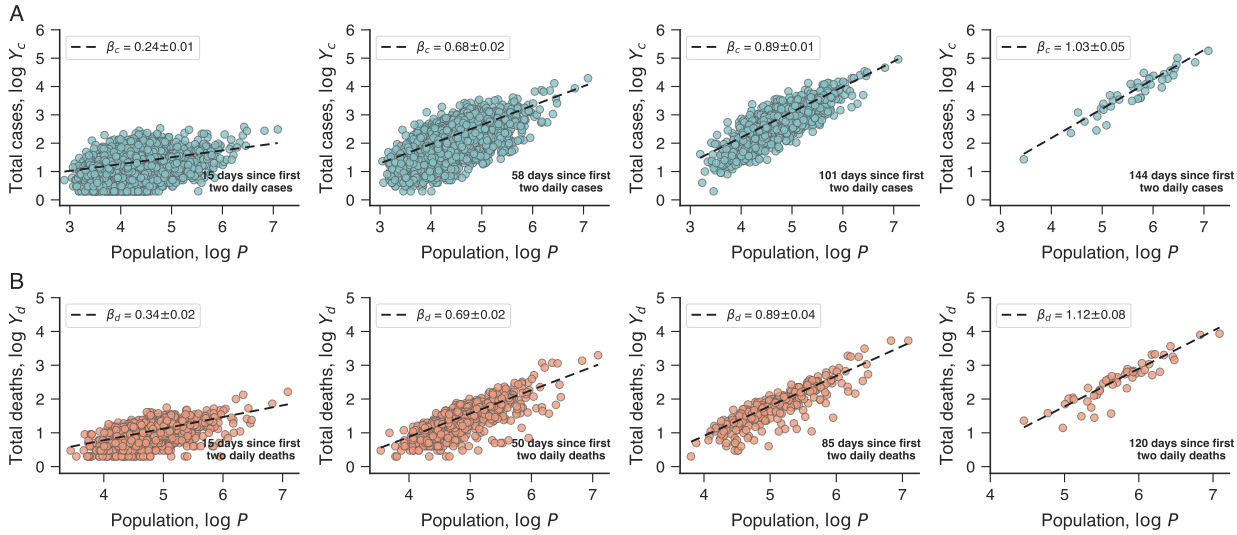

Figura 2.1: Relações de escala urbana de casos e mortes por COVID-19. (A) Relação entre o número total de casos de COVID-19 (Y_c) e a população urbana (P) em escala logarítmica. Painéis mostram as relações de escala para o número de casos em um dia particular após os primeiros dois casos diários reportados em cada cidade (quatro valores igualmente espaçados de t_c entre os primeiros 15 dias e o maior valor com pelo menos 50 cidades, como indicado nos gráficos). **(B)** Relação entre o número total de mortes por COVID-19 (Y_d) e população urbana (P) em escala logarítmica. Painéis mostram as relações de escala para o número de mortes em um dia particular após as primeiras duas mortes diárias reportadas em cada cidade (quatro valores igualmente espaçados de t_d entre os primeiros 15 dias e o maior valor com pelo menos 50 cidades, como indicado nos gráficos). Em todos os painéis, os marcadores representam cidades e as linhas tracejadas são as relações de escala ajustadas com os expoentes do melhor ajuste indicados em cada gráfico (β_c para casos e β_d para mortes).

superlinear ($\beta_c > 1$) conforme o número de dias t_c cresce. O comportamento dinâmico do expoente de escala para mortes β_d é similar ao β_c, todavia, β_d parece estar se aproximando a um platô mais alto do que aquele observado para β_c.

A evolução dos expoentes de escala para casos e mortes indica que pequenas cidades são proporcionalmente mais afetadas pela COVID-19 durante os quatro primeiros meses. Porém, essa aparente vantagem urbana desaparece com o tempo, tornando-se uma desvantagem depois de aproximadamente quatro meses. Podemos constatar isso diretamente estimando o número de casos *per capita* a partir da Eq. (2.4), isto é, $Y_c/P \sim P^{\beta_c - 1}$. Similarmente, podemos estimar o número de mortes *per capita* a partir da Eq. (2.5), isto é, $Y_d/P \sim P^{\beta_d - 1}$. Nessa perspectiva, esperamos que o número de casos *per capita* ou mortes *per capita* por COVID-19 diminuam com a população urbana se $\beta_c < 1$ e $\beta_d < 1$; de modo contrário, esses números *per capita* aumentam com a população se $\beta_c > 1$ e $\beta_d > 1$. Por exemplo, como $\beta_c \approx 0.77$ e $\beta_d \approx 0.85$ após 75 dias após os primeiros dois casos e duas mortes, o número de casos e de mortes *per capita* diminui com a população como $Y_c/P \sim P^{-0.23}$ e $Y_c/P \sim P^{-0.15}$. Para esses valores particulares de t_c e t_d, um aumento de 1% na população está associado

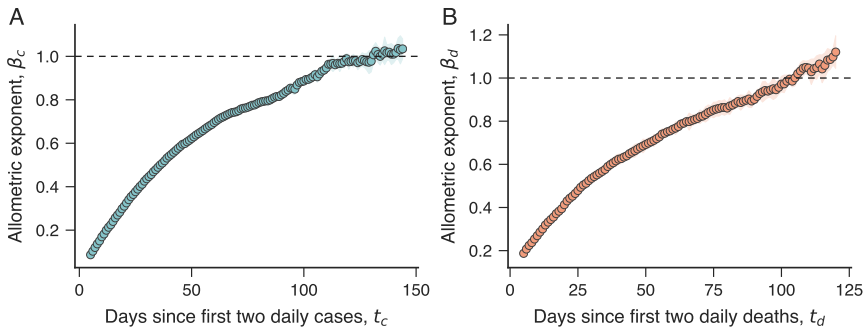

Figura 2.2: Dependência temporal dos expoentes de escala para casos e mortes por COVID-19. (A) Dependência do expoente β_c no número de dias após os primeiros dois casos diários de COVID-19 (t_c). **(B)** Dependência do expoente β_d no número de dias após as primeiras duas mortes diárias por COVID-19 (t_d). As regiões sombreadas em todos os painéis representam erros padrões estimados via *bootstrap* e as linhas tracejadas horizontais indicam a escala isométrica ($\beta_c = \beta_d = 1$). Notamos que β_c e β_d aumentam com o tempo e parecem se aproximar de valores assintóticos maiores do que um.

com uma diminuição de aproximadamente $\approx 0.23\%$ na incidência de casos de COVID-19 e uma diminuição de aproximadamente $\approx 0.15\%$ na incidência de mortes. Num exemplo concreto para $t_c = t_d = 75$ dias, esperamos que uma metrópole como São Paulo (com ≈ 12 milhões de pessoas) tenha $\approx 54\%$ menos casos e $\approx 39\%$ menos mortes *per capita* do que uma cidade de tamanho médio como Maringá/PR (com ≈ 420 mil pessoas, $\approx 1/30$ de São Paulo), que, por sua vez, deve ter $\approx 41\%$ menos casos e $\approx 29\%$ menos mortes *per capita* do que uma cidade de pequeno porte como Paranaíba/MS (com ≈ 42 mil pessoas, $\approx 1/10$ de Maringá).

Entretanto, ambos os expoentes de escala crescem com o tempo. Dessa forma, essa "vantagem urbana" desaparece e torna-se uma desvantagem em períodos posteriores da pandemia. Considerando as últimas estimativas dos expoentes de escala, encontramos $\beta_c \approx 1.04$ ($t_c = 144$ dias) e $\beta_d \approx 1.12$ ($t_d = 120$ dias). Nesses valores particulares de t_c e t_d, esperamos que o número de casos *per capita* aumente levemente com a população ($Y_c/P \sim P^{0.04}$) e o número de fatalidades *per capita* aumente com a população ($Y_d/P \sim P^{0.12}$). Portanto, para $\beta_d \approx 1.12$ em $t_d = 120$ dias, esperamos que uma metrópole como São Paulo (≈ 12 milhões de pessoas) tenha $\approx 50\%$ mais mortes *per capita* do que Maringá/PR (≈ 420 mil pessoas), que, por sua vez, deve ter $\approx 32\%$ mais mortes *per capita* do que Paranaíba/MS (≈ 42 mil pessoas). As Figuras A.24-A.29 mostram que as relações de escala para o número de casos e mortes *per capita* sustentam as discussões anteriores.

As últimas estimativas de β_c para os casos de COVID-19 são menores que aquelas reportadas para a pandemia de H1N1 de 2009 no Brasil ($\beta_c \approx 1.20$) e de HIV no Brasil e nos Estados Unidos ($\beta_c \approx 1.44$) [93]. De maneira parecida com o que observamos para casos de COVID-19, o expoente alométrico para os casos de HIV no Brasil era inicialmente sublinear durante os anos 80, tornou-se superlinear depois dos anos 90 e começou a se aproximar de

um platô superlinear depois dos anos 2000 [93]. Entretanto, a evolução da alometria do HIV tem sido muito mais lenta do que a observada para a COVID-19. Outro ponto interessante reportado por Rocha, Thorson, e Lambiotte [93] é que o número de casos de H1N1 no Brasil começou a escalar linearmente com a população urbana em 2010 (um ano após o primeiro surto). Os autores argumentam que a redução no expoente de escala possivelmente reflete a melhor resposta para o espalhamento da H1N1 após o surto pandêmico. Se o comportamento observado na pandemia de H1N1 de 2009 generalizar (pelo menos em partes) para a atual pandemia da COVID-19, esperamos que haja uma diminuição nos valores de β_c no futuro. As últimas estimativas de β_d para mortes por COVID-19 são maiores que os reportados para diabetes ($\beta_d \approx 0.97$), ataque cardíaco ($\beta_d \approx 1.04$) e acidente vascular cerebral ($\beta_d \approx 1.00$) no Brasil depois dos anos 2000 [93]. Em contrapartida, expoentes de escala relacionados à mortalidade por doenças no Brasil mostraram uma tendência decrescente com o tempo e valores tão altos como 1.25 foram observados para diabetes em 1996 ($\beta_d \approx 1.22$) e ataque cardíaco em 1981 ($\beta_d \approx 1.25$) [93]. A convergência desses expoentes para regimes lineares ou sublineares pode refletir o aumento crescente ao acesso a serviços de saúde em áreas urbanas [93].

Baseado nos dados disponíveis mais recentes (Figura 2.2), é difícil ter certeza de que os valores de β_c e β_d permanecerão maiores do que um em períodos posteriores da pandemia. Porém, a persistência desse comportamento indica que cidades grandes podem ser mais afetadas ao fim do surto da COVID-19. Parte desse comportamento pode ser explicado devido ao maior proporção de testagem em cidades grandes quando comparado a cidades pequenas. Resultados para os Estados Unidos apontam que estados mais rurais têm menores taxas e detectam desproporcionalmente menos casos de COVID-19 [101]. Como cidades brasileiras também são suscetíveis a esse viés, esperamos uma diminuição no expoente de escala β_c depois de observar tendências crescentes nos números de casos dependendo da magnitude desse efeito (isto é, à medida em que cidades pequenas aumentam sua capacidade de testagem, seu número de casos tende a aumentar e "dobrar" a lei de escala para baixo).

Por outro lado, percebemos que cidades grandes foram proporcionalmente menos afetadas durante os meses iniciais (desde os primeiros dois casos ou duas mortes) da pandemia. Acreditamos que existem ao menos duas explicações para esse comportamento. Primeiro, o comportamento pode refletir uma "crescente vantagem urbana" em que as cidades grandes possuem maior acesso a serviços de saúde e, assim, apresentam maior a chance de fornecer um tratamento adequado para a COVID-19. A segunda explicação pode estar associada com mudanças na demografia etária que varia de acordo com o tamanho da população urbana; especificamente, uma proporção reduzida de idosos em alto risco de enfermidades severas e/ou morte por COVID-19 acarreta um número reduzido de mortes *per capita*. Uma terceira possibilidade está relacionada às diferenças estratégicas de resposta à COVID-19 entre cidades pequenas e cidades grandes. Diferentes ações acarretam diferentes graus de eficácia

na contenção da pandemia. No contexto dos Estados Unidos, essas respostas são altamente heterogêneas a nível nacional [102, 103] bem como entre seus condados [104]. Dentre essas três possibilidades, não exploramos os possíveis efeitos das diferentes estratégias adotadas pelas cidades contra a COVID-19, mas, em face das investigações nos Estados Unidos [104], esse efeito provavelmente tem um papel importante no caso brasileiro e merece atenção em pesquisas futuras.

Para testar a maior vantagem urbana para o tratamento da COVID-19 durante o espalhamento inicial da doença, investigamos as relações de escala entre o número de leitos de unidade de tratamento intensivo (UTI) e a população da cidade. Dado que pacientes em estado crítico frequentemente necessitam de ventilação mecânica [105, 106], o número de leitos de UTI provaram ser cruciais no tratamento da COVID-19. A Figura 2.3 mostra a relação alométrica entre o número de leitos de UTI dos sistemas de saúde público e privado (Y_{icu}, no período de abril de 2020) e a população urbana (P). Uma relação superlinear entre essas duas variáveis emerge com expoente de escala $\beta_{icu} \approx 1.16$. A escala superlinear dos leitos de UTI indica que cidades grandes no Brasil são melhores estruturadas para lidar com pacientes em situação crítica. Esse fato pode explicar parcialmente a redução de mortes *per capita* com tamanho da cidade durante os três ou quatro meses iniciais desde as primeiras duas mortes diárias. É importante notar que o Sistema Único de Saúde (SUS) é descentralizado e composto por "regiões de saúde", grupos contíguos de cidades geralmente formados por uma cidade grande e suas cidades vizinhas [107]. Cidades dentro da mesma região de saúde podem repartir serviços médicos, o que pode, em contrapartida, explicar parcialmente a redução de vantagens estruturais de grandes centros urbanos em períodos posteriores da pandemia.

Além disso, investigamos como a distribuição da demografia etária muda conforme o tamanho populacional das cidades. Estimativas demostraram que a taxa de fatalidade da COVID-19 é substancialmente maior em pessoas com mais de 60 anos (0.32% para pessoas com menos de 60 anos *versus* 6.5% para pessoas com mais de 60 anos [108]). Em vista disso, a demografia etária das cidades representa um fator importante para o número de mortes causadas por COVID-19. As Figuras 2.3B e 2.3C mostram como o número de pessoas mais velhas (P_{hr}, a população de alto risco) e de pessoas mais novas (P_{lr}, a população de baixo risco) do que 60 anos muda com o total da população (P). Notamos que a população de alto risco aumenta sublinearmente com o tamanho da cidade com um expoente $\beta_{hr} \approx 0.91$, enquanto a população de baixo risco escala linearmente ($\beta_{lr} \approx 1$) com o tamanho da cidade. Esse resultado mostra que cidades grandes tem menor prevalência de adultos com mais de 60 anos, de tal forma que um aumento de 1% na população urbana é associada com um aumento de 0.91% na população de alto risco. Num exemplo mais concreto, esperamos que uma cidade com um milhão de pessoas tenha proporcionalmente $\approx 19\%$ menos adultos com mais de 60 anos quando comparado com uma cidade de 100 mil habitantes. Logo, a menor

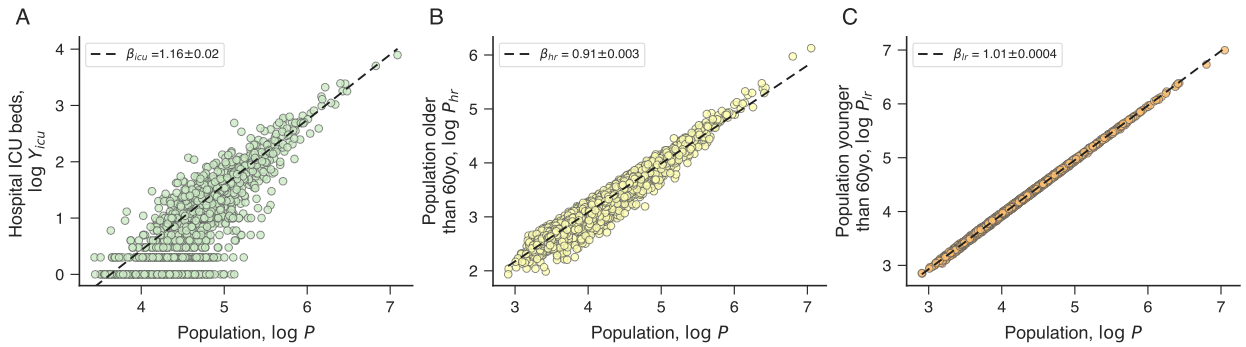

Figura 2.3: Escala urbana dos leitos de UTI e das populações de alto e de baixo risco. **(A)** Relação entre o número de leitos de UTI (Y_{icu}) e a população das cidades (P) em escala logarítmica. O número de leitos de UTI escala superlinearmente com o tamanho da cidade ($\beta_{icu} = 1.16 \pm 0.02$), indicando uma vantagem urbana na cobertura dos serviços de saúde. **(B)** Relação entre a população de alto risco (P_{hr}, definida como adultos com idade superior a 60 anos) e a população das cidades (P) em escala logarítmica. A população de alto risco escala sublinearmente ($\beta_{hr} = 0.910 \pm 0.003$), indicando que cidades grandes tendem a ter menores frações de idosos do que cidades pequenas. **(C)** Relação entre a população de baixo risco (P_{lr}, definida como adultos com idade inferior a 60 anos) e a população das cidades (P) em escala logarítmica. A população de baixo risco escala quase que linearmente ($\beta_{hr} = 1.0100 \pm 0.0004$) com o tamanho da cidade. O comportamento conjunto dessas três quantidades explica parcialmente a diminuição inicial do número de mortes *per capita* com a população ($\beta_d < 1$ para $t_d \lesssim 100$ dias). Veja a Figura A.30 para as relações de escala envolvendo as quantidades *per capita*.

prevalência de idosos em grandes centros urbanos pode também parcialmente explicar a redução inicial do número de mortes *per capita* com o aumento da população da cidade.

Além de estudar a escala urbana de casos e mortes por COVID-19, também investigamos associações entre as taxas de crescimento de casos e mortes e a população das cidades (Figuras A.31-A.37). Como mencionado anteriormente, o trabalho de Stier, Berman e Bettencourt [95] mostra que as taxas de crescimento iniciais de casos de COVID-19 em áreas metropolitanas dos Estados Unidos escala como uma função lei de potência da população com expoente entre 0.11 e 0.20. Utilizando nosso conjunto de dados, estimamos a taxa de crescimento dos casos (r_c) e das mortes (r_d) para cidades brasileiras. Em concordância com o caso dos Estados Unidos, nossos resultados indicam que casos de COVID-19 inicialmente crescem mais rapidamente em cidades grandes (Figura A.38) de tal forma que $r_c \sim P^{\beta_{r_c}}$ com β_{r_c} entre ≈ 0.1 e ≈ 0.3 durante os três primeiros meses após os dois primeiros casos diários ($t_c \lesssim 90$, Figura A.38). Além do mais, encontramos um comportamento similar para a taxa de crescimento do número de mortes r_d. A relação lei de potência $r_d \sim P^{\beta_{r_d}}$ é uma descrição razoável para os dados empíricos com expoente de escala β_{r_d} entre ≈ 0.1 e ≈ 0.5 nos primeiros três meses após as duas primeiras mortes diárias ($t_d \lesssim 90$, Figura A.38).

A taxa de crescimento ilustra uma etapa do processo de espalhamento da COVID-19 e sua associação com tamanho da cidade pode mudar durante a evolução a longo prazo da

pandemia. Essas mudanças nas taxas de crescimento podem refletir as diferentes ações adotadas por cada cidade a fim de encarar a pandemia da COVID-19. No caso do espalhamento nos Estados Unidos, Heroy [109] reporta que cidades grandes parecem entrar num regime de espalhamento exponencial mais cedo do que as pequenas cidades. Para melhor investigar essas possibilidades em nosso conjunto de dados, estimamos a relação média entre a taxa de crescimento do número de casos (r_c) e mortes (r_d) e o ranque s da cidade ($s = 1$ representa a maior cidade dos dados, $s = 2$ a segunda maior cidade e assim por diante) em diferentes períodos. A Figura 2.4A mostra os resultados para as taxas de crescimento no número de casos (r_c). Em acordo com a associação por lei de potência entre r_c e a população da cidade P (Figuras A.31-A.37), notamos que os valores inferiores do ranque s da cidade estão associados com maiores taxas de crescimento r_c nos dias posteriores aos primeiros dois casos diários. Contudo, com o passar do tempo, as taxas de crescimento de casos começam a diminuir em cidades grandes (valores de ranque pequeno) e a aumentar em cidades pequenas (valores de ranque grandes). Esse resultado parece concordar com as recentes descobertas de Heroy [109] pois existe um atraso na emergência de altas taxas de crescimento de casos para as cidades pequenas. A Figura 2.4B mostra a mesma análise para a taxa de crescimento no número de mortes por COVID-19. Ao mesmo tempo que observamos uma redução em r_d para cidades grandes e um aumento para cidades pequenas, as diferenças em r_d são menos pronunciadas em comparação com r_c. Esses resultados também emergem no estudo de expoentes de escala associados com a taxa de crescimento de casos (β_{r_c}) e de mortes (β_{r_d}). Os resultados da Figura A.38 mostram que esses expoentes começam a diminuir ao redor de $t_c \approx t_d \approx 100$ dias e tornam-se negativos em nossas últimas estimativas. É interessante lembrar que o tempo t_c (ou t_d) é medido em dias desde os primeiros dois casos diários (ou primeiras duas mortes diárias) para cada cidade; posto isto, os resultados da Figura 2.4 não refletem atrasos na emergência do primeiro caso em cada cidade.

2.3 Conclusões

Estudamos as relações de escala para o número de casos e mortes por COVID-19 em cidades brasileiras. Similarmente ao que acontece para outras doenças, descobrimos que o número de casos e mortes são leis de potência relacionadas com a população das cidades. Durante os primeiros três a quatro meses desde os primeiros dois casos diários ou primeiras duas mortes diárias, encontramos uma associação sublinear entre casos e mortes por COVID-19, o que significa que os número de casos e mortes *per capita* tendem a decrescer com a população nesse estágio inicial da pandemia. Acreditamos que esse comportamento pode ser parcialmente explicado por uma "crescente vantagem urbana" em que cidades grandes tem proporcionalmente mais leitos de UTI do que as pequenas. Além disso, mudanças na demografia etária com o tamanho da cidade mostram que cidades grandes têm proporcio-

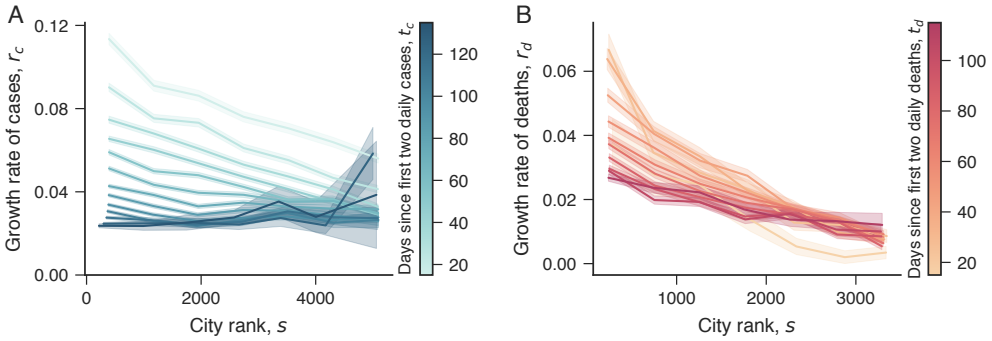

Figura 2.4: Associação entre as taxas de crescimento e o tamanho das cidades. **(A)** Relação entre a taxa de crescimento nos casos de COVID-19 (r_c) e o ranque s das cidades. As diferentes curvas mostram os valores médios de r_c *versus* s para diferentes números de dias desde os primeiros dois casos diários de COVID-19 (t_c, como indicado pela escala de cor). **(B)** Relação entre a taxa de crescimento nas mortes por COVID-19 (r_c) e o ranque s das cidades. As diferentes curvas mostram os valores médios de r_d *versus* s para diferentes números de dias desde as primeiras duas mortes diárias por COVID-19 (t_d, como indicado pela escala de cor).

nalmente menos idosos, isto é, o grupo de risco mais suscetível a atingir um estado crítico e falecer por COVID-19. Isso pode parcialmente explicar a redução inicial das fatalidades *per capita* com a população da cidade. Ademais, argumentamos que estratégias e politicas contra a COVID-19 de cidades grandes e pequenas podem ser diferentes, levando a níveis discrepantes de eficiência na contenção da pandemia.

Entretanto, encontramos que a "vantagem urbana" desaparece em períodos posteriores da pandemia de tal maneira que a associação entre casos e mortes por COVID-19 com a população torna-se superlinear nas nossas últimas estimativas desde os primeiros dois casos diários ou desde as primeiras duas mortes diárias. A persistência desse padrão indica que cidades grandes serão proporcionalmente mais afetadas ao fim da pandemia de COVID-19. Esse resultado está de acordo com investigações realizadas para outras doenças infecciosas [91,93] e provavelmente reflete a existência de um maior grau de interação entre as pessoas em cidades grandes [85,94]. Como medidas de distanciamento social eram as únicas medidas de mitigação da COVID-19 na época em que os dados foram extraídos, os nossos resultados sugerem que cidades grandes poderiam precisar de graus mais severos de políticas de distanciamento social.

Em linha com os resultados para as áreas metropolitanas dos Estados Unidos [95], encontramos que cidades grande geralmente apresentam taxas de crescimento no número de casos maiores durante o espalhamento inicial da COVID-19. Não obstante, nossos resultados também mostram que essas taxas de crescimento tendem a decrescer em cidades grandes e a crescer nas cidades pequenas em fases posteriores da pandemia. Esse comportamento sugere a existência de um atraso na emergência de altas taxas de crescimento entre cidades grandes e pequenas. Um comportamento similar foi encontrado nos Estados Unidos [109],

onde cidades grandes parecem entrar num regime de crescimento exponencial mais cedo do que as cidades pequenas. A existência desse atraso sugere que o ritmo lento de espalhamento inicial da COVID-19 em cidades pequenas é um comportamento transiente.

Em conjunto com as descobertas de Stier-Berman-Bettencourt [95] e Heroy [109] para os Estados Unidos, bem como aqueles de Cardoso e Gonçalves [96] para os Estados Unidos, Brasil e Alemanha, nossos resultados sugerem que políticas de distanciamento social e outras ações contra a pandemia devem levar em conta os efeitos não lineares do tamanho de cidades no espalhamento da COVID-19.

CAPÍTULO 3

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

A crescente disponibilidade de informação e o desenvolvimento de economias baseadas em conhecimento têm instigado esforços interdisciplinares em direção de um melhor entendimento quantitativo da empreitada científica: uma ciência da ciência [110,111]. 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 práticas e políticas de sucesso, da escolha de prioridades científicas nacionais até a seleção de projetos de pesquisa e a contratação de professores. Atualmente, o progresso científico é fortemente dependente dos processos de avaliação, pois esses regulam o fluxo de ideias viabilizando projetos de pesquisa por meio da alocação de recursos financeiros [111–114]. Nesse contexto, a revisão por pares (do inglês, *peer review*) é considerada a abordagem padrão para avaliar performance acadêmica [115]. Entretanto, esse processo é laborioso e tem várias desvantagens desde vieses e falta de consistência a fraudes [115–118]. Além disso, o número crescente de publicações científicas [119] e o aumento da massa trabalhadora acadêmica [120] acarretam mais limitações para o método de revisão por pares [114]. Como consequência direta dessas dificuldades, houve um crescimento no uso de índices bibliométricos (ou bibliometrias) para a classificação da performance acadêmica [114,121], especialmente depois dos anos 2000 [122].

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 [123,124]. Para além da questão sobre a viabilidade da avaliação,

o uso de bibliometrias impõe uma enorme pressão aos cientistas (particularmente aos mais jovens [125]) para publicar em grandes quantidades, em jornais de prestígio[1] e para desenvolver pesquisas altamente citadas [114, 126, 127]. Pelos fatores enumerados anteriormente, o uso de bibliometrias tem sido alvo de muitas críticas recentes [115, 128–130]. É nesse contexto controverso que a produtividade e as medidas de impacto são frequente e amplamente utilizadas para quantificar a 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. Independentemente da métrica utilizada, a avaliação da pesquisa por meio de bibliometrias tem levantado um debate sobre "qualidade *versus* quantidade" desde sua concepção [131–146] e ainda não existe consenso sobre a natureza da associação entre essas duas variáveis. Por exemplo, enquanto Larivière e Costas [141] encontraram uma associação positiva entre a produtividade e o número de artigos altamente citados, Bornmann e Tekles [145] 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 altos 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*". 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".

Aqui, 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 controversa [128, 147]. No entanto, essa abordagem ainda permanece difundida e largamente utilizada [148], especialmente no Brasil em que várias universidades usam o prestígio do jornal (ou indicadores derivados) para diversas tarefas, desde a contratação de professores [149] até a concessão de recursos financeiros [150]. Trabalhos recentes demostram que métricas em nível de jornal carregam informação sobre a performance acadêmica [151–155] e são correlacionadas com citações, indicando que essas duas métricas podem ser consideradas como substitutas parciais. Apesar de não termos uma resposta definitiva se métricas em nível de jornal (ou até mesmo as citações) são apropriadas para avaliação de pesquisas, fato é que essas métricas ainda permanecem importantes para a comunidade acadêmica. Dessa forma, novas investigações podem trazer à luz aspectos relevantes para melhorar o processo de avaliação.

Nossa pesquisa examina padrões na associação entre produtividade e métricas de impacto

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

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. Além disso, a medida padronizada referente ao prestígio médio dos jornais também corrigiu vieses relacionados à sua escala. 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 brasileiros 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 de carreira de produtividade e prestígio de jornal são específicos para cada disciplina. O fato em comum é que a produtividade média cresce com o tempo para todas as disciplinas. Ao estudar aspectos da produtividade e do impacto de jornal relacionados ao estágio da carreira e às disciplinas acadêmicas, acreditamos que nosso trabalho pode contribuir significantemente para um processo de avaliação de pesquisa mais compreensivo e mais justo.

3.1 Métodos

3.1.1 Dados

O conjunto de dados utilizado em nosso estudo foi obtido por meio da Plataforma Lattes[2]. Esse sistema de informação tem sido mantido pelo governo brasileiro desde 1999, abrigando o currículo oficial dos acadêmicos brasileiros. O currículo Lattes é amplamente utilizado para avaliação tanto individual quanto institucional. Por isso, 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) e obtivemos seu registro completo de publicações (um total de 1.121.652 artigos científicos). 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 conclusão do doutorado, reduzindo o número para 14.146 pesquisadores. Para completar essa base de dados, incluímos informações faltantes sobre o ano de publicação e 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 entre 1997 e 2015 nos relatórios de citação de jornais da Clarivate (*Clarivate's Journal Citation*

[2] http://lattes.cnpq.br/

Reports). Combinamos ambos 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 dos jornais). Finalmente, agrupamos as séries temporais por disciplina e selecionamos as 14 disciplinas com pelo menos 50 pesquisadores com artigos publicados em cada ano entre 1997 e 2015. Esse processo nos leva ao nosso conjunto final de dados com 6.028 pesquisadores de 14 disciplinas e 312.881 artigos (Figura A.39). 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. 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.40).

Enquanto o JIF de um jornal é simplesmente definido como o número de citações recebidas por artigos dos dois anos anteriores dividido pelo número total de artigos publicados nesses mesmos anos, o SJR é uma medida baseada em rede [156] (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 $\approx$0.85 (Figura A.41). 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 predominância de concessões de bolsas produtividade para pesquisadores dessas disciplinas.

3.1.2 Inflação e medidas robustas de padronização

O volume de produção científica tem crescido com o tempo em nível global e individual [157, 158]. Esse crescimento acarreta um efeito de inflação temporal na produtividade e no impacto médio dos jornais, impossibilitando a comparação de observações de períodos distintos (Figuras A.42 e A.43). As disciplinas também têm volumes de publicação e dinâmicas de citações distintos [159, 160]. Como consequência, também existe dificuldade em comparar a produtividade e o impacto médio dos jornais entre diferentes disciplinas. Além disso, o impacto médio dos jornais sofre um efeito de escala. Especificamente, o valor médio diminui sua variabilidade com o aumento da produtividade. Esse efeito foi anteriormente verificado no estudo sobre os fatores de impacto de jornais com diferentes números totais de publicação [161, 162] e, como argumenta Antonoyiannakis [161, 162], é uma consequência direta do 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 dos jornais. Considere que $p_j^k(y)$ e $i_j^k(y)$ representem, respectivamente, o número de artigos e o prestígio médio dos jornais de publicações de um

pesquisador j da disciplina k no ano y. Calculamos os z-scores de produtividade como

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

em que $\mathbb{E}[p_j^k(y)]$ e $\mathbb{S}[p_j^k(y)]$ 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

$$I_j^k(y) = \frac{i_j^k(y) - \mathbb{E}[i_{\text{rnd}}^k(y, p_j^k(y))]}{\mathbb{S}[i_{\text{rnd}}^k(y, p_j^k(y))]},$$

em que $i_{\text{rnd}}^k(y, p)$ é o impacto médio dos jornais de uma amostra aleatória de p publicações da disciplina k no ano y e $\mathbb{E}[i_{\text{rnd}}^k(y, p)]$ e $\mathbb{S}[i_{\text{rnd}}^k(y, p)]$ são, respectivamente, a média e o desvio padrão de $i_{\text{rnd}}^k(y, p)$ estimadas a partir de 1.000 realizações independentes. Esta última definição é uma adaptação do índice Φ proposto por Antonoyiannakis [161, 162] para ranquear jornais de diferentes tamanhos; aqui, levamos em consideração o fato de que pesquisadores pouco prolíficos tem alta variabilidade nos valores médios de prestígio de jornal, enquanto pesquisadores muito prolíficos mostram variabilidade significativamente menor (Figura A.44).

Aqui, usamos os estimadores robustos de Huber (veja a Seção 1.6) para média (localização) e desvio padrão (escala) em vez dos estimadores usuais [61] devido à existência de valores extremos ("*outliers*") para $p_j^k(y)$ e $i_j^k(y)$ (Figura A.45). Ou seja, $\mathbb{E}[\dots]$ e $\mathbb{S}[\dots]$ representam, respectivamente, os estimadores de localização e escala de Huber (conforme implementado no pacote de Python *statsmodels* [163]).

3.1.3 Regressões logísticas

A fim de quantificar o efeito de performar como *outlier* em produtividade na chance de ser *outlier* em impacto de jornal (perfeccionista), usamos o seguinte modelo logístico (veja a Seção 1.2)

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

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* [163]) aos nossos dados considerando todos os pesquisadores que foram *outliers* em impacto de jornal ou produtividade pelo menos em algum ponto de suas carreiras. Também ajustamos o mesmo modelo agrupando pesquisadores por disciplina. A Figura 3.1C mostra $\Pi_{\text{perfectionist}}$ como uma função de Y_P para cada disciplina do conjunto de dados JIF e quando considerando todas as disciplinas

agregadas (a inserção discrimina os valores de α_1). Similarmente, a Figura A.46C mostra os resultados correspondentes ao conjunto de dados SJR.

Utilizamos também um modelo logístico para estimar a probabilidade de ser perfeccionista como uma função do 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 A.47 mostra $\Pi_{\text{perfectionist}}$ como função de L para ambos conjuntos de dados. 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 (Figura A.47A) e $\theta_0 = 1.921 \pm 0.108$ e $\theta_1 = -0.054 \pm 0.005$ para o conjunto de dados SJR (Figura A.47B).

3.1.4 Modelo hierárquico bayesiano

Usamos um modelo hierárquico bayesiano (veja a Seção 1.3) para estimar o efeito da produtividade no impacto médio dos jornais de pesquisadores não *outliers*. 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 (aqui 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

$$I_j \sim \mathcal{N}(c_j + \beta_j P_j, \varepsilon)\,, \tag{3.1}$$

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 [164] a fim de não enviesar a estimativa da *posteriori*, isto é, consideramos

$$\begin{aligned} \varepsilon &\sim \mathcal{U}(0, 10^2) \\ \mu_c &\sim \mathcal{N}(0, 10^5) \\ \mu_P &\sim \mathcal{N}(0, 10^5) \\ \sigma_c &\sim \text{Inv-}\Gamma(10^{-3}, 1) \\ \sigma_P &\sim \text{Inv-}\Gamma(10^{-3}, 1) \end{aligned} \quad , \tag{3.2}$$

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 A.48 mostra uma representação gráfica desse modelo.

Também utilizamos uma versão generalizada do modelo definido na Eq. (3.1) em que o ano da carreira A_j também é considerado como uma variável independente no modelo hierárquico, resultando em

$$I_j \sim \mathcal{N}(c_j + \beta_j P_j + \gamma_j A_j, \varepsilon)\,, \tag{3.3}$$

em que γ_j é a inclinação da associação linear entre o ano da carreira e o prestígio de jornal. Assume-se que esse coeficiente linear 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 Eq. (3.3) com as mesmas distribuições não informativas a *priori* definidas na Eq. (3.2) e usamos

$$\begin{aligned} \mu_A &\sim \mathcal{N}(0, 10^5) \\ \sigma_A &\sim \text{Inv-}\Gamma(10^{-3}, 1) \end{aligned} \tag{3.4}$$

como as distribuições não informativas a *priori* para os parâmetros adicionais relacionados aos efeitos da idade da carreira. A Figura A.49 mostra a representação gráfica desse modelo generalizado que leva em consideração possíveis efeitos de confusão da idade da carreira na associação entre impacto médio dos jornais e produtividade.

Implementamos esses dois modelos usando a abordagem do pacote *PyMC3* via método de gradiente com amostrador de Monte Carlo Hamiltoniano NUTS (*No-U-Turn-Sampler*, veja as seções 1.5 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 a estatística de convergência de Gelman-Rubin (R chapéu, veja a seção 1.5) para todas as análises de regressão e os resultados foram todos pertos de um, ou seja, uma indicação da convergência do método de

amostragem.

3.2 Resultados

Plano do prestígio de jornal *versus* produtividade

Para investigar a associação entre produtividade e prestígio de jornal, conforme já mencionado, obtivemos os currículos acadêmicos de 6.028 pesquisadores brasileiros com Bolsa Produtividade em Pesquisa do CNPq de 14 disciplinas de acordo com seu *status* em maio de 2017. A bolsa produtividade tem sido concedida em reconhecimento à eminente produção científica de determinados pesquisadores desde os anos 70 pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Os bolsistas do CNPq são considerados a elite dos cientistas brasileiros. Além disso, obtivemos o fator de impacto dos jornais (JIF, *Journal Impact Factor*) entre 1997 e 2015 do *Journal Citation Reports* da *Clarivate*. Em seguida, combinamos os dois conjuntos de dados para atribuir os valores de JIF variando no tempo para 312.881 artigos publicados pelos bolsistas do CNPq entre 1997 e 2015. Consideramos o número de artigos publicados por ano como o indicador de produtividade e o JIF médio como indicador de prestígio de jornal. Para efeitos de comparação e robusteza, também realizamos uma análise considerando o ranque de jornais SCImago (SJR, *SCImago Journal Rank*) da *Scopus* como indicador de prestígio de jornal. Apesar da diferença substancial na definição das duas medidas, ambos indicadores de prestígio de jornal são fortemente correlacionados (Figura A.41) e produzem resultados muito similares. Optamos por apresentar os resultados para o JIF no texto principal e nos referimos ao Apêndice A de Figuras Adicionais para comparações com o SJR.

Começamos nossa investigação percebendo que o número de artigos e citações têm crescido ao longo do tempo [157,158]. Esse efeito produz inflação na produtividade e nas medidas de impacto de jornal que precisa ser considerada para uma comparação justa entre diferentes anos de publicação. Nossos resultados indicam que a produtividade média dos bolsistas CNPq tem crescido numa taxa de ≈ 1.57 artigos/ano por década. Similarmente, o JIF médio dessas publicações tem crescido ≈ 0.72 unidades por década (Figuras A.42 e A.43). O fenômeno de inflação é diferente entre as disciplinas: 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. Assim, não utilizamos os números brutos de produtividade devido ao efeito de inflação e às diferenças nos padrões de publicação entre disciplinas. Em vez disso, utilizamos medidas padronizadas robustas (z-scores) relativas à disciplina e ao ano de publicação. Por sua vez, as medidas padronizadas robustas para o prestígio médio dos jornais também são relativas ao nível de produtividade dos pesquisadores além da disciplina

e do ano de publicação. Essa normalização adicional leva em consideração o fato de que, quanto mais produtivo é um pesquisador em certo ano, mais estreito é o intervalo de variação de seu prestígio médio dos jornais. Um efeito de tamanho similar foi observado por Antonoyiannakis [161, 162] ao comparar o fator de impacto de jornais de diferentes tamanhos. Aqui, adaptamos o método de reescala para ranquear jornais de Antonoyiannakis [161, 162] a fim de considerar os efeito de escala para o prestígio médio dos jornais (veja a discussão da Seção 3.1.2).

A Figura 3.1A mostra um diagrama de dispersão do prestígio médio dos jornais *versus* produtividade para todos anos de carreira de pesquisadores em nosso conjunto de dados (veja a Figura A.46 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 médio dos jornais representa a performance de 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 médio dos jornais (I). O setor $IP{+}{+}$ contém anos de carreira em que pesquisadores foram *outliers* simultaneamente em produtividade e prestígio de jornal ($I > 3.5$ and $P > 3.5$). Similarmente, os setores $I{+}{+}$ e $P{+}{+}$ indicam anos de carreira *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 $P < 0$).

Pesquisadores *outliers* e não *outliers*

Uma das características mais surpreendentes do plano mostrado na Figura 3.1A 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. Esses anos *outliers* da carreira são relativamente raros e representam apenas 7,7% dos 76.454 anos de carreira totais (Figura A.50A). 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 mais raros ainda e correspondem a apenas 7% do total de anos *outliers*. Resultados similares foram obtidos para o conjunto de dados SJR (Figura A.50B).

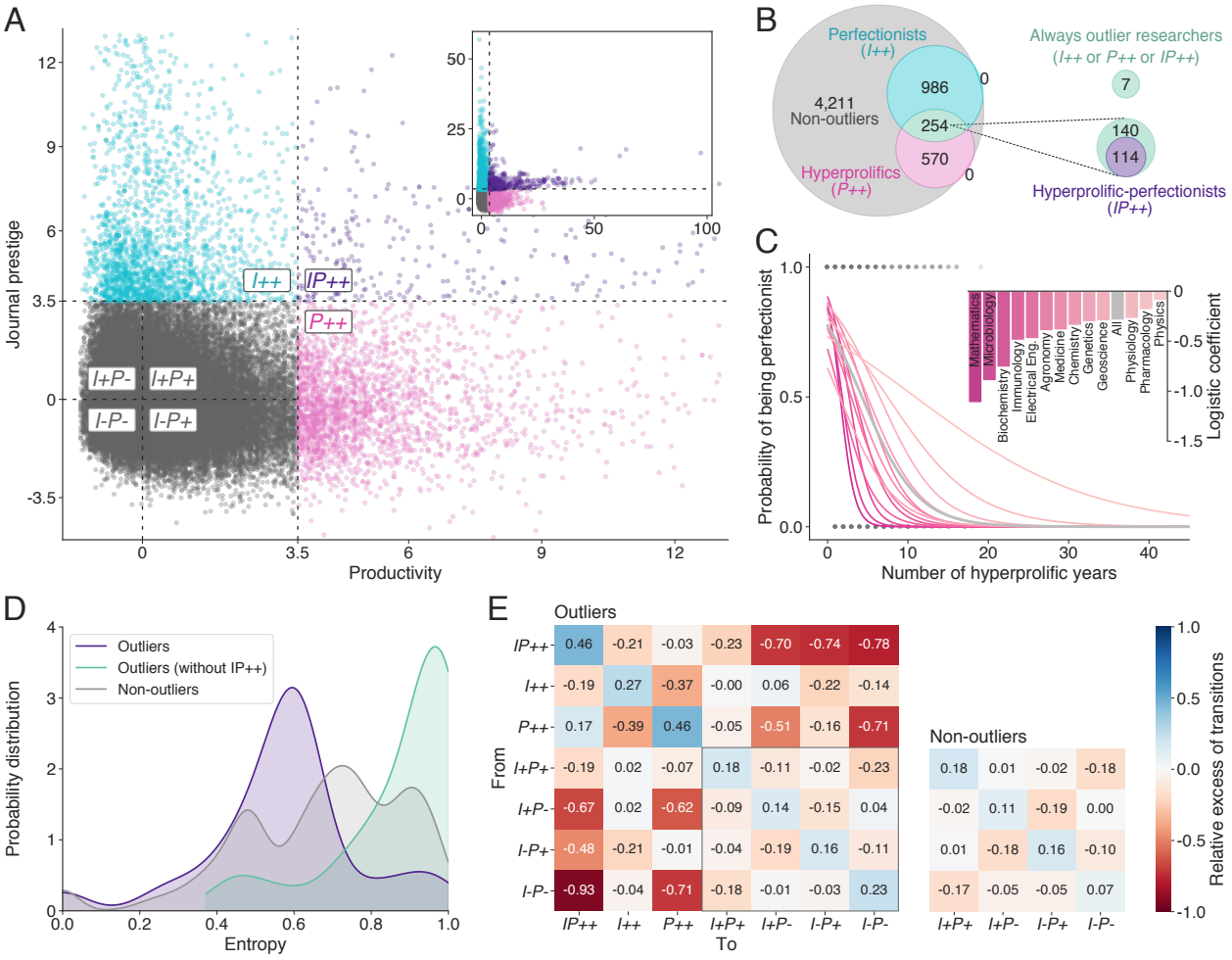

Figura 3.1: Prestígio de jornal *versus* produtividade. (A) Relação entre impacto médio dos jornais e produtividade em unidades padronizadas (a inserção mostra o intervalo completo do plano). Os marcadores representam anos da carreira de pesquisadores de 14 disciplinas em nosso estudo. **(B)** Diagrama de Venn mostrando o conjunto de relações entre as quatro categorias de pesquisadores. **(C)** 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 agregado para todas as disciplinas. A disciplina de engenharia dos materiais (omitida nesse painel) é a única disciplina que não apresenta uma associação significativa. **(D)** Distribuição de probabilidade da entropia normalizada de Shannon associada com a ocupação dos setores do plano para as carreiras individuais dos pesquisadores. A curva em roxo mostra os resultados da 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 para pesquisadores *outliers* (esquerda) e não *outliers* (direita). Cada célula representa o excesso relativo de transições entre dois setores comparado com o modelo nulo, que corresponde às versões embaralhadas das carreiras dos pesquisadores para 10.000 realizações.

Anos *outliers* também representam apenas uma pequena fração das carreiras dos pesquisadores da nossa base de dados (Figura A.51A). 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 3.1B 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 em nosso conjunto de dados nos três setores *outliers*. Resultados similares foram encontrados para o conjunto de dados SJR (Figura A.51B).

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 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 recentemente por Bornmann e Tekles [145] para a associação entre produtividade e o número de artigos no *top*-1% mais citados. Assim, nosso resultado indica que é extremamente difícil publicar frequentemente em jornais de alto prestígio e, concomitantemente, manter altíssimos níveis de produtividade. De modo intrigante, observamos que anos da carreira extremamente hiperprolíficos ($P > 27.7$) estão todos no setor $IP{+}{+}$. Esse resultado mostra que, a despeito de muito raros, existem dezesseis pesquisadores capazes de manter performances extremas em produtividade e prestígio de jornal.

Para reforçar esse resultado, usamos uma regressão logística para estimar o efeito de anos hiperprolíficos na probabilidade de performar como um pesquisador perfeccionista. A Figura 3.1C mostra a probabilidade de ser um pesquisador perfeccionista como uma função do número de anos hiperprolíficos e os coeficientes logísticos ao considerar todas as disciplinas conjuntamente e separadamente. A disciplina de engenharia dos materias não mostra uma associação significativa (p-valor > 0.05) e foi omitida da Figura 3.1C. 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 previnem a existência de pesquisadores perfeccionistas na matemática, existe uma proba-

bilidade 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.46C), reafirmando que existe uma associação negativa entre essas dois comportamentos.

O grupo de 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, aquele 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 em cada setor e calculamos a entropia normalizada de Shannon. Valores de entropia perto de um representam comportamento alternante, enquanto valores perto de zero indicam representam preferência por dado setor *outlier*. A Figura 3.1D 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 mais subpovoado), a distribuição de entropia desloca na direção de valores mais elevados com pico ao redor 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 mostrado na Figura 3.1D (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.46D).

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 e $h = \{IP{+}{+}, I{+}{+}, P{+}{+}, I{\pm}P{\pm}, I{\pm}P{\mp}\}$). 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 e $h = \{IP{+}{+}, I{+}{+}, P{+}{+}, \cancel{I}\cancel{P}{\pm}, \cancel{I}\cancel{P}{\mp}\}$). A partir desse processo, estimamos o excesso relativo para todas as transições possíveis a

partir de

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

A Figura 3.1E mostra as matrizes de transição ao agrupar os pesquisadores em categorias *outliers* em não *outliers*. A primeira característica que observamos é que as matrizes são aproximadamente simétricas, indicando que a maioria das transições não tem direção preferencial. Além disso, os elementos diagonais positivos estão entre os maiores valores absolutos, ou seja, permanecer no mesmo setor é uma 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 na matriz não *outlier* 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 oposta, 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 frequentemente do que 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.46E). 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 ($I{+}{+}$), a diferença não é tão substancial quanto para o conjunto de dados JIF. Todas as outras transições apresentam aproximadamente o mesmo comportamento. Para testar a robusteza, constatamos que os resultados para o conjunto de dados SJR são consistentes mesmo quando consideramos apenas as disciplinas presentes na base de dados JIF (Figura A.52).

Efeitos da idade da carreira

Investigamos os efeitos da idade da carreira acadêmica no prestígio médio dos jornais e na produtividade. Com esse objetivo, consideramos o ano após a obtenção do título de doutor como o primeiro ano da carreira dos pesquisadores. Em seguida, calculamos os valores médios da produtividade e do prestígio médio dos jornais em janelas móveis de 5 anos a partir do agrupamento das carreiras de cada disciplina. A Figura 3.2 mostra os valores médios como função do ano da carreira dos pesquisadores. Observamos uma tendência crescente na produtividade média ao longo da carreira para todas as disciplinas (Figura A.53A), seguida por um platô ou pequeno decréscimo no período final da carreira. Para o prestígio médio dos jornais, 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 (Figura A.53B); não obstante, determinadas disciplinas mostram padrões mais complexos. As Figuras A.54 e A.55 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 na próxima seção.

Para continuar caracterizando os efeitos da idade 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 médio dos jornais *versus* produtividade como uma função da idade da carreira. A Figura 3.3 mostra essas frações para todas as disciplinas em nossa investigação. Nessa representação matricial, colunas indicam intervalos da carreira, linhas indicam diferentes planos do setor e os códigos de cor 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 maior do que o número de anos no conjunto de dados JIF (19 anos).

A Figura 3.3 indica que as tendências de ocupação no plano prestígio de jornal *versus* produtividade variam entre as disciplinas (veja a Figura A.56 para os resultados baseados no 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 produ-

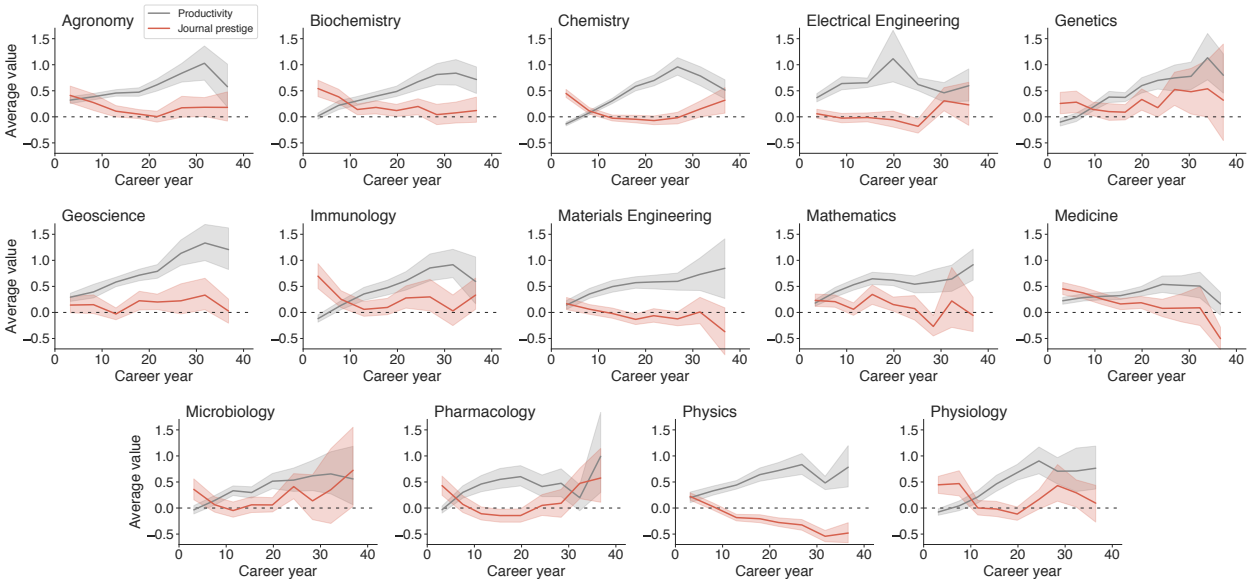

Figura 3.2: Valores médios da produtividade e do impacto médio dos jornais ao longo da carreira dos pesquisadores para diferentes disciplinas. Essas visualizações mostram os valores médios da produtividade (curva em cinza) e do prestígio médio dos jornais (curva em vermelho) calculados a partir de médias 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*. A produtividade média aumenta com a progressão da carreira para todas as disciplinas (Figura A.53A) e mostra um platô ou pequeno decréscimo em estágios posteriores da carreira para a maioria das disciplinas. Apesar de algumas disciplinas apresentarem padrões mais complexos, o valor médio do prestígio médio dos jornais mostra uma tendência sutil decrescente e é usualmente maior nos estágios iniciais da carreira para a maioria das disciplinas (Figura A.53B).

tividade ($I{+}P{+}$ e $I{-}P{+}$) em estágios posteriores da carreira dos 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 para todas as disciplinas em estágios posteriores. 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 3.2.

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 médio dos jornais nos primeiros anos de carreira para maioria das disciplinas (Figura 3.2). 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*

Figura 3.3: Tendências de ocupação do plano prestígio de jornal *versus* produtividade ao longo das carreiras dos pesquisadores. Os painéis mostram a fração dos anos das carreiras em cada setor não *outlier* e nos setores *outliers* $I{+}{+}$ e $P{+}{+}$ como uma função da idade 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. Os setores de baixa produtividade são mais povoados durante os anos iniciais da carreira. Além disso, há uma tendência de mudança para setores de alta produtividade em estágios posteriores da carreira para a maioria das disciplinas. Apenas intervalos de 5 anos com pelo menos 20 pesquisadores são mostrados nessas visualizações.

nessa categoria mais frequentemente. De fato, 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 (Figura A.47A). 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 apenas inclui pesquisadores pertencentes à elite científica brasileira. Os resultados para o conjunto de dados SJR corroboram esse resultado (Figura A.47B) e indicam tendências muito similares não apenas para discipli-

nas presentes em ambos os conjuntos de dados mas também para disciplinas exclusivas do conjunto de dados SJR.

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 70% dos pesquisadores em nosso conjunto de dados e podem exibir comportamentos heterogêneos. Esta última característica limita a emergência de uma clara associação agregada no nível da disciplina. Para levar em conta os padrões individuais diversos, selecionamos apenas anos produtivos de pesquisadores não *outliers* com carreiras maiores do que cinco anos (Tabelas B.1 e B.2) a fim de aplicar um modelo bayesiano hierárquico para examinar a associação entre produtividade e prestígio de jornal. Assumimos uma relação linear entre essas duas quantidades em que a distribuição do coeficiente linear relacionado a cada pesquisador tem média retirada de outra distribuição com valor médio μ_P (veja a seção 3.1.4).

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 médio dos jornais 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 médio dos jornais para maioria dos pesquisadores.

A Figura 3.4A mostra que a distribuição de μ_P (curvas preenchidas coloridas) variam significativamente entre as disciplinas. Todas as disciplinas exceto matemática têm distribuições 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 $\approx$0.242 unidades de impacto médio dos jornais 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 negativas).

Os resultados ilustrados nas Figuras 3.2 e 3.3 demonstraram que a idade da carreira afeta os valores médios da produtividade e do prestígio médio dos jornais quando se agregam pesquisadores por suas respectivas disciplinas. Sendo assim, podemos esperar que a

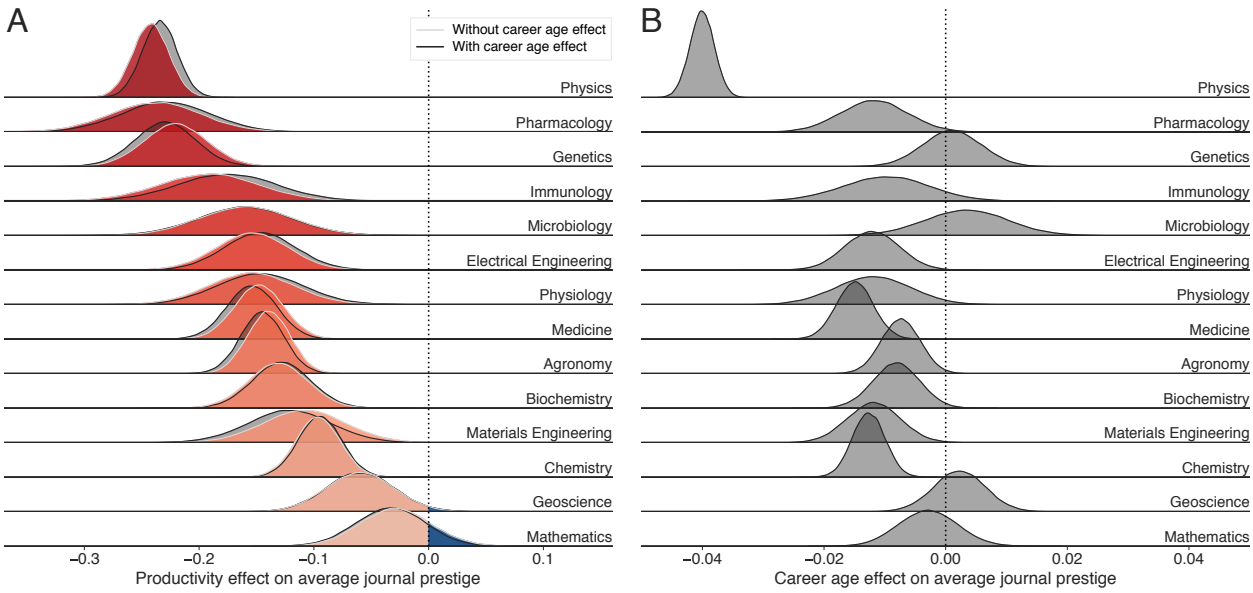

Figura 3.4: Efeito da produtividade no prestígio de jornal para pesquisadores não *outliers*. **(A)** Distribuições de probabilidade a *posteriori* do valor médio do coeficiente linear (μ_P) ao considerar a associação entre produtividade e impacto de jornal para pesquisadores não *outliers* de cada disciplina. As curvas preenchidas coloridas representam os resultados sem levar em consideração os efeitos da idade da carreira, enquanto as curvas preenchidas em cinza mostram as distribuições de μ_P após incluir a idade 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 da idade da carreira no impacto médio dos jornais para pesquisadores não *outliers* de cada disciplina.

idade 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 3.4A 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 da idade da carreira na associação entre prestígio de jornal e produtividade, incluímos a idade da carreira como um preditor do impacto médio dos jornais no modelo bayesiano hierárquico (veja a seção 3.1.4). Nesse caso, a distribuição do coeficiente linear relacionado com o efeito da idade da carreira para cada pesquisador é retirado de outra distribuição com valor médio μ_A. A Figura 3.4B 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 médio dos jornais 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 10 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 química e física (Figura A.57). Mais importante, a Figura 3.4A mostra que a distribuições de μ_P com (curvas coloridas) e sem (curvas em cinza) o efeito da idade da carreira mudam pouco. Dessa forma, o efeito de confusão da idade 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 da idade da carreira.

O conjunto de dados SJR (Figuras A.58 e A.59) estende essa análise para mais disciplinas e apresenta resultados semelhantes para as disciplinas presentes em ambos os conjuntos de dados.

3.3 Conclusões

Investigamos a associação entre a produtividade científica anual e o impacto médio dos jornais 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 dos jornais e práticas específicas de cada disciplina por meio de *scores* robustos de padronização. Esse procedimento permitiu a construção do plano de prestígio de jornal *versus* produtividade: uma representação direta e coerente das performances dos pesquisadores em impacto médio dos jornais e produtividade. Dessa representação, categorizamos os pesquisadores entre *outliers* e não *outliers* e, mais, 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 em nosso conjunto de dados, sendo a performance como *outlier* em apenas um ano da carreira (47,6% dos casos) o comportamento mais comum. Entre os *outliers*, a vasta maioria dos pesquisadores é exclusivamente hiperprolífica ou exclusivamente perfeccionista. Apesar disso, 16 pesquisadores extremamente hiperprolíficos apresentam anos da carreira apenas no setor *IP++* quando têm performances 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 um perfeccionista para acadêmicos que não simultaneamente performam excepcionalmente acima da média em ambas as categorias para todas as disciplinas em nosso conjunto de dados, exceto para engenharia dos materiais. Essa associação negativa varia entre disciplinas, com matemática apresentando o maior efeito negativo e 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 médio dos jornais. Com esse objetivo, estimamos o excesso de transições entre setores do plano prestígio de jornal *versus* produtividade durante anos consecutivos de carreira 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. Transições entre níveis similares de produtividade e prestígio de jornal são tão frequentes como o acaso. Por outro lado, transições entre setores do plano com níveis diferentes de produtividade e impacto de jornal ocorrem muito menos frequentemente que o acaso, indicando que pesquisadores são aversos a mudanças simultâneas de seus níveis de produtividade e impacto médio dos jornais 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 relação a esses dois indicadores sugerem a adoção de determinadas estratégias de publicação e de pesquisa em que os pesquisadores optam por táticas focadas em produtividade ou focadas em impacto de jornal [144]. A fim de manter os níveis de produtividade, os acadêmicos podem adotar estratégias baseadas em expandir colaborações, evitar jornais de alto impacto, dividir seus resultados em vários artigos e selecionar temas de pesquisa mais tradicionais [144]. De outra forma, estratégias focadas em impacto podem apoiar-se em realizar colaborações apenas quando necessário e benéfico para pesquisa, selecionar jornais de alto impacto como a primeira opção, publicar os resultados com maximização do entendimento e impacto em mente e escolher temas de pesquisa mais inovadores [144]. Além disso, nossos resultados ainda podem indicar que as 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 esses hábitos e a adoção dessas estratégias.

Investigamos o efeito agregado médio da idade da careira no prestígio de jornal e na produtividade para todas as disciplinas. Primeiramente, identificamos que o valor médio do prestígio médio dos jornais é ligeiramente maior nos estágios iniciais da carreira com uma

tendência decrescente sutil ao longo dos anos para a maioria das disciplinas. A produtividade média, por sua vez, tende a crescer com a progressão da carreira para todas as disciplinas. Estudamos também o efeito da idade 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 ocupações 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 em estágios iniciais da carreira. É importante ressaltar que tanto a tendência de apresentar maior prestígio de jornal em anos iniciais da carreira bem como a maior probabilidade de encontrar pesquisadores ocupando o setor $I{+}{+}$ no período inicial da carreira podem 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 ou não é uma questão interessante que pesquisas futuras podem abordar. O aumento da produtividade com idade da carreira também foi verificada por Sinatra *et al.* [158], podendo refletir uma série de conquistas que tendem a ser habituais na progressão de carreiras científicas tal como maior familiaridade com os temas de pesquisa [133], maior disponibilidade de recursos financeiros [133, 165] e maior quantidade de convites para elaboração de artigos de revisão [133]. 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, as ciências médicas e da vida) de incluir a chefia de laboratórios científicos em todas as publicações [166].

Os nossos resultados também mostraram que a relação entre produtividade e impacto médio dos jornais 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 os comportamentos heterogêneos individuais dos pesquisadores e identifica um padrão emergente para cada disciplina. Encontramos uma associação geral negativa para maioria das disciplinas ao considerar apenas pesquisadores não *outliers*, um resultado que se alinha com a associação negativa observada em níveis *outliers* de produtividade. No entanto, a intensidade dessa associação varia entre as disciplinas. A física apresenta a associação mais negativa e a matemática apresenta o efeito mais brando da produtividade no prestígio médio dos jornais. Verificamos que, mesmo que a idade da carreira também seja negativamente correlacionada 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.* para criatividade denominada "modelo do caminho duplo para criatividade" ("dual pathway to creativity model" [137]), 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 do mesmo tema exaustivamente). De acordo com essa teoria, os pesquisadores com alta produtividade deveriam estar explorando e associando vários temas e assim permitindo a criação de novas ideias criativas pelo caminho da flexibilidade ou trabalhando e publicando intensivamente na mesma temática até que ideias criativas são criadas pelo caminho da persistência. Desse modo, como a produtividade não se correlaciona positivamente com o impacto médio dos jornais, os indicadores JIF e SJR podem não ser os indicadores mais adequados para a avaliação da criatividade de trabalhos acadêmicos.

APÊNDICE A

Figuras adicionais

Neste apêndice, apresentamos todas as figuras adicionais ao texto principal.

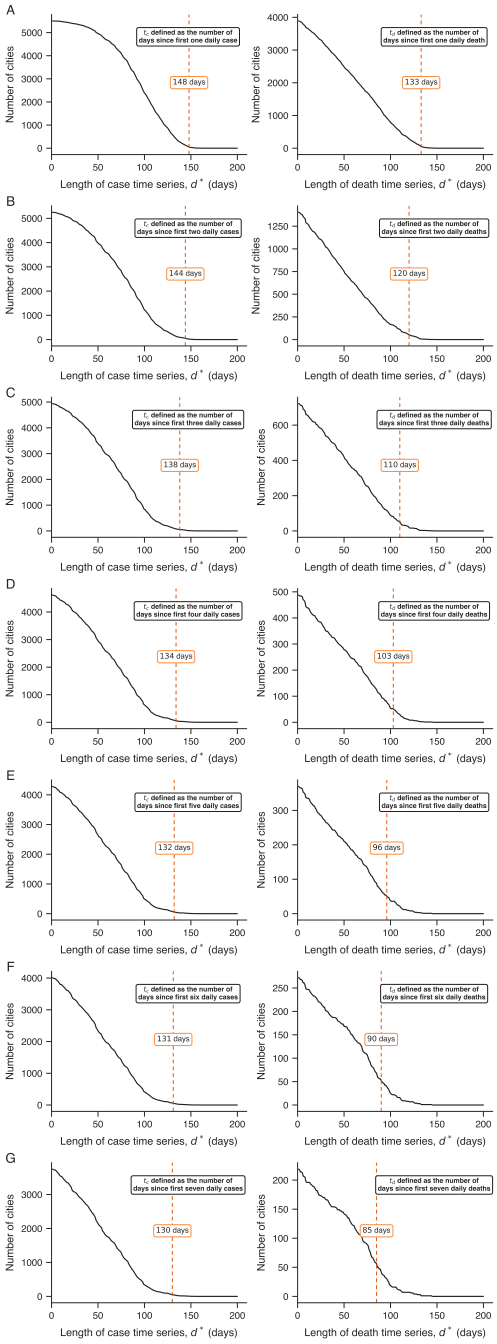

Figura A.1: Número de cidades com série temporal mais longa do que um número particular de dias. Os painéis da esquerda mostram o número de cidades reportando casos de COVID-19 com série temporal maior do que d^* dias. As linhas verticais tracejadas indicam que existem 50 cidades com série temporal de casos confirmados mais longa que um número particular de dias indicado dentro dos gráficos. Os painéis da direita mostram o número de cidades reportando mortes por COVID-19 com série temporal maior do que d^* dias. As linhas verticais tracejadas indicam que existem 50 cidades com série temporal de mortes mais longa que um número particular de dias indicado dentro dos gráficos. Esses limites foram utilizados para garantir que as relações de escala são estimadas para tamanhos de amostra de pelo menos 50 cidades. Os painéis (A)-(G) mostram os resultados considerando os primeiros 1-7 casos diários e as primeiras 1-7 mortes diárias como pontos de referência.

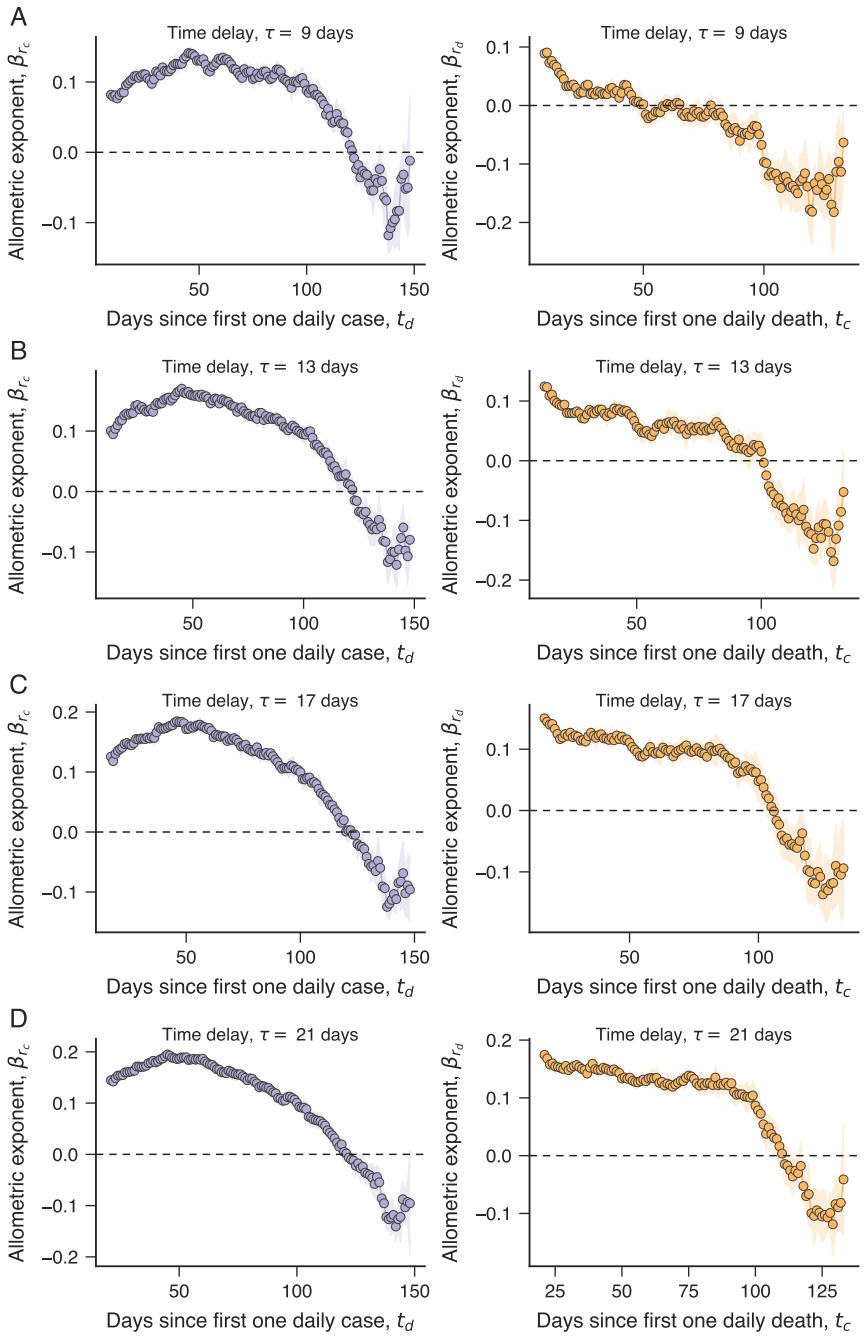

Figura A.2: Variações nos expoentes de escala para taxas de crescimento de casos e mortes sob diferentes escolhas do atraso no tempo τ. Painéis (A)-(D) mostram a dependência do expoente β_{r_c} (painéis à esquerda) e β_{r_d} (painéis à direita) em relação ao número de dias desde o primeiro caso diário (t_c) ou primeira morte diária (t_d) para diferentes valores de τ.

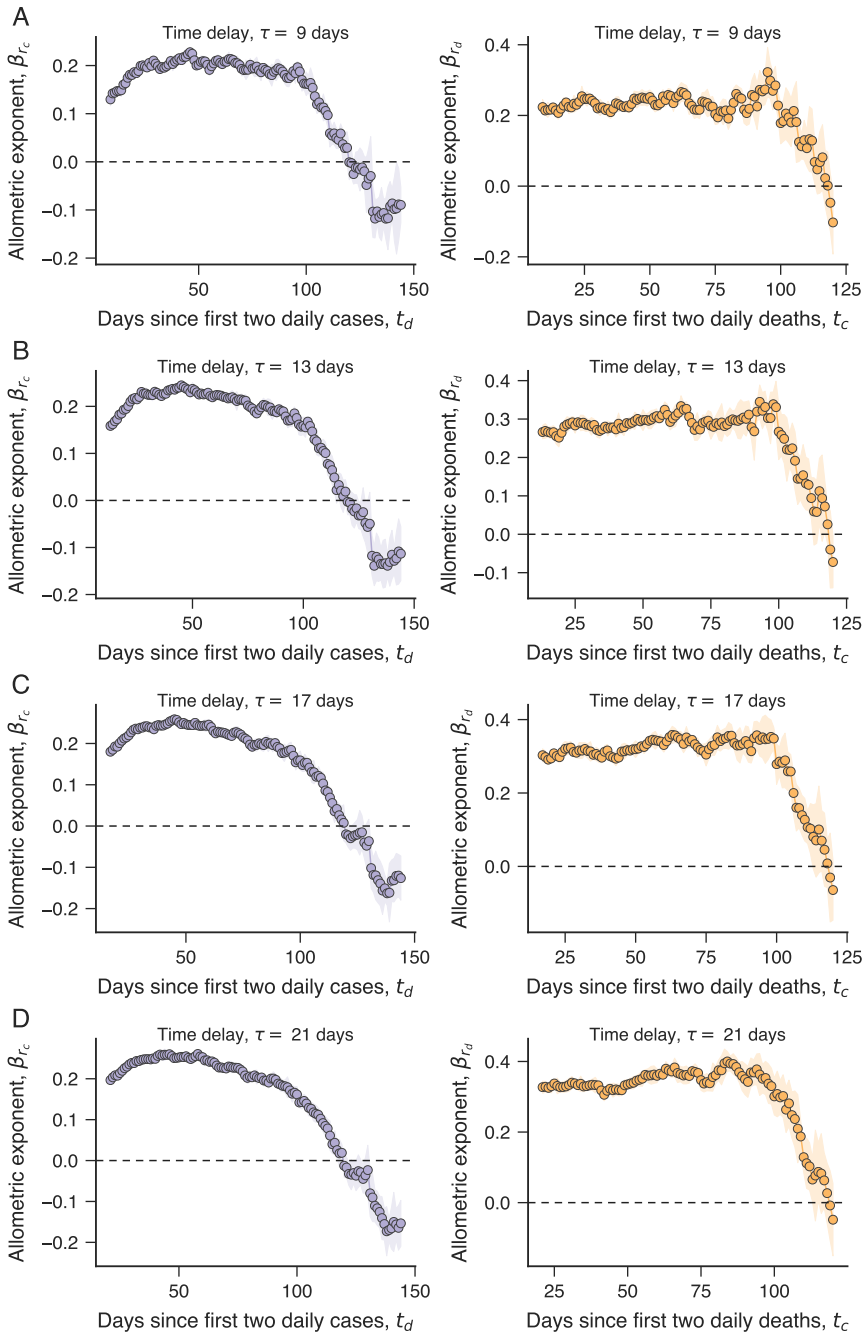

Figura A.3: Variações nos expoentes de escala para taxas de crescimento de casos e mortes sob diferentes escolhas do atraso no tempo τ. Painéis (A)-(D) mostram a dependência do expoente β_{r_c} (painéis à esquerda) e β_{r_d} (painéis à direita) em relação ao número de dias desde os primeiros dois casos (t_c) ou duas primeiras mortes (t_d) diárias para diferentes valores de τ.

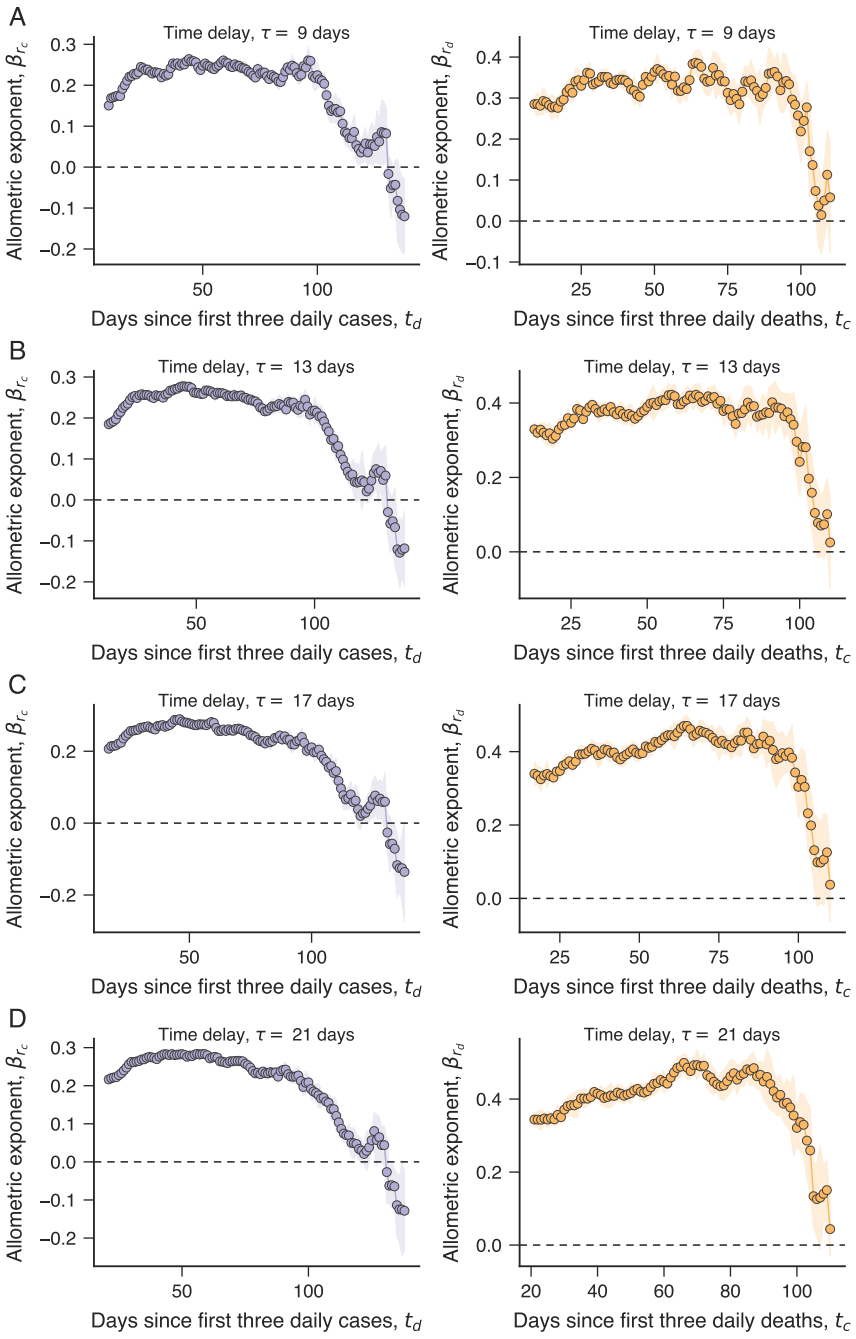

Figura A.4: Variações nos expoentes de escala para taxas de crescimento de casos e mortes sob diferentes escolhas do atraso no tempo τ. Painéis (A)-(D) mostram a dependência do expoente β_{r_c} (painéis à esquerda) e β_{r_d} (painéis à direita) em relação ao número de dias desde os primeiros três casos diários (t_c) ou primeiras três mortes diárias (t_d) para diferentes valores de τ.

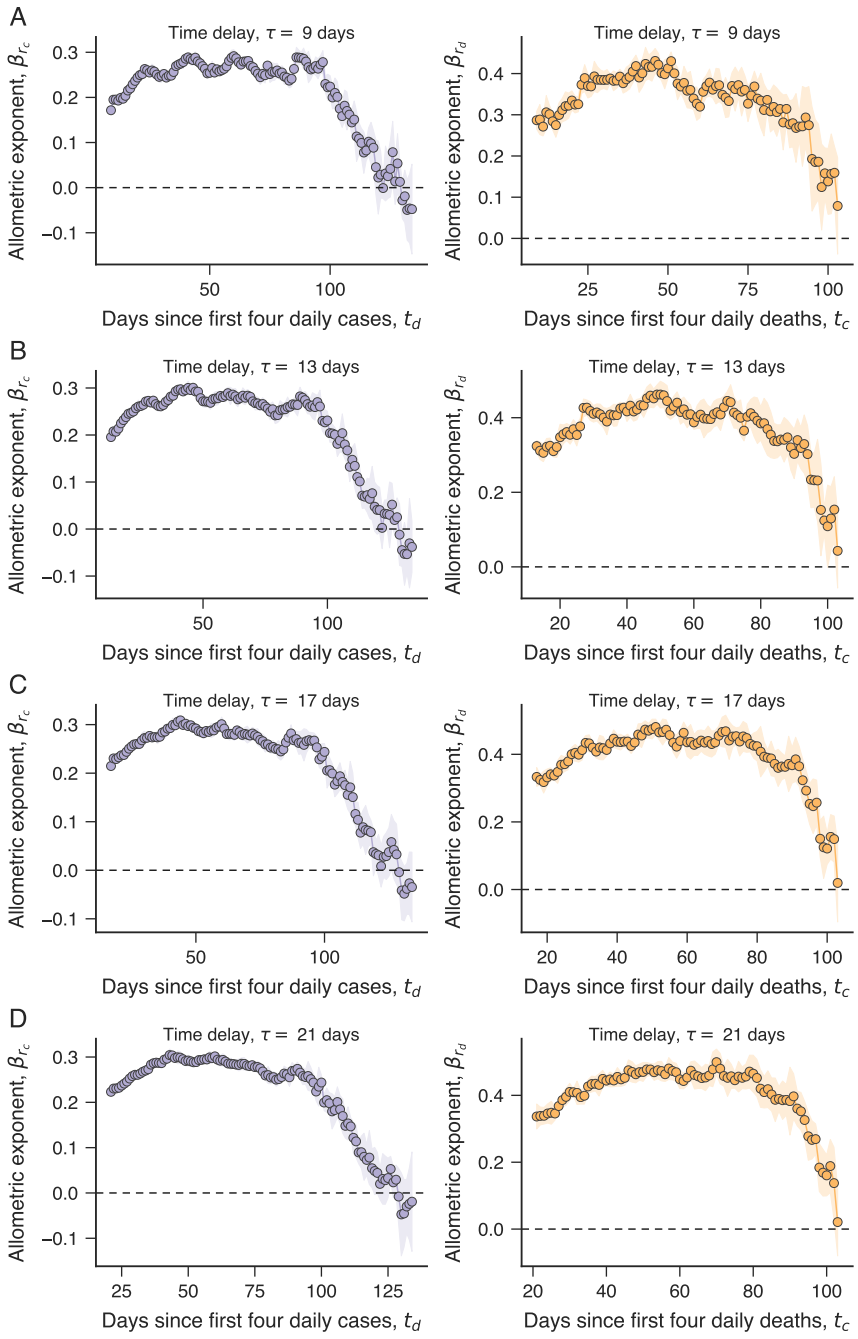

Figura A.5: Variações nos expoentes de escala para taxas de crescimento de casos e mortes sob diferentes escolhas do atraso no tempo τ. Painéis (A)-(D) mostram a dependência do expoente β_{r_c} (painéis à esquerda) e β_{r_d} (painéis à direita) em relação ao número de dias desde os primeiros quatro casos diários (t_c) ou primeiras quatro mortes diárias (t_d) para diferentes valores de τ.

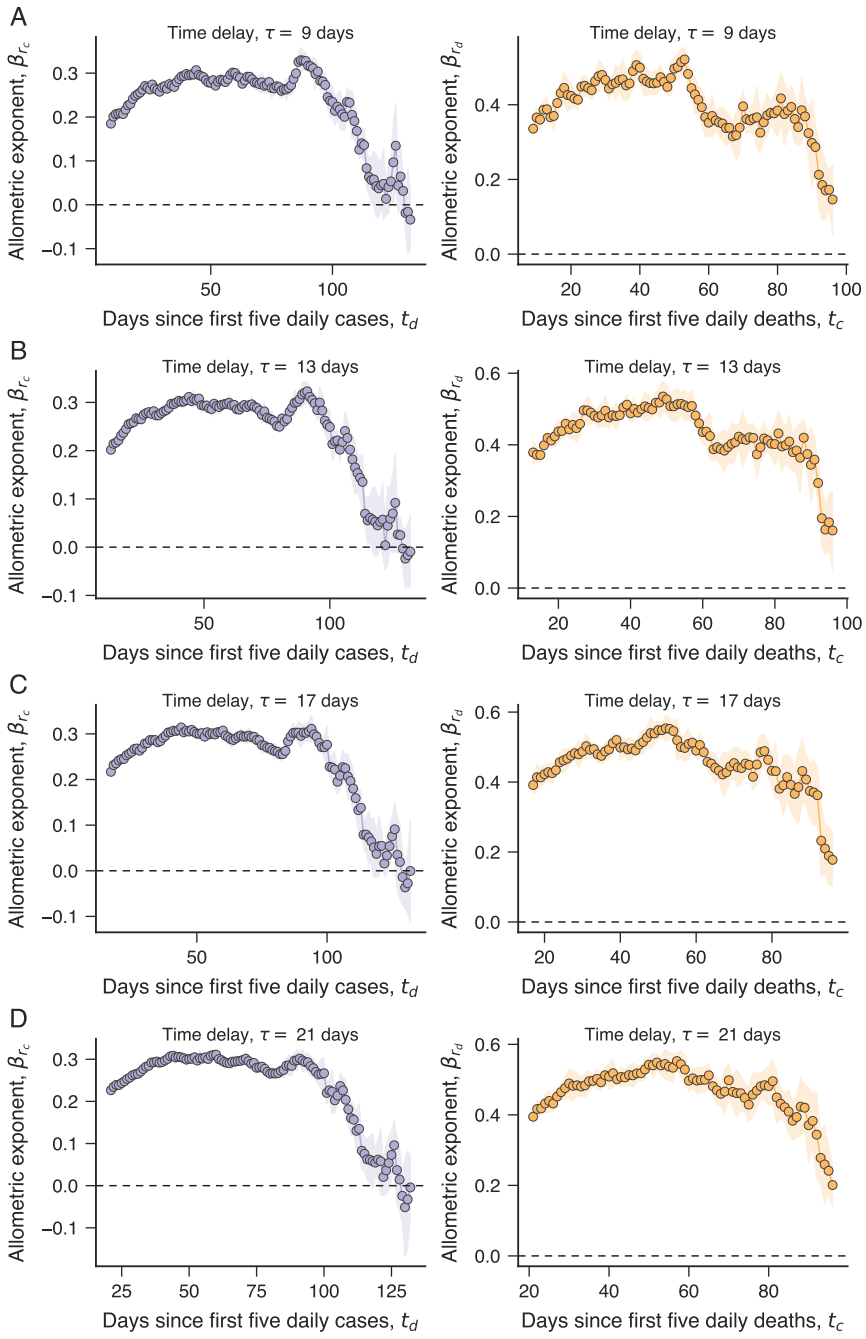

Figura A.6: Variações nos expoentes de escala para taxas de crescimento de casos e mortes sob diferentes escolhas do atraso no tempo τ. Painéis (A)-(D) mostram a dependência do expoente β_{r_c} (painéis à esquerda) e β_{r_d} (painéis à direita) em relação ao número de dias desde os primeiros cinco casos diários (t_c) ou primeiras cinco mortes diárias (t_d) para diferentes valores de τ.

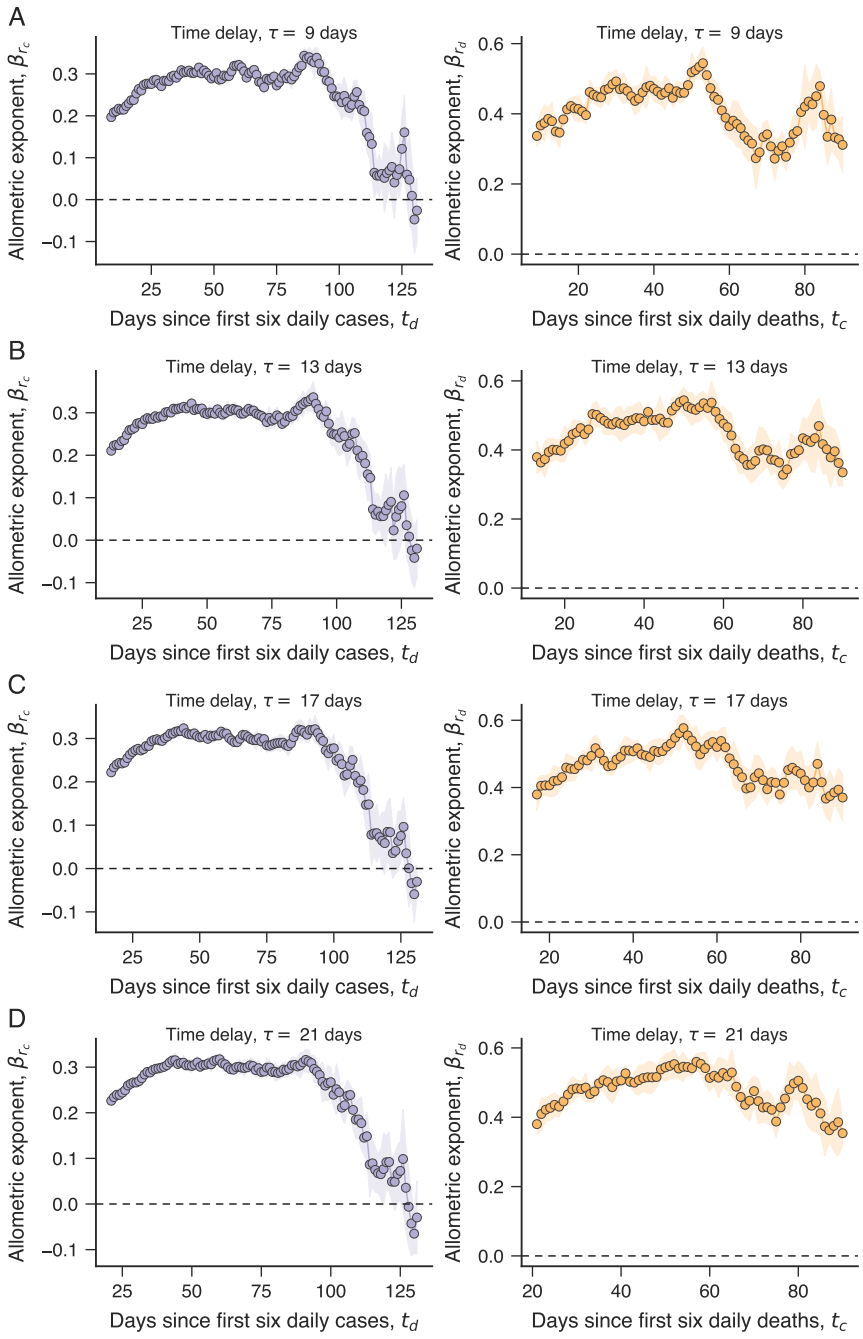

Figura A.7: Variações nos expoentes de escala para taxas de crescimento de casos e mortes sob diferentes escolhas do atraso no tempo τ. Painéis (A)-(D) mostram a dependência do expoente β_{r_c} (painéis à esquerda) e β_{r_d} (painéis à direita) em relação ao número de dias desde os primeiros seis casos diários (t_c) ou primeiras seis mortes diárias (t_d) para diferentes valores de τ.

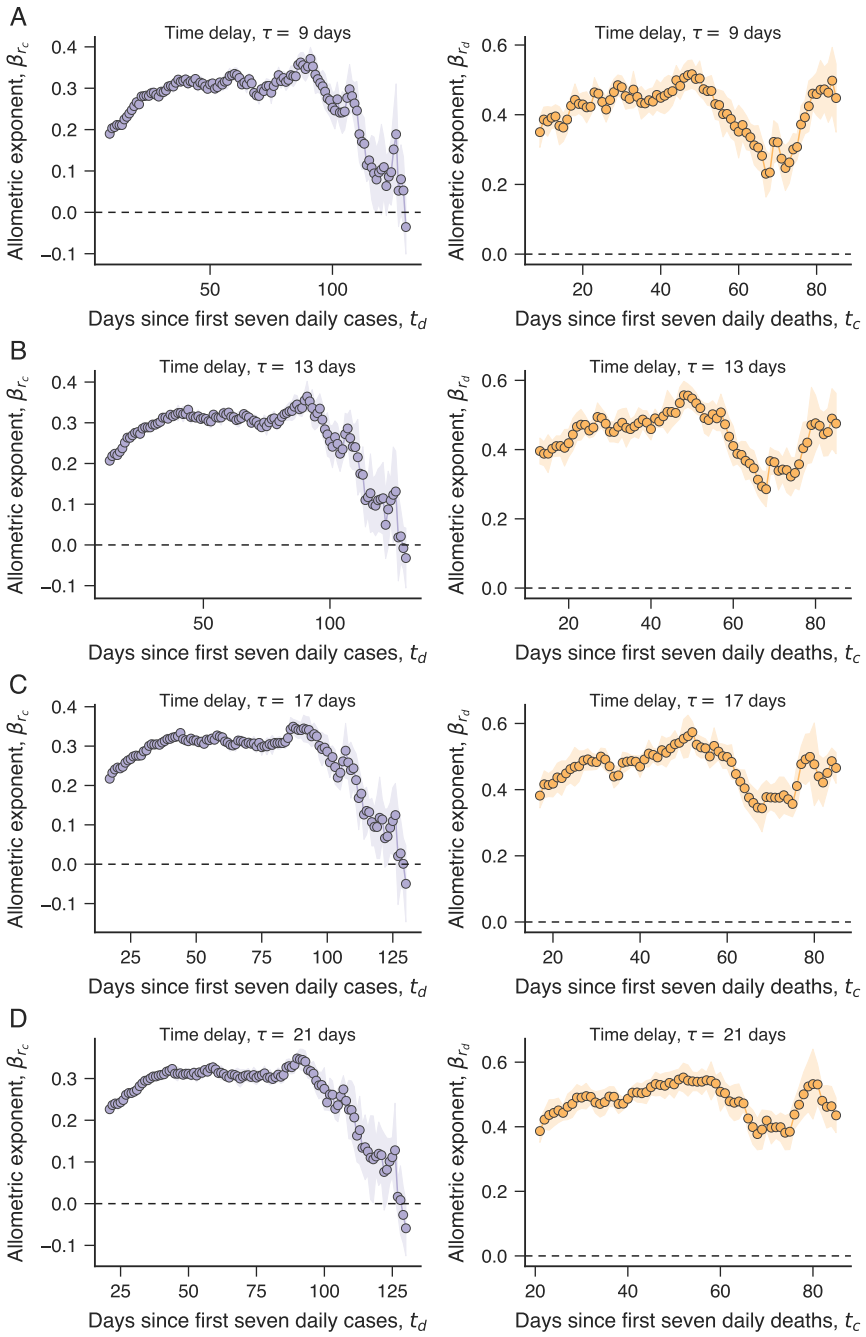

Figura A.8: Variações nos expoentes de escala para taxas de crescimento de casos e mortes sob diferentes escolhas do atraso no tempo τ. Painéis (A)-(D) mostram a dependência do expoente β_{r_c} (painéis à esquerda) e β_{r_d} (painéis à direita) em relação ao número de dias desde os primeiros sete casos diários (t_c) ou primeiras sete mortes diárias (t_d) para diferentes valores de τ.

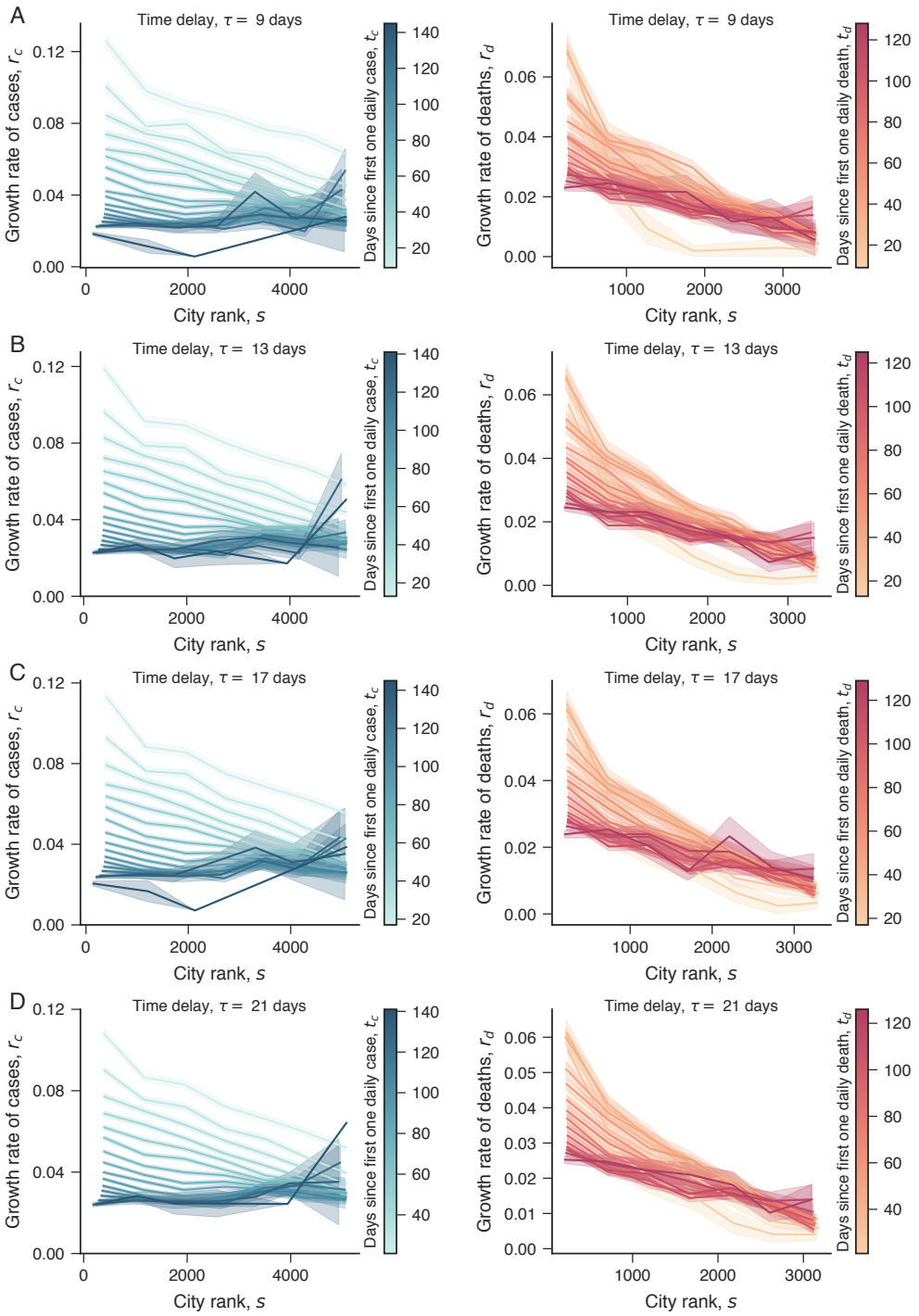

Figura A.9: Variações na associação entre as taxas de crescimento e o ranque das cidades sob diferentes escolhas do atraso no tempo τ. Painéis (A)-(D) mostram a relação média entre as taxas de crescimento dos casos (r_c, painéis à esquerda) e mortes (r_d, painéis à direita) por COVID-19 e o ranque das cidades s para o número de dias desde o primeiro caso diário (t_c) ou primeira morte diária (t_d) e para diferentes valores de τ (indicado dentro dos gráficos).

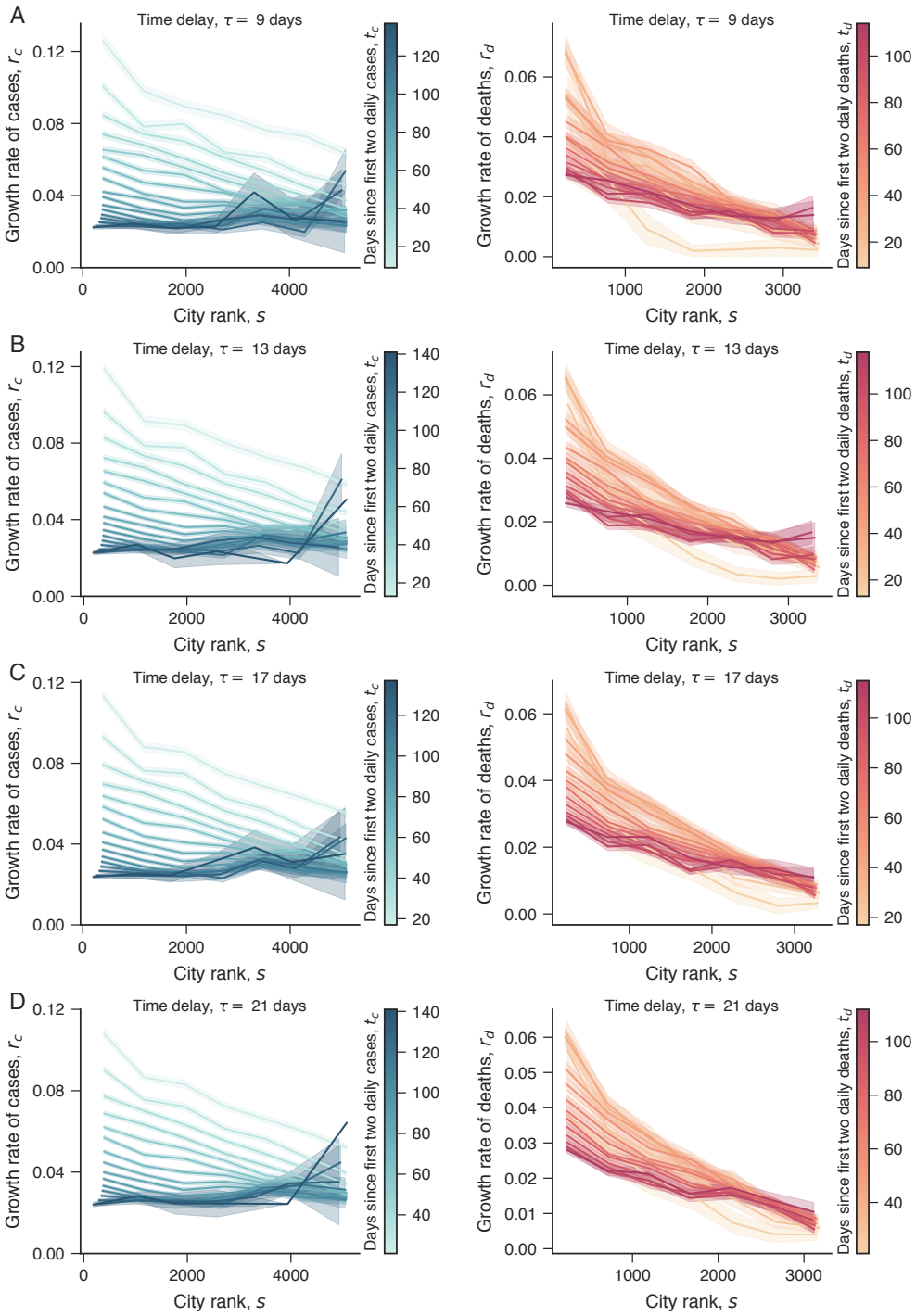

Figura A.10: Variações na associação entre as taxas de crescimento e o ranque das cidades sob diferentes escolhas do atraso no tempo τ. Painéis (A)-(D) mostram a relação média entre as taxas de crescimento dos casos (r_c, painéis à esquerda) e mortes (r_d, painéis à direita) por COVID-19 e o ranque das cidades s para o número de dias desde os primeiros dois casos diários (t_c) ou primeiras duas mortes diárias (t_d) e para diferentes valores de τ (indicado dentro dos gráficos).

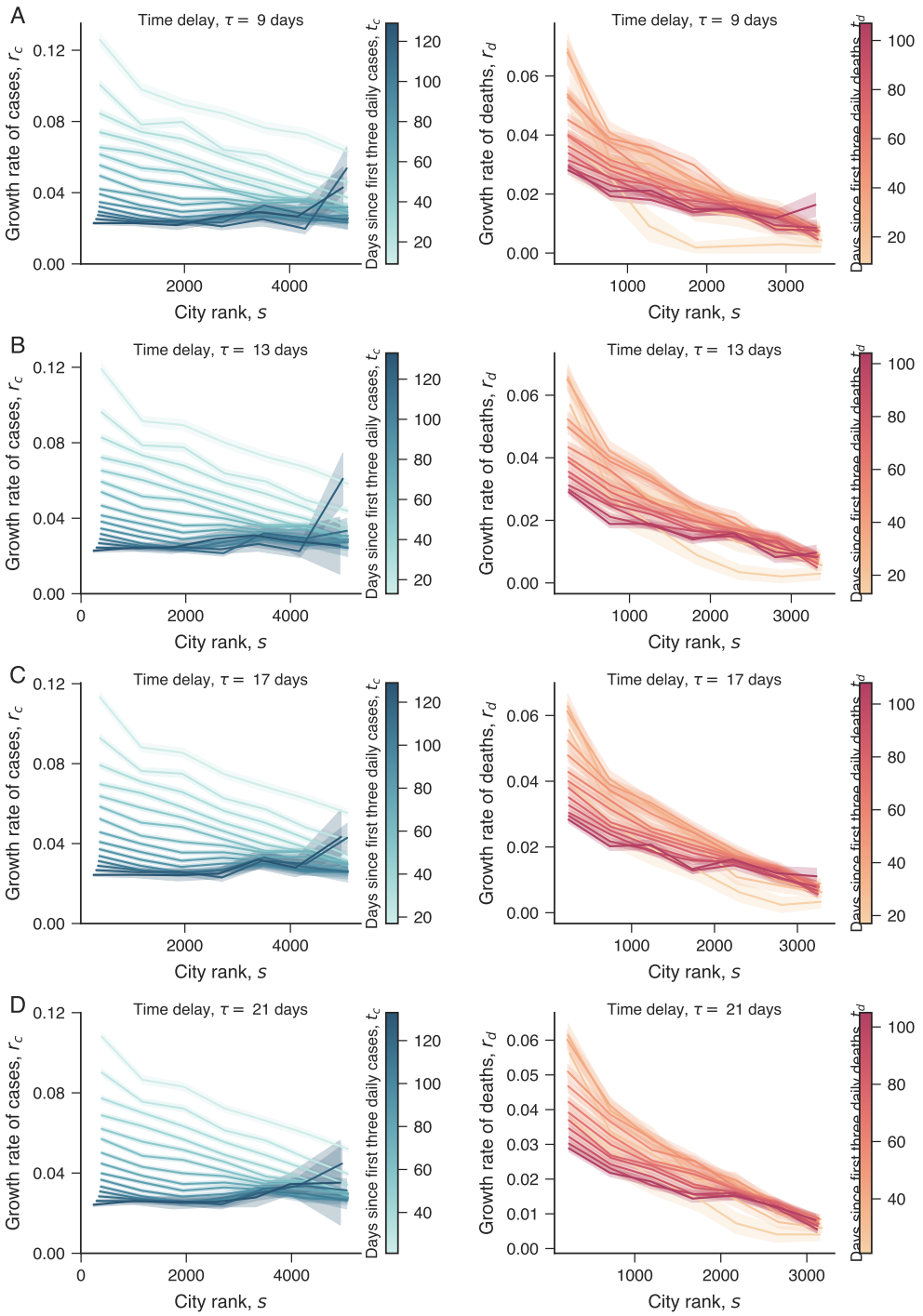

Figura A.11: Variações na associação entre as taxas de crescimento e o ranque das cidades sob diferentes escolhas do atraso no tempo τ. Painéis (A)-(D) mostram a relação média entre as taxas de crescimento dos casos (r_c, painéis à esquerda) e mortes (r_d, painéis à direita) por COVID-19 e o ranque das cidades s para o número de dias desde os primeiros três casos diários (t_c) ou primeiras três mortes diárias (t_d) e para diferentes valores de τ (indicado dentro dos gráficos).

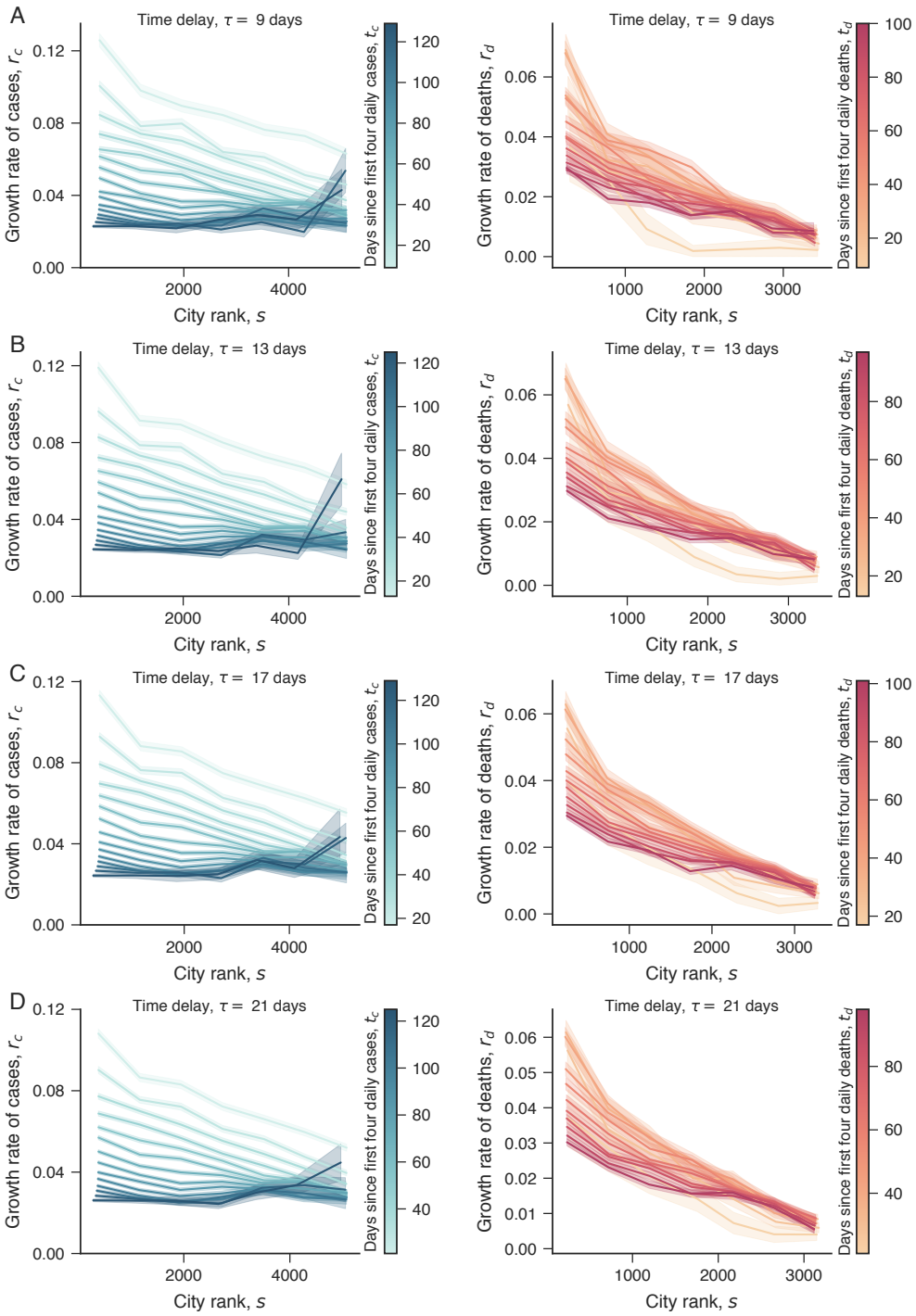

Figura A.12: Variações na associação entre as taxas de crescimento e o ranque das cidades sob diferentes escolhas do atraso no tempo τ. Painéis (A)-(D) mostram a relação média entre as taxas de crescimento dos casos (r_c, painéis à esquerda) e mortes (r_d, painéis à direita) por COVID-19 e o ranque das cidades s para o número de dias desde os primeiros quatro casos diários (t_c) ou primeiras quatro mortes diárias (t_d) e para diferentes valores de τ (indicado dentro dos gráficos).

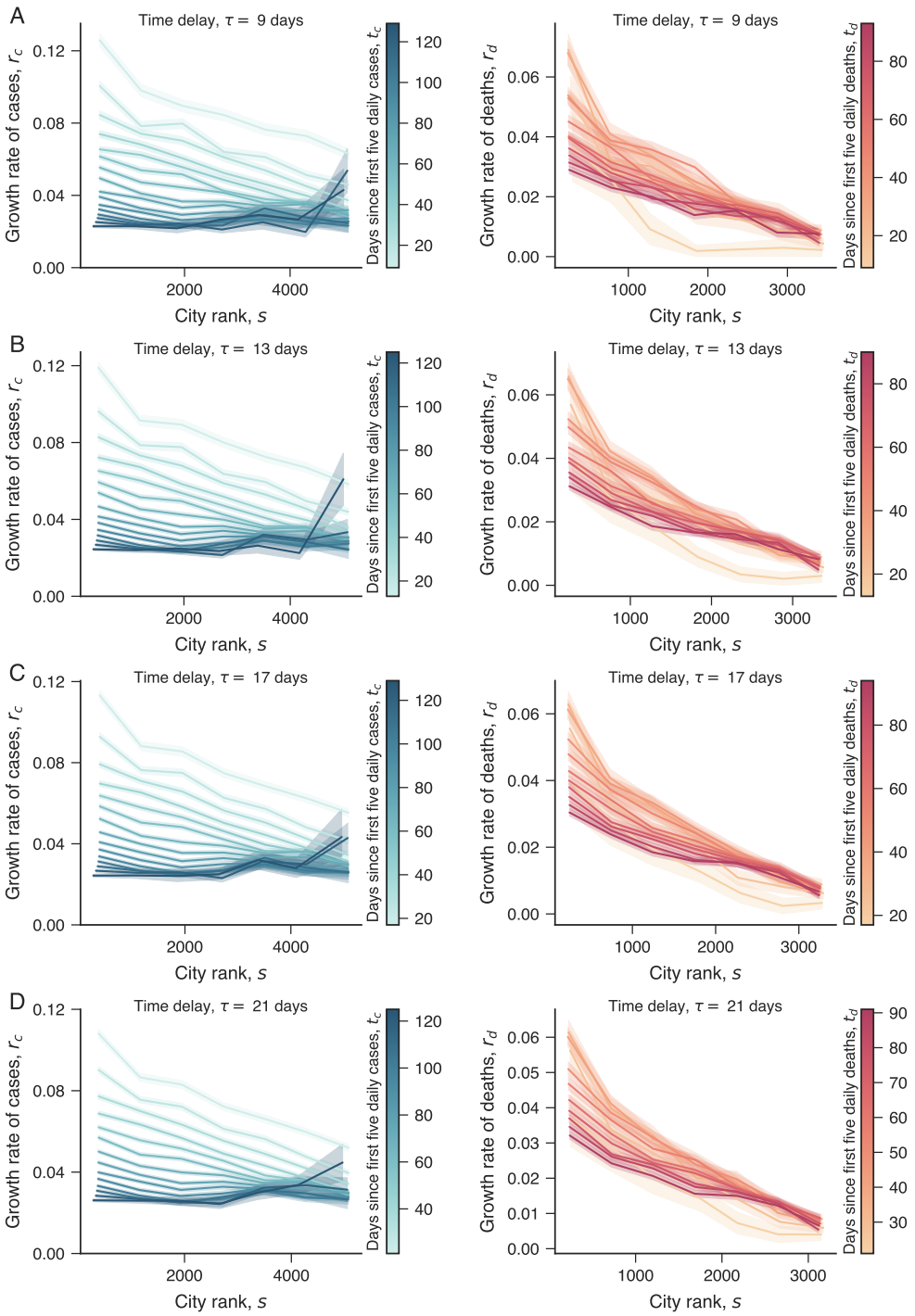

Figura A.13: Variações na associação entre as taxas de crescimento e o ranque das cidades sob diferentes escolhas do atraso no tempo τ. Painéis (A)-(D) mostram a relação média entre as taxas de crescimento dos casos (r_c, painéis à esquerda) e mortes (r_d, painéis à direita) por COVID-19 e o ranque das cidades s para o número de dias desde os primeiros cinco casos diários (t_c) ou primeiras cinco mortes diárias (t_d) e para diferentes valores de τ (indicado dentro dos gráficos).

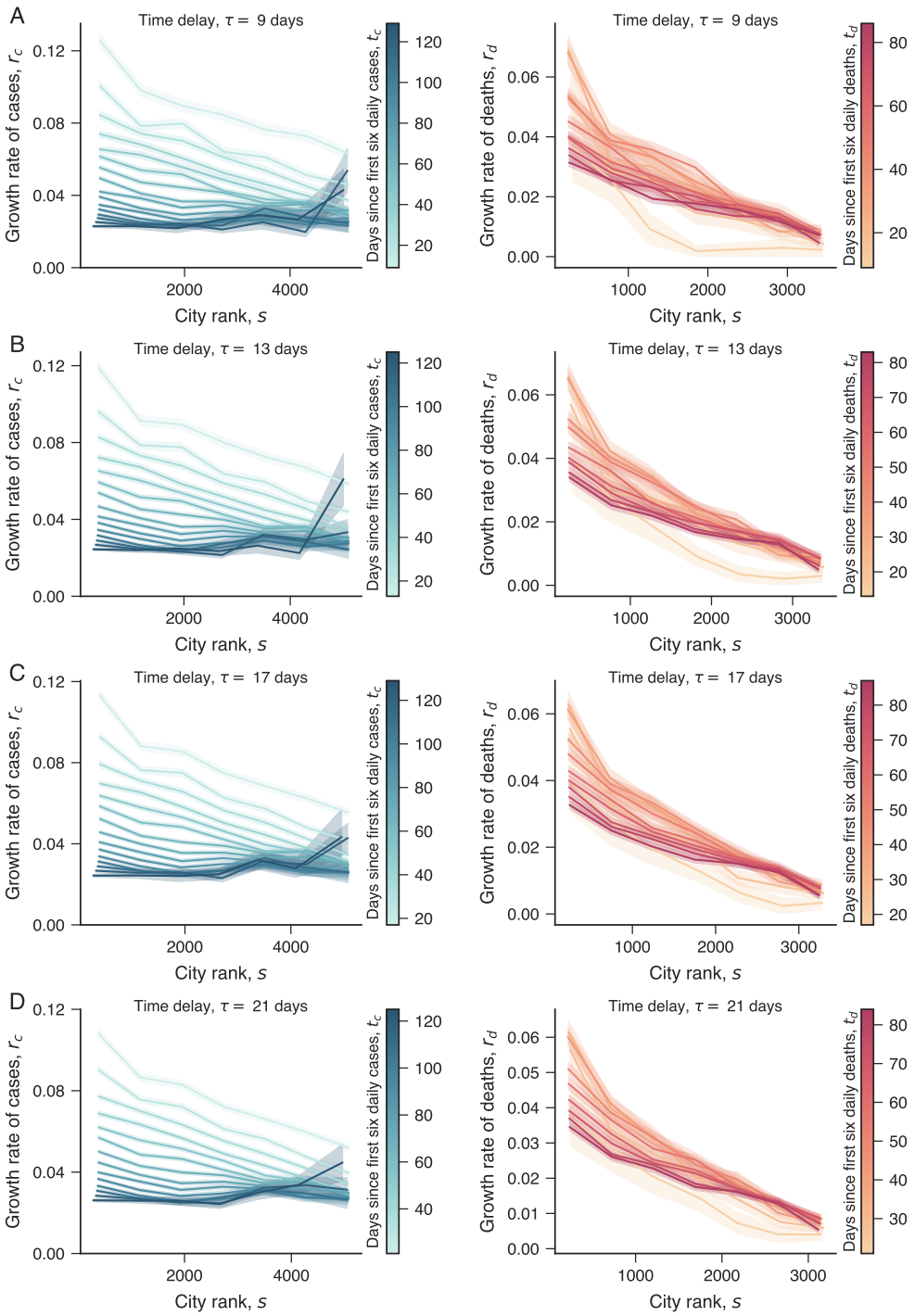

Figura A.14: Variações na associação entre as taxas de crescimento e o ranque das cidades sob diferentes escolhas do atraso no tempo τ. Painéis (A)-(D) mostram a relação média entre as taxas de crescimento dos casos (r_c, painéis à esquerda) e mortes (r_d, painéis à direita) por COVID-19 e o ranque das cidades s para o número de dias desde os primeiros seis casos diários (t_c) ou primeiras seis mortes diárias (t_d) e para diferentes valores de τ (indicado dentro dos gráficos).

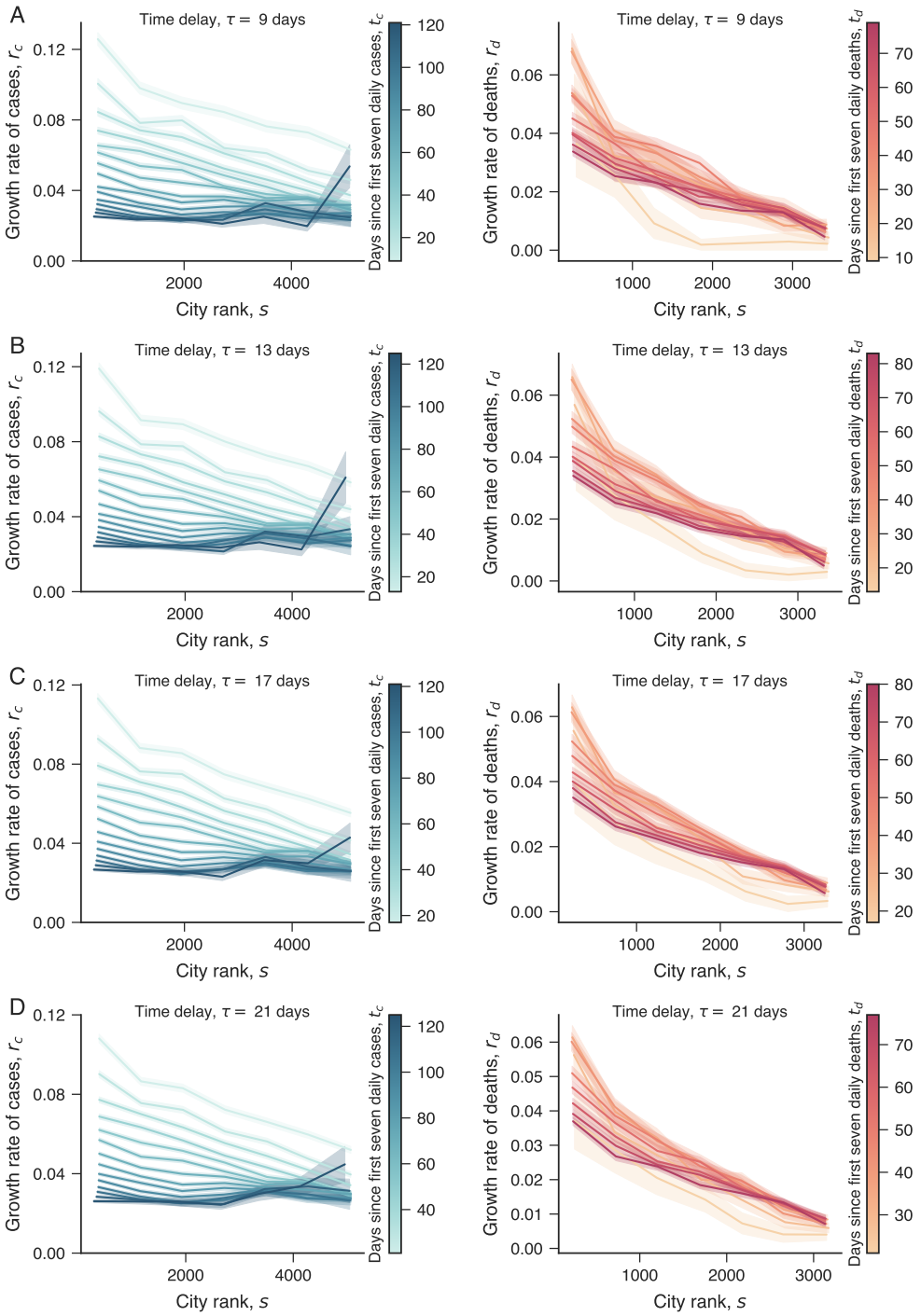

Figura A.15: Variações na associação entre as taxas de crescimento e o ranque das cidades sob diferentes escolhas do atraso no tempo τ. Painéis (A)-(D) mostram a relação média entre as taxas de crescimento dos casos (r_c, painéis à esquerda) e mortes (r_d, painéis à direita) por COVID-19 e o ranque das cidades s para o número de dias desde os primeiros sete casos diários (t_c) ou primeiras sete mortes diárias (t_d) e para diferentes valores de τ (indicado dentro dos gráficos).

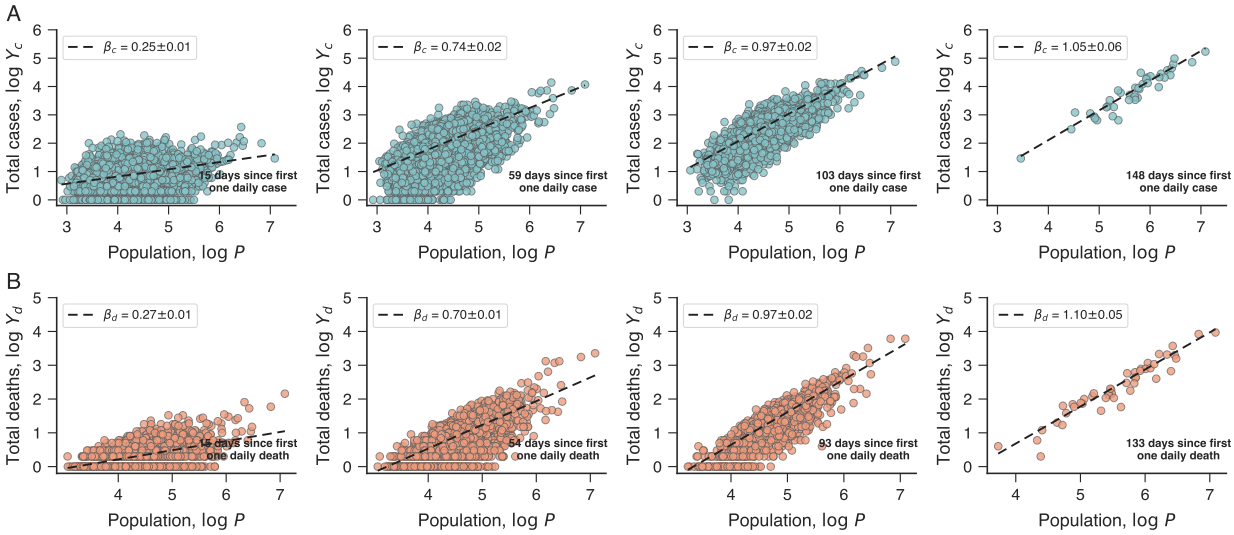

Figura A.16: Relações de escala urbana de casos e mortes por COVID-19 sob diferentes escolhas de valores para o número de casos diários ou mortes diárias como ponto de referência. Os mesmos gráficos da Figura 2.1 do texto principal mas considerando o primeiro caso diário e primeira morte diária como pontos de referência.

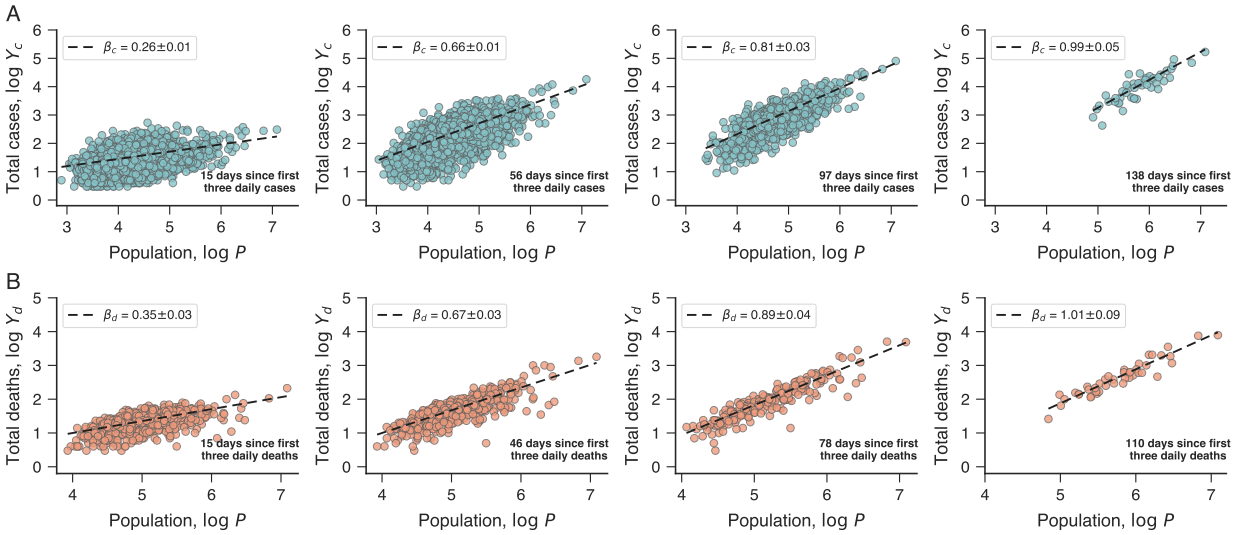

Figura A.17: Relações de escala urbana de casos e mortes por COVID-19 sob diferentes escolhas de valores para o número de casos diários ou mortes diárias como ponto de referência. Os mesmos gráficos da Figura 2.1 do texto principal mas considerando os primeiros três casos diários e primeiras três mortes diárias como pontos de referência.

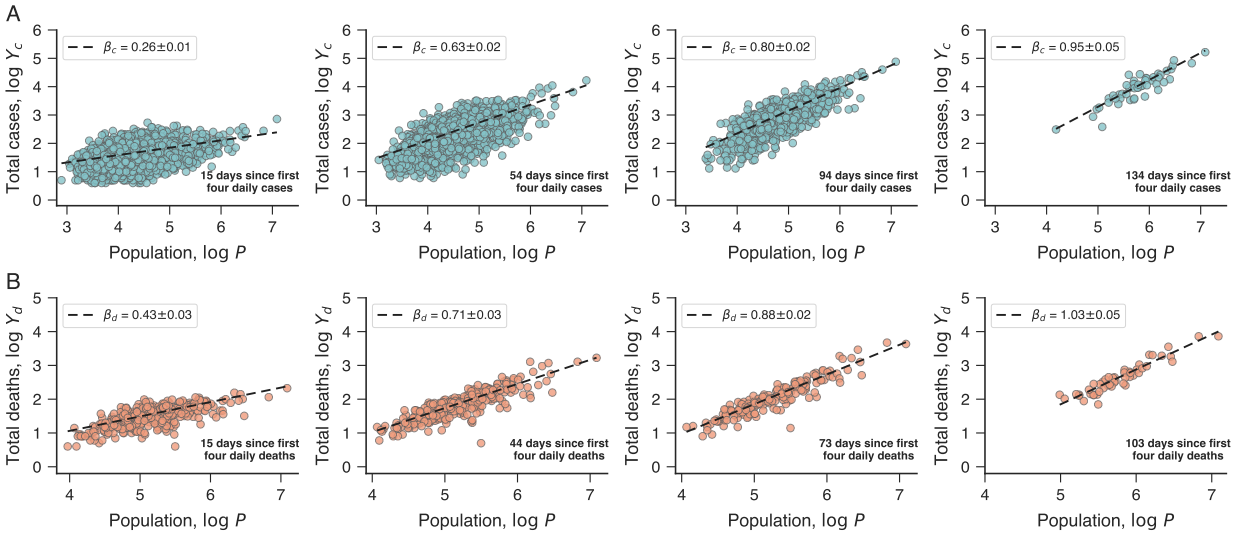

Figura A.18: Relações de escala urbana de casos e mortes por COVID-19 sob diferentes escolhas de valores para o número de casos diários ou mortes diárias como ponto de referência. Os mesmos gráficos da Figura 2.1 do texto principal mas considerando os primeiros quatro casos diários e primeiras quatro mortes diárias como pontos de referência.

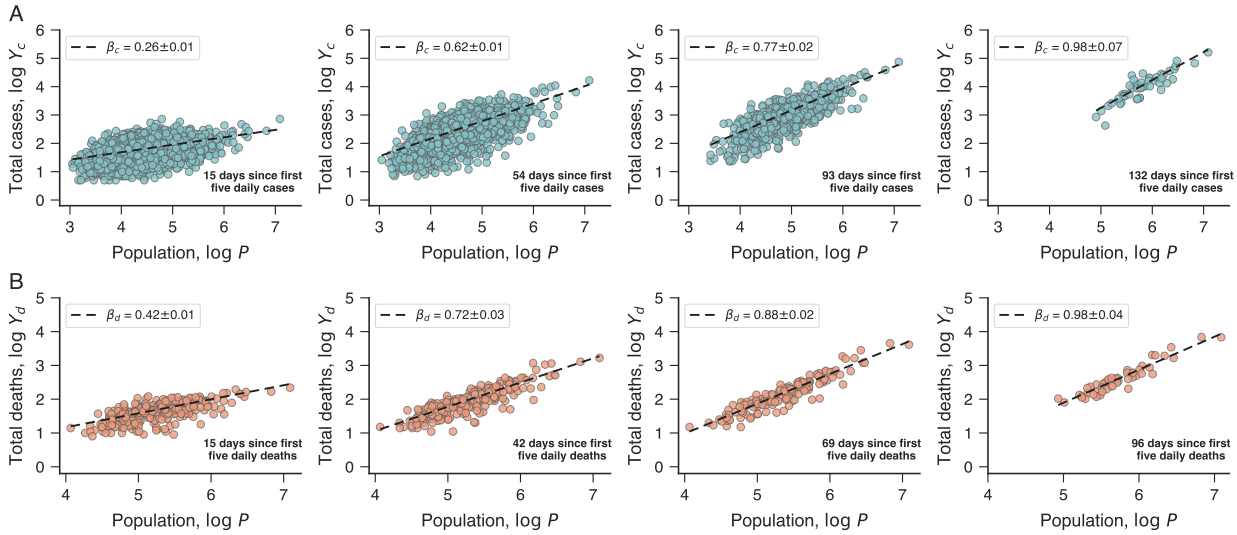

Figura A.19: Relações de escala urbana de casos e mortes por COVID-19 sob diferentes escolhas de valores para o número de casos diários ou mortes diárias como ponto de referência. Os mesmos gráficos da Figura 2.1 do texto principal mas considerando os primeiros cinco casos diários e primeiras cinco mortes diárias como pontos de referência.

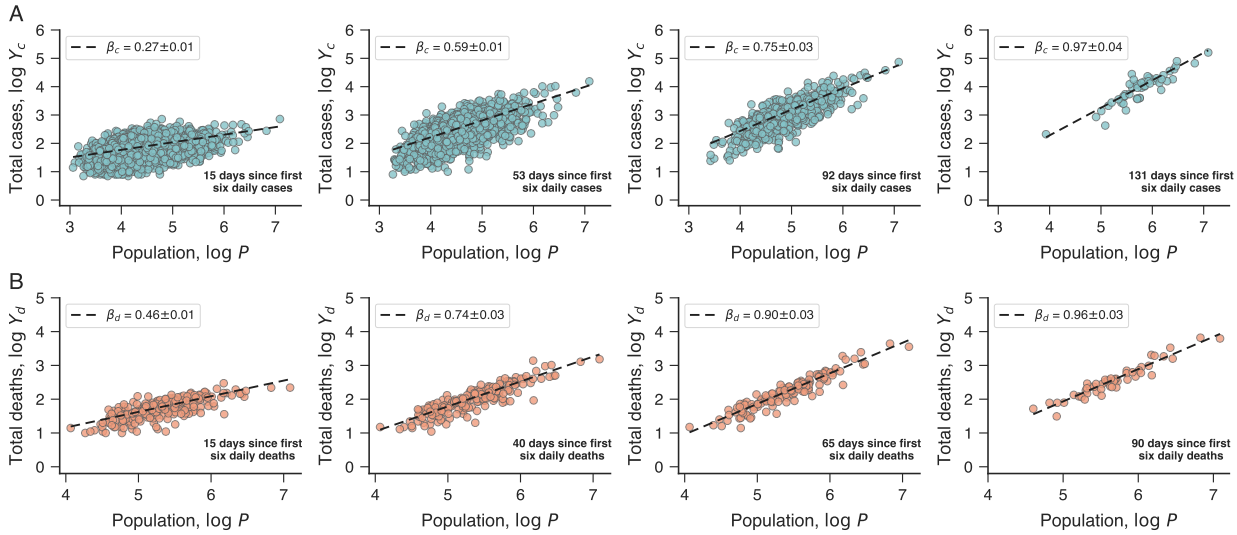

Figura A.20: Relações de escala urbana de casos e mortes por COVID-19 sob diferentes escolhas de valores para o número de casos diários ou mortes diárias como ponto de referência. Os mesmos gráficos da Figura 2.1 do texto principal mas considerando os primeiros seis casos diários e primeiras seis mortes diárias como pontos de referência.

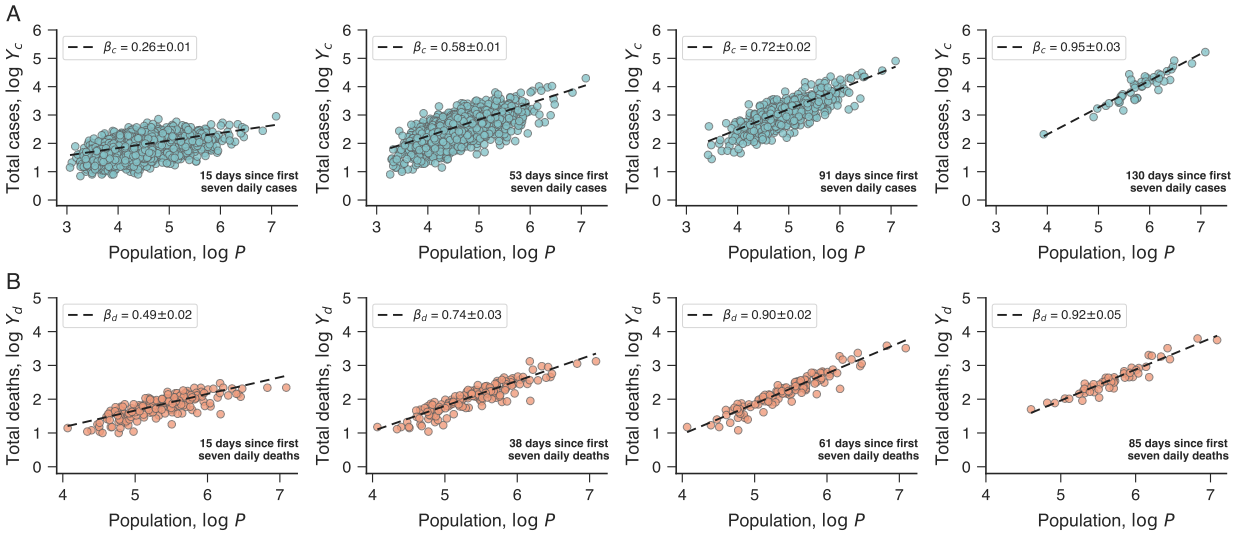

Figura A.21: Relações de escala urbana de casos e mortes por COVID-19 sob diferentes escolhas de valores para o número de casos diários ou mortes diárias como ponto de referência. Os mesmos gráficos da Figura 2.1 do texto principal mas considerando os primeiros sete casos diários e primeiras sete mortes diárias como pontos de referência.

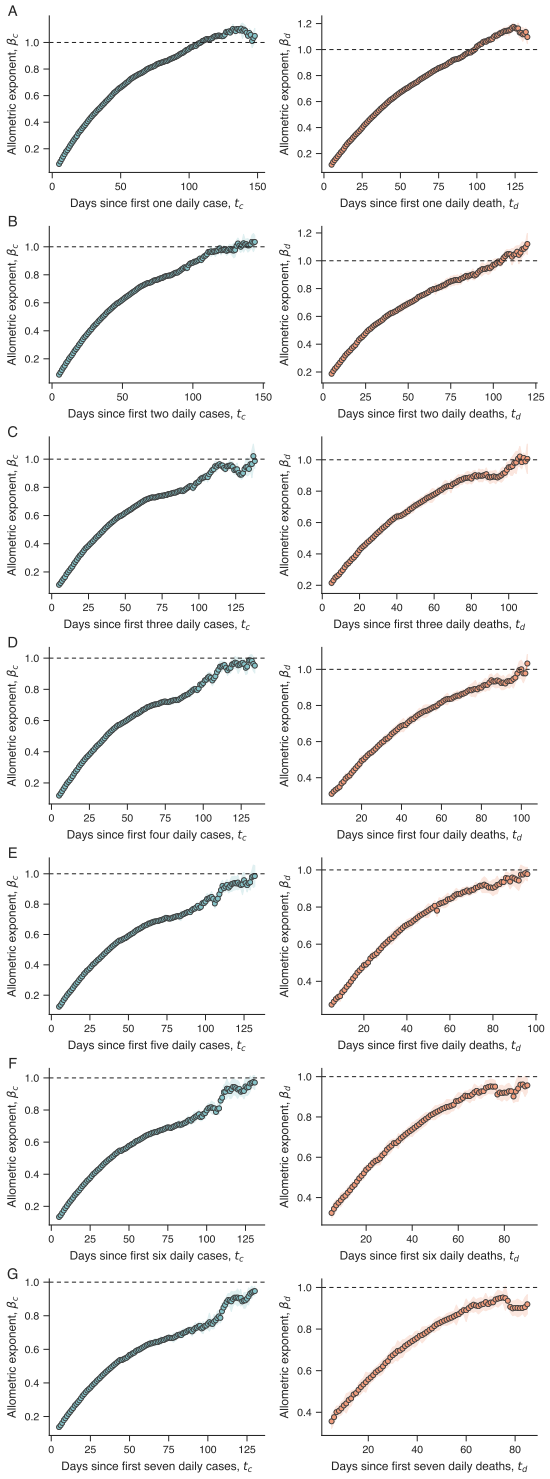

Figura A.22: Dependência temporal dos expoentes de escala para casos e mortes por COVID-19 sob diferentes escolhas de valores para o número de casos diários ou mortes diárias como ponto de referência. Painéis (A)-(G) mostram a dependência dos expoentes β_c (esquerda) e β_d (direita) em relação a t_c e t_d considerando os primeiros 1-7 casos diários e as primeiras 1-7 mortes diárias como pontos de referência. As regiões sombreadas representam o desvio padrão e as linhas tracejadas horizontais representam $\beta_c = \beta_d = 1$. Notamos que o comportamento observado para números grandes de ponto de referência tende a seguir o comportamento dos números pequenos na evolução do espalhamento da COVID-19.

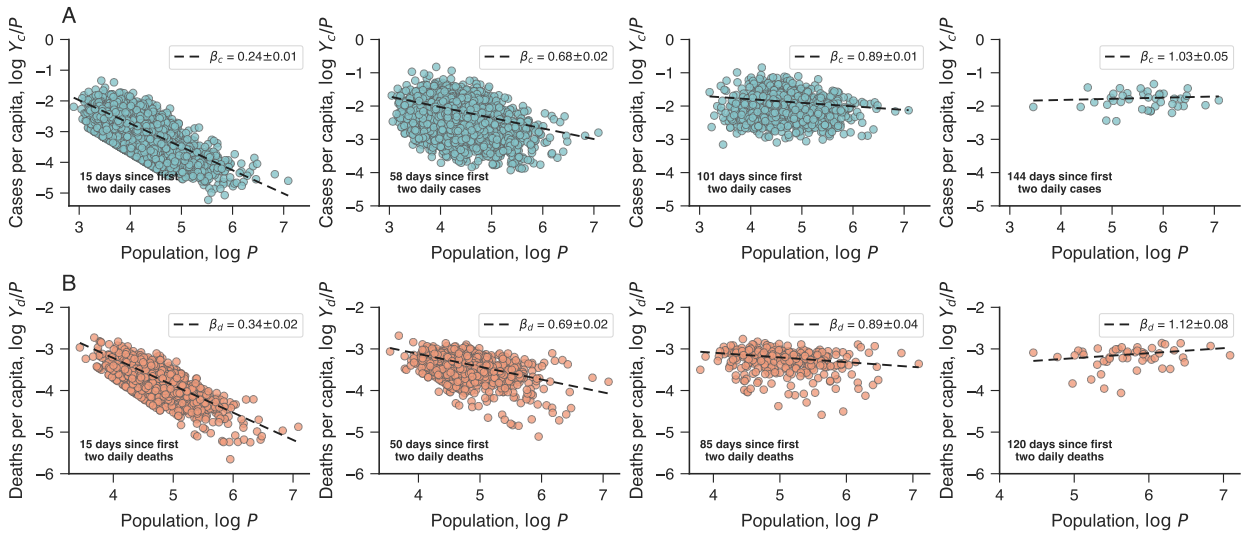

Figura A.23: Relações de escala urbana de casos e mortes por COVID-19 *per capita*. **(A)** Relação entre o número total de casos confirmados de COVID-19 *per capita* (Y_c/P) e a população das cidades (P) em escala logarítmica. **(B)** Relação entre o número total de mortes por COVID-19 *per capita* (Y_d/P) e a população das cidades (P) em escala logarítmica. Os painéis mostram as relações de escala no dia particular após os primeiros dois casos diários ou após as primeiras duas mortes diárias. As linhas tracejadas representam as relações de escala com expoentes indicados indicados em cada gráfico ($\beta_c - 1$ para casos *per capita* e $\beta_d - 1$ para mortes *per capita*).

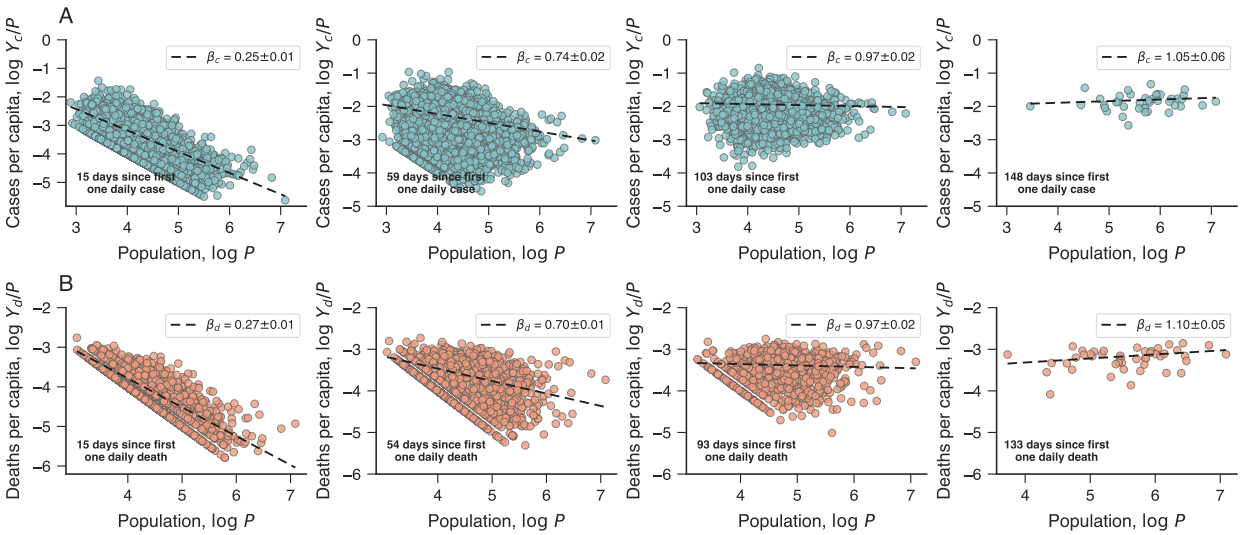

Figura A.24: Relações de escala urbana de casos e mortes por COVID-19 *per capita*. O mesmo que a Figura A.23 mas considerando o primeiro caso diário e a primeira morte diária como pontos de referência.

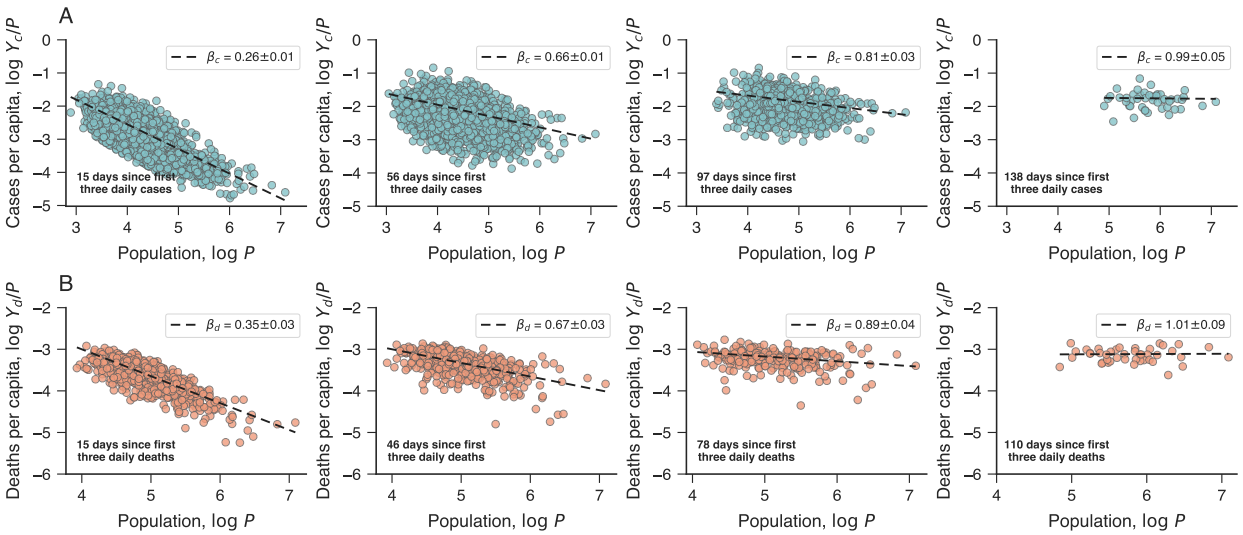

Figura A.25: Relações de escala urbana de casos e mortes por COVID-19 *per capita*. O mesmo que a Figura A.23 mas considerando os primeiros três casos diários e as primeiras três mortes diárias como pontos de referência.

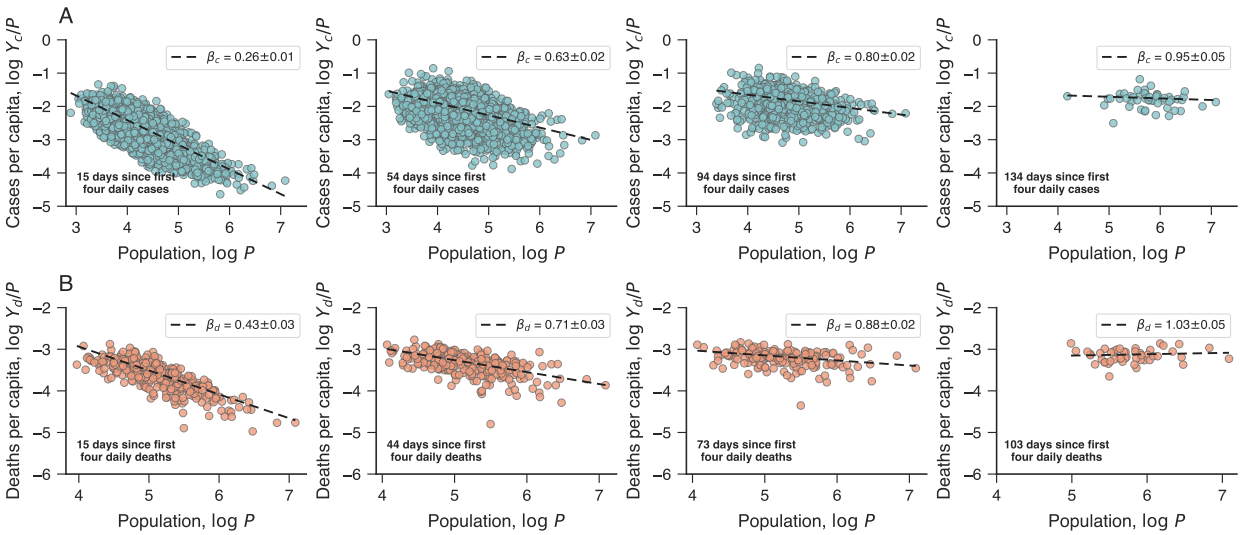

Figura A.26: Relações de escala urbana de casos e mortes por COVID-19 *per capita*. O mesmo que a Figura A.23 mas considerando os primeiros quatro casos diários e as primeiras quatro mortes diárias como pontos de referência.

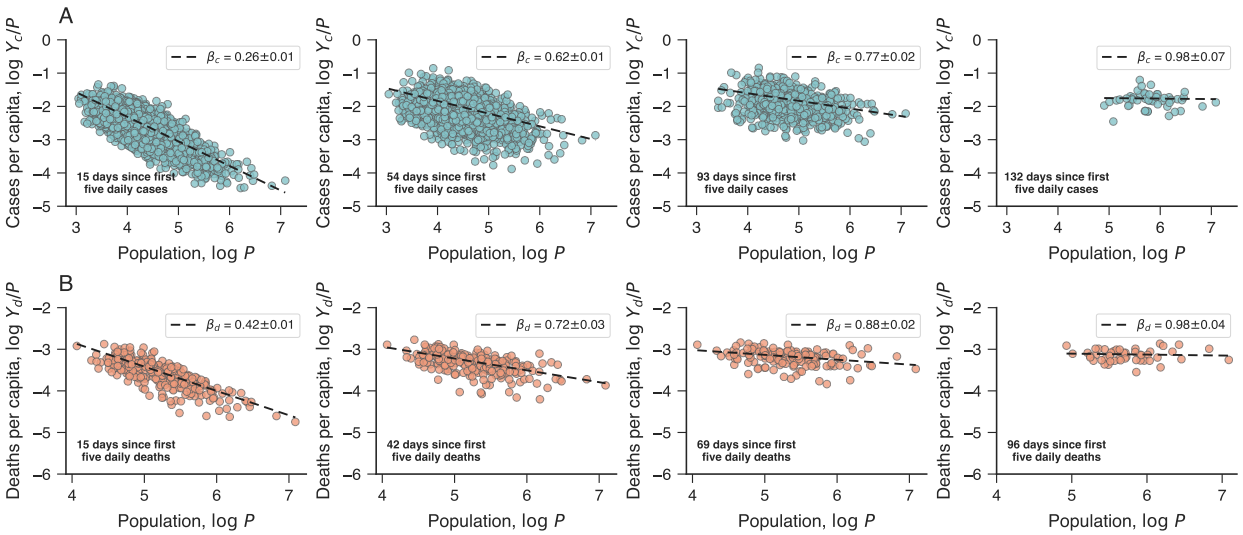

Figura A.27: Relações de escala urbana de casos e mortes por COVID-19 *per capita*. O mesmo que a Figura A.23 mas considerando os primeiros cinco casos diários e as primeiras cinco mortes diárias como pontos de referência.

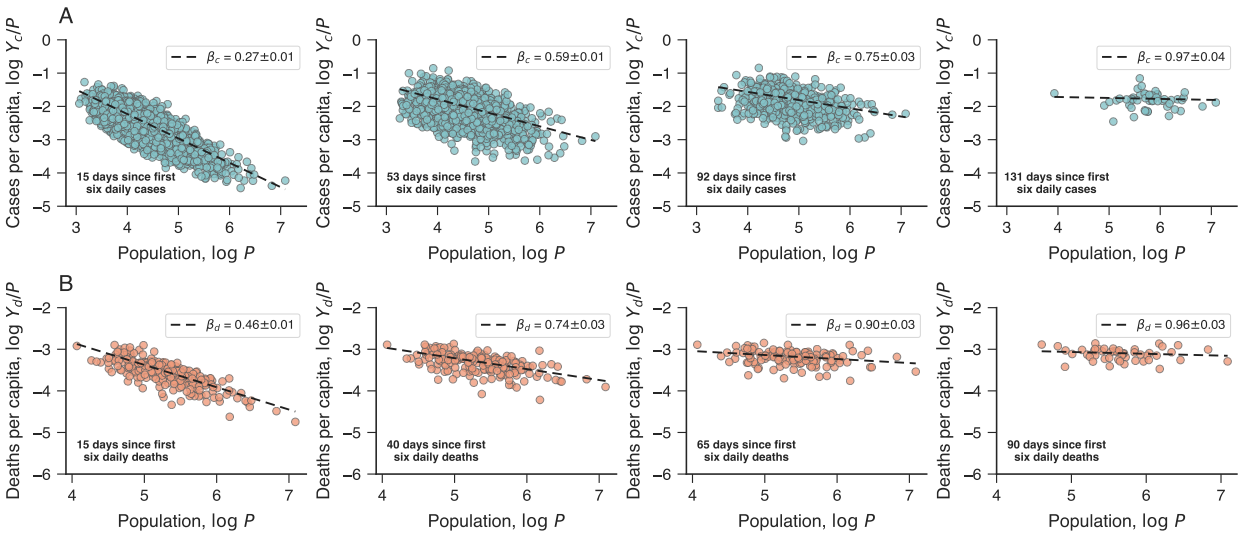

Figura A.28: Relações de escala urbana de casos e mortes por COVID-19 *per capita*. O mesmo que a Figura A.23 mas considerando os primeiros seis casos diários e as primeiras seis mortes diárias como pontos de referência.

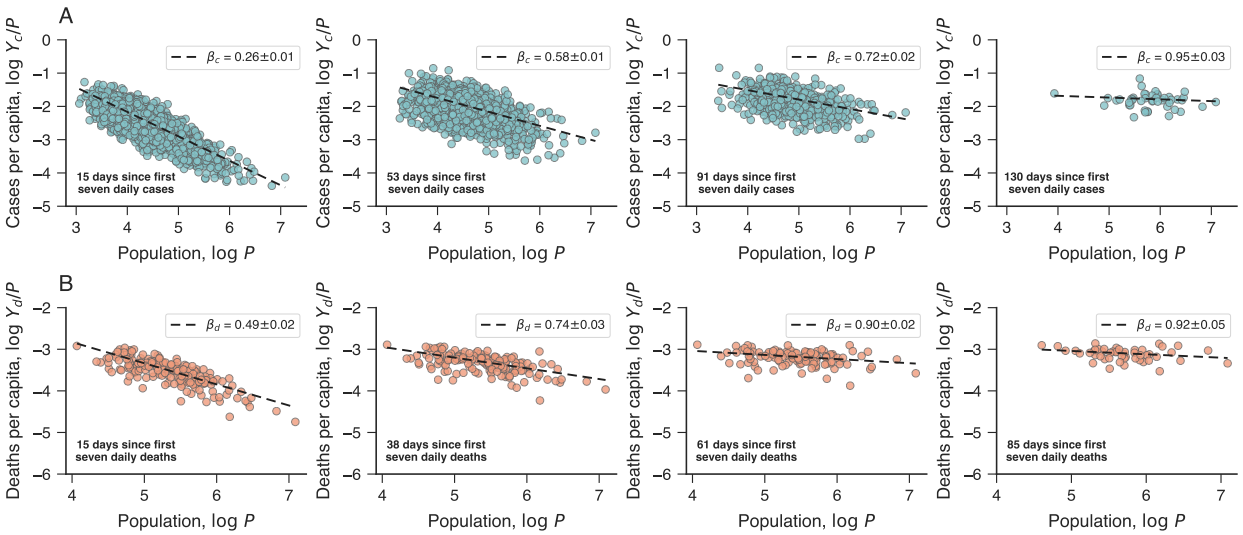

Figura A.29: Relações de escala urbana de casos e mortes por COVID-19 *per capita*. O mesmo que a Figura A.23 mas considerando os primeiros sete casos diários e as primeiras sete mortes diárias como pontos de referência.

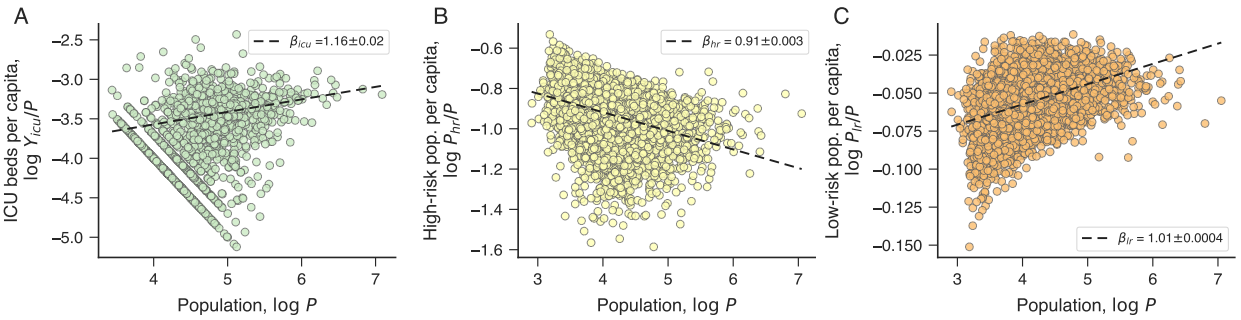

Figura A.30: Escala urbana dos leitos de UTI e das populações de alto e de baixo risco *per capita*. **(A)** Relação entre o número de leitos de UTI *per capita* (Y_{icu}/P) e a população das cidades (P) em escala logarítmica. **(B)** Relação entre a população de alto risco *per capita* (P_{hr}/P) e a população das cidades (P) em escala logarítmica. **(C)** Relação entre a população de baixo risco *per capita* (P_{lr}/P) e a população das cidades (P) em escala logarítmica. Em todos os painéis, as linhas tracejadas indicam as relações de escala com expoentes especificados em cada gráfico ($\beta_{icu} - 1$ para leitos de UTI *per capita*, $\beta_{hr} - 1$ para população de alto risco *per capita* e $\beta_{lr} - 1$ para população de baixo risco *per capita*).

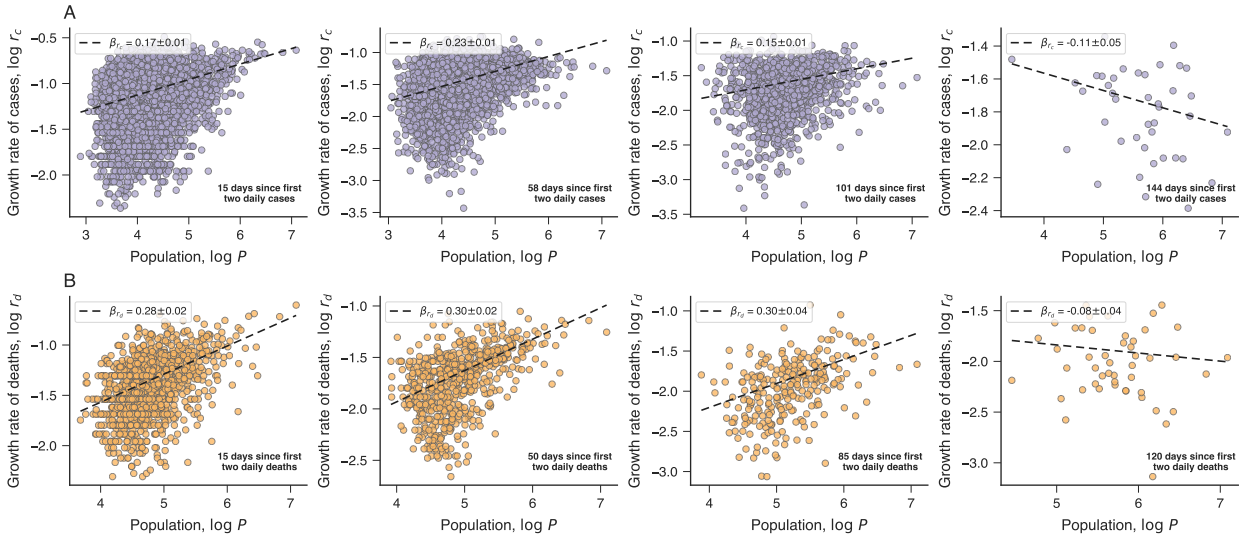

Figura A.31: Relações de escala urbana das taxas de crescimento de casos e mortes por COVID-19. (A) Relação entre a taxa de crescimento de casos (r_c) e a população das cidades (P) em escala logarítmica. Painéis mostram as relações de escala para valores de r_c estimados após um dado número de dias a partir dos primeiros dois casos diários (quatro valores de t_c igualmente espaçados entre 15 dias e o maior valor com pelo menos cinquenta cidades, como indicado nos painéis). **(B)** Relação entre a taxa de crescimento de mortes (r_d) e a população das cidades (P) em escala logarítmica. Painéis mostram as relações de escala para valores de r_d estimados após um dado número de dias a partir das primeiras duas mortes diárias (quatro valores de t_d igualmente espaçados entre 15 dias e o maior valor com pelo menos cinquenta cidades, como indicado nos painéis). Os marcadores em (A) e (B) representam cidades e as linhas tracejadas são as relações de escala com os expoentes de melhor ajuste indicados em cada gráfico (β_{r_c} para a taxa de crescimento de casos e β_{r_d} para a taxa de crescimento de mortes). Todas as taxas foram estimadas usando $\tau = 14$ como definido na Eq. (2.3).

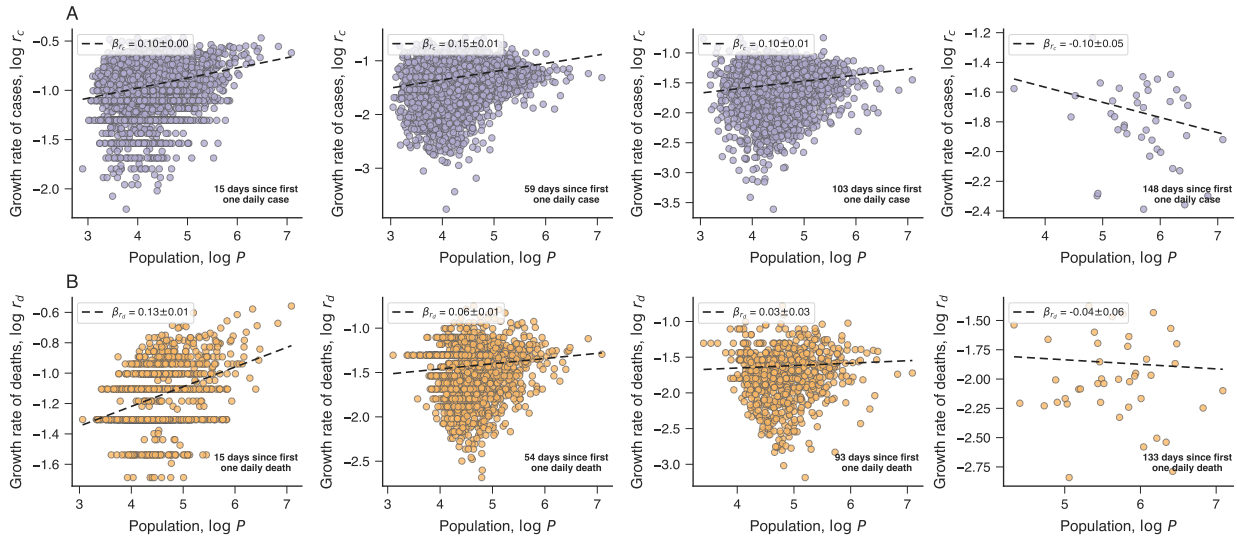

Figura A.32: Relações de escala urbana das taxas de crescimento de casos e mortes por COVID-19. O mesmo que a Figura A.31 mas considerando o primeiro caso diário e a primeira morte diária como pontos de referência.

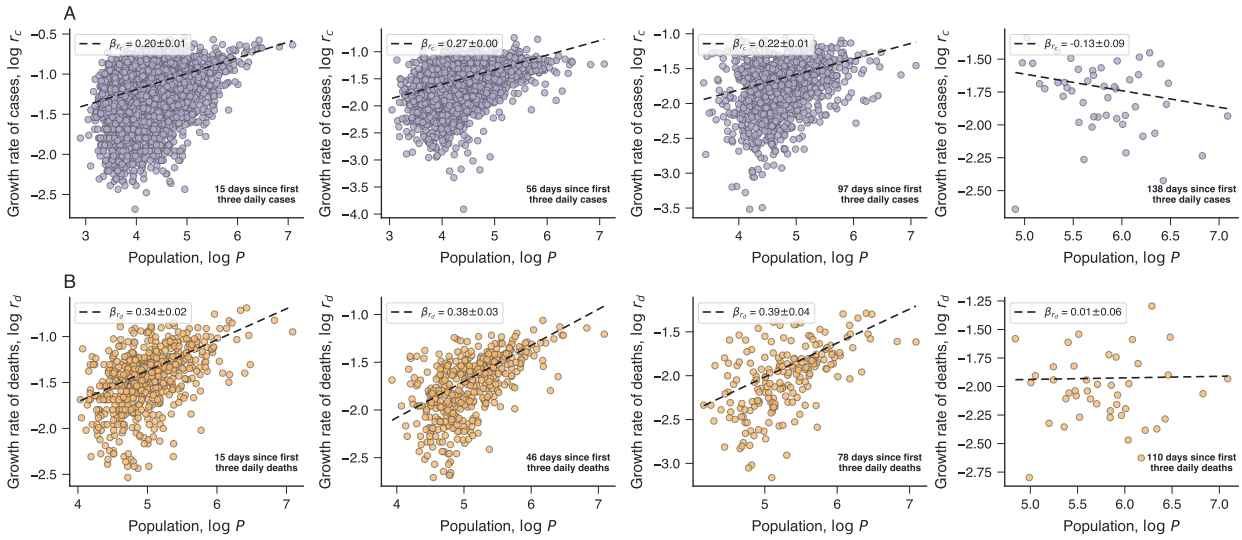

Figura A.33: Relações de escala urbana das taxas de crescimento de casos e mortes por COVID-19. O mesmo que a Figura A.31 mas considerando os primeiros três casos diários e as primeiras três mortes diárias como pontos de referência.

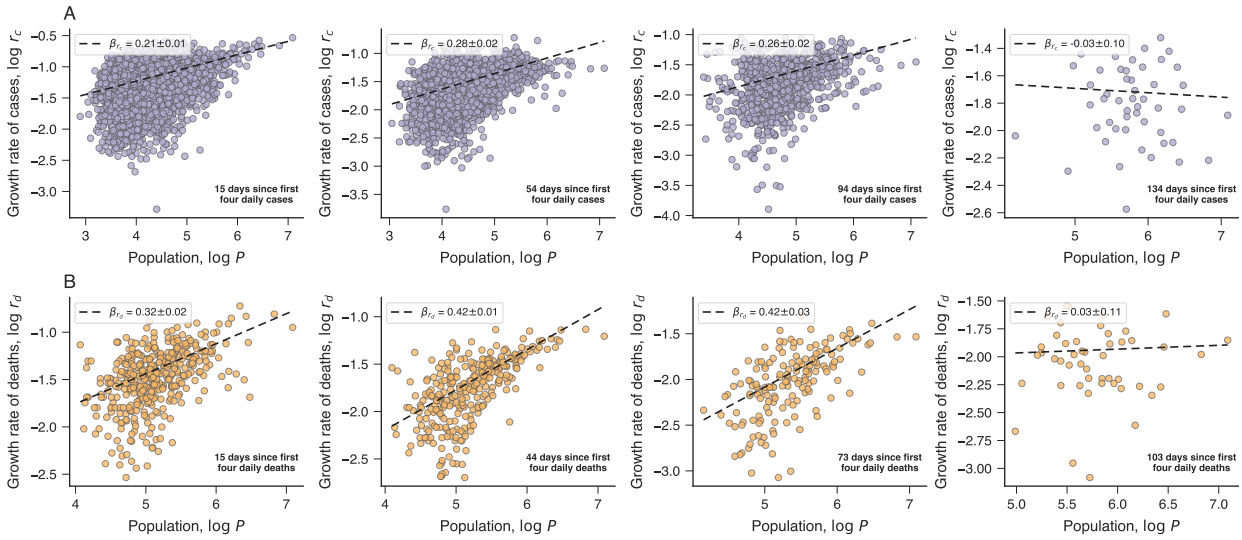

Figura A.34: Relações de escala urbana das taxas de crescimento de casos e mortes por COVID-19. O mesmo que a Figura A.31 mas considerando os primeiros quatro casos diários e as primeiras quatro mortes diárias como pontos de referência.

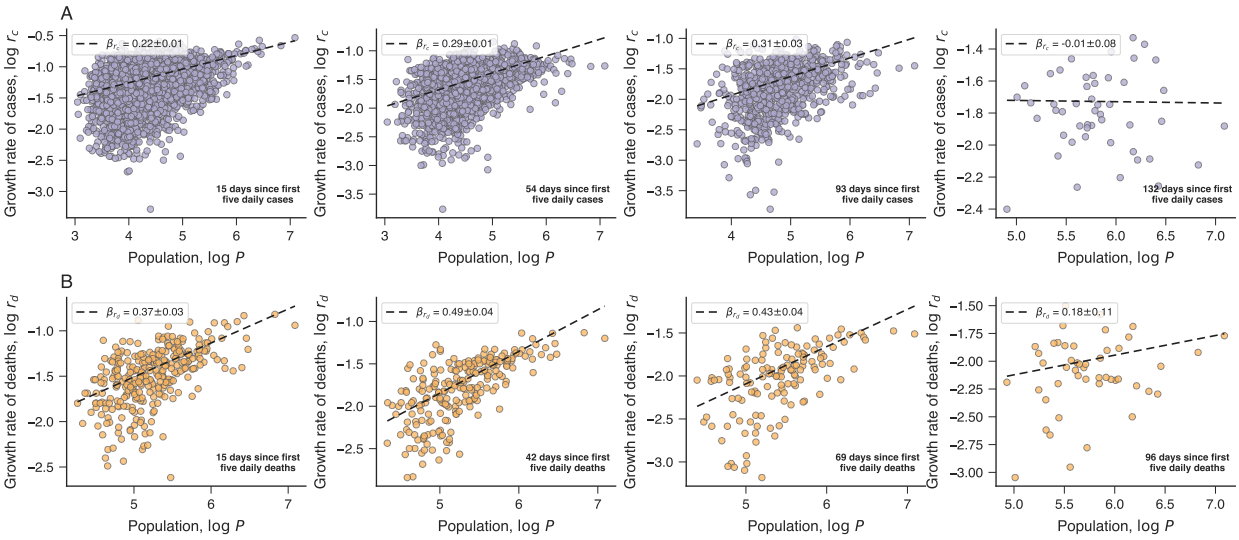

Figura A.35: Relações de escala urbana das taxas de crescimento de casos e mortes por COVID-19. O mesmo que a Figura A.31 mas considerando os primeiros cinco casos diários e as primeiras cinco mortes diárias como pontos de referência.

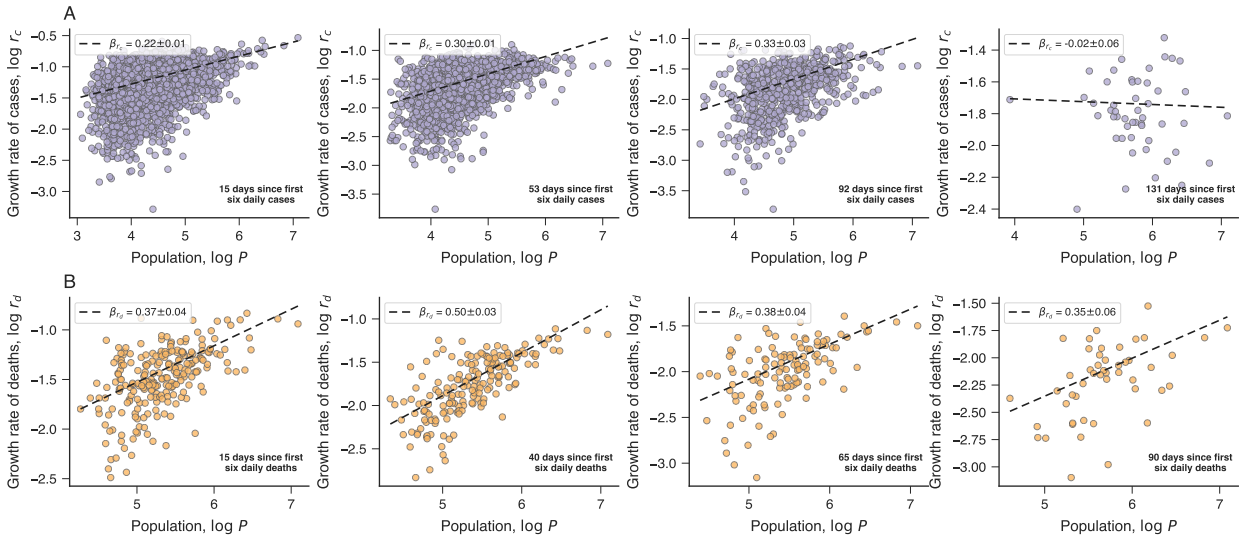

Figura A.36: Relações de escala urbana das taxas de crescimento de casos e mortes por COVID-19. O mesmo que a Figura A.31 mas considerando os primeiros seis casos diários e as primeiras seis mortes diárias como pontos de referência.

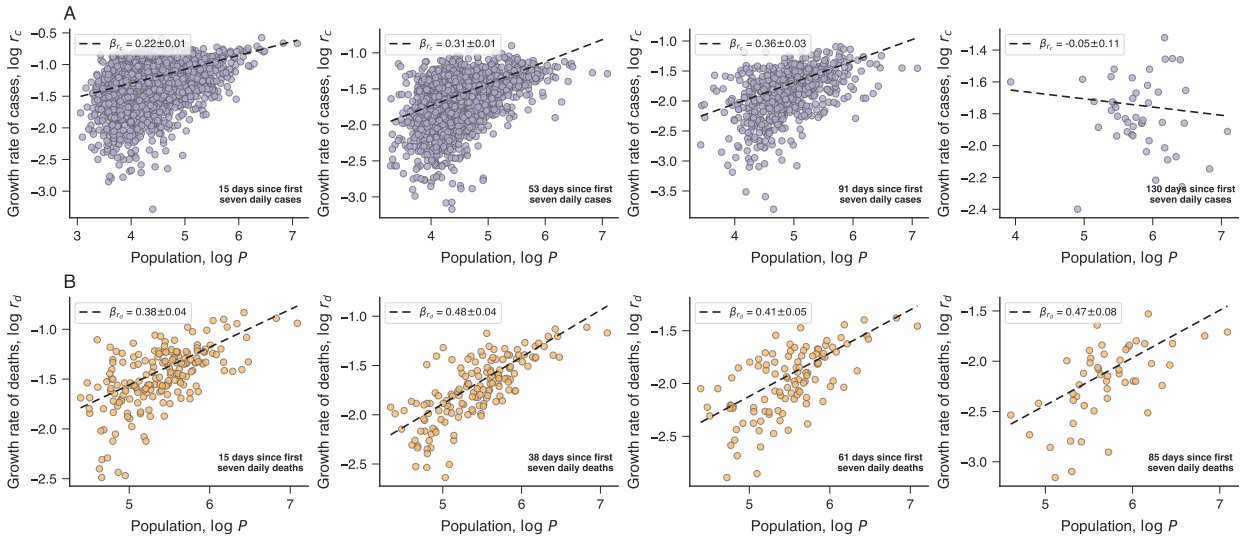

Figura A.37: Relações de escala urbana das taxas de crescimento de casos e mortes por COVID-19. O mesmo que a Figura A.31 mas considerando os primeiros sete casos diários e as primeiras sete mortes diárias como pontos de referência.

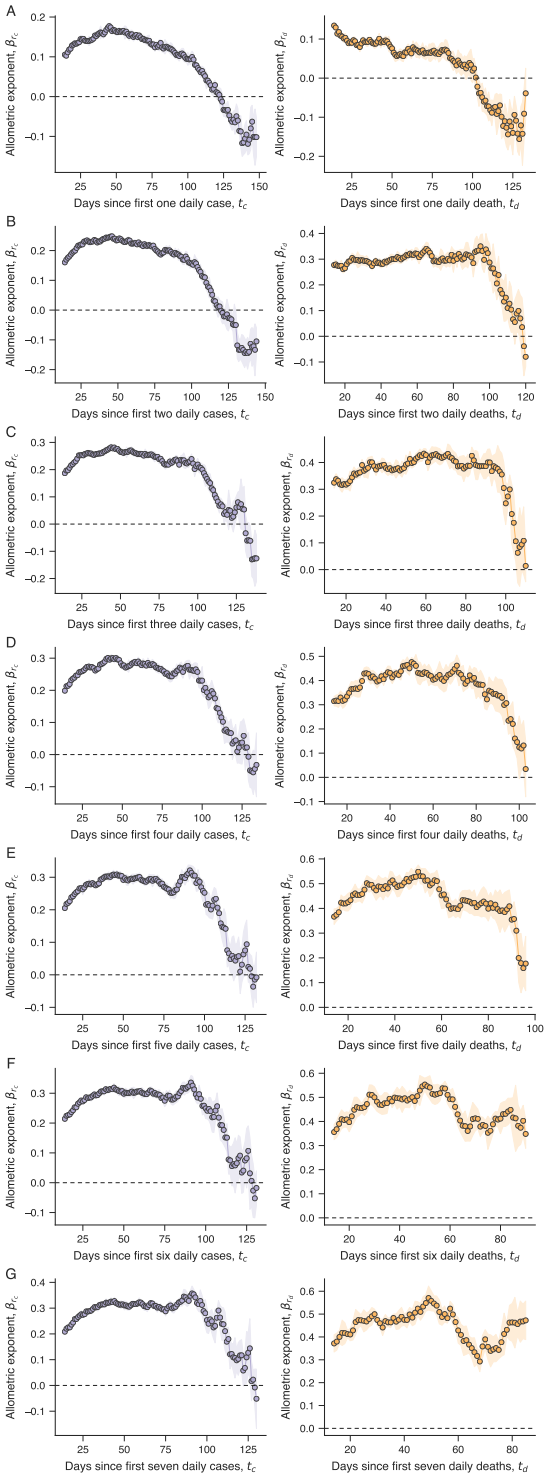

Figura A.38: Dependência temporal dos expoentes de escala para as taxas de crescimento de casos e mortes por COVID-19 sob diferentes escolhas de valores para o número de casos diários ou mortes diárias como ponto de referência. Painéis (A)-(G) mostram a dependência dos expoentes β_{r_c} (esquerda) e β_{r_d} (direita) em relação a t_c e t_d considerando os primeiros 1-7 casos diários e as primeiras 1-7 mortes diárias como pontos de referência. As regiões sombreadas representam o desvio padrão e as linhas tracejadas horizontais representam $\beta_{r_c} = \beta_{r_d} = 0$. Notamos que o comportamento observado para números grandes de ponto de referência tende a seguir o comportamento dos números pequenos na evolução do espalhamento da COVID-19.

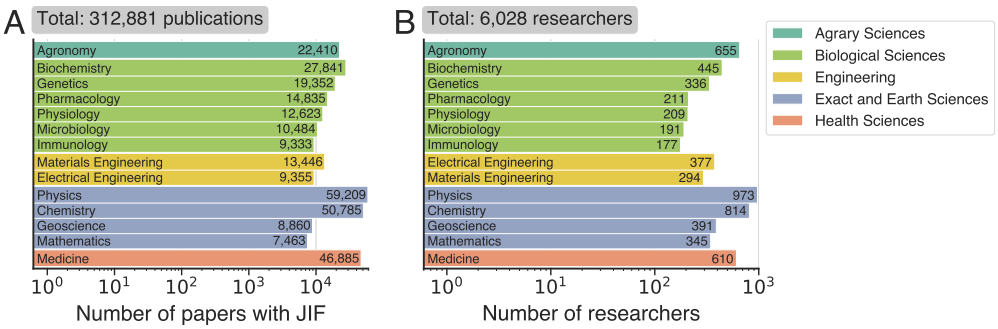

Figura A.39: Número de publicações e pesquisadores no conjunto de dados JIF. O painel **(A)** mostra o número total de artigos e o painel **(B)** mostra o número total de pesquisadores para cada disciplina no conjunto de dados JIF. As cores das barras representam os diferentes campos da ciência em nosso conjunto de dados.

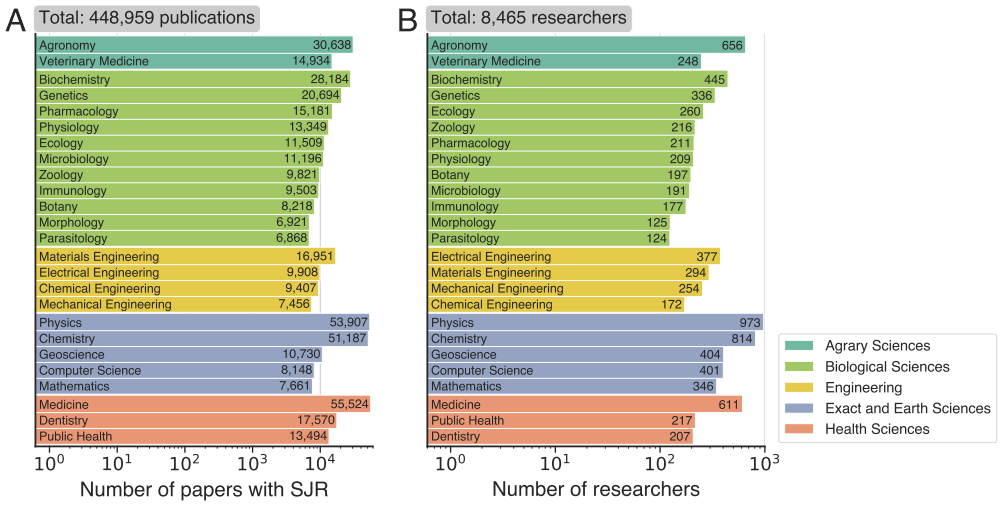

Figura A.40: Número de publicações e pesquisadores no conjunto de dados SJR. O painel **(A)** mostra o número total de artigos e o painel **(B)** mostra o número total de pesquisadores para cada disciplina no conjunto de dados SJR. As cores das barras representam os diferentes campos da ciência em nosso conjunto de dados.

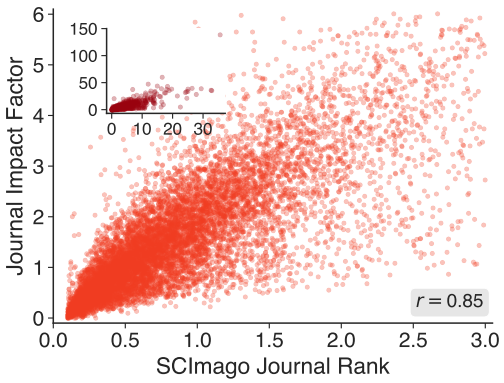

Figura A.41: Fator de impacto de jornais (JIF) e ranque de jornais SCImago (SJR) são correlacionados. Gráfico de dispersão do SJR *versus* JIF para 11.055 jornais presentes em ambos os conjuntos de dados para o ano de 2015. A inserção mostra o gráfico de dispersão considerando o intervalo completo em 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.

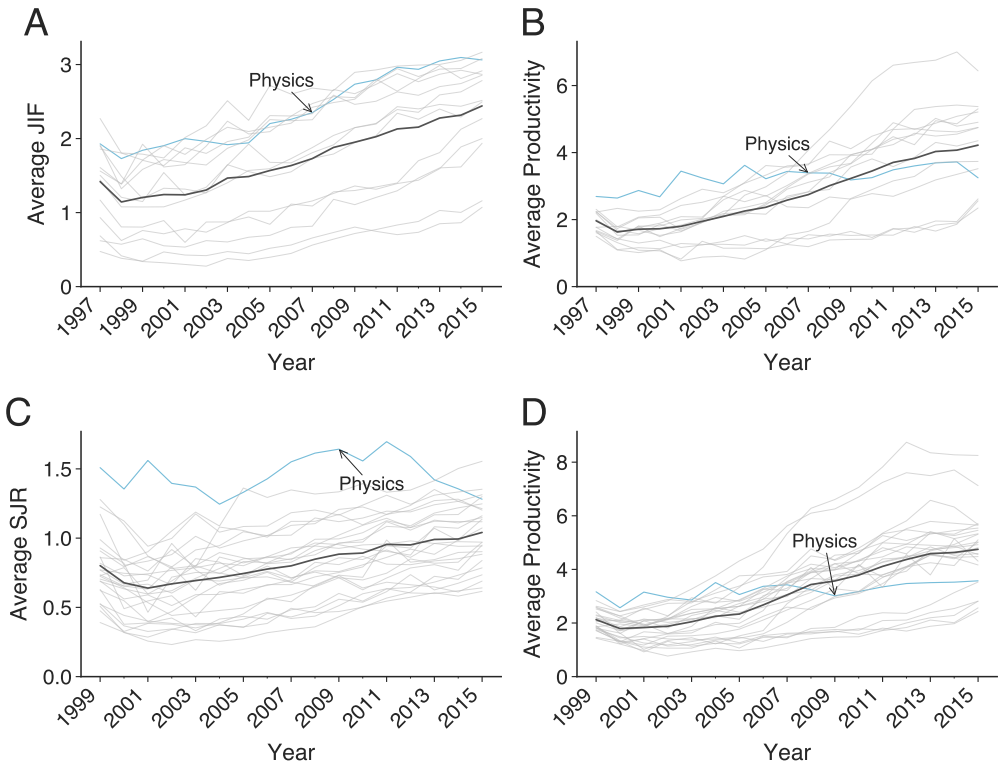

Figura A.42: Evolução temporal do prestígio médio dos jornais e da produtividade. As curvas em cinza mostram a evolução temporal dos valores médios do **(A)** prestígio médio dos jornais e da **(B)** produtividade para o conjunto de dados JIF para todas as disciplinas em nosso estudo. Os painéis **(C)** e **(D)** mostram a mesma informação para o conjunto de dados SJR. As curvas em preto representam o comportamento médio agregado de todas as disciplinas, enquanto as curvas em azul ilustram o comportamento médio da disciplina de física. Os valores médios foram estimados utilizando o estimador de localização Huber.

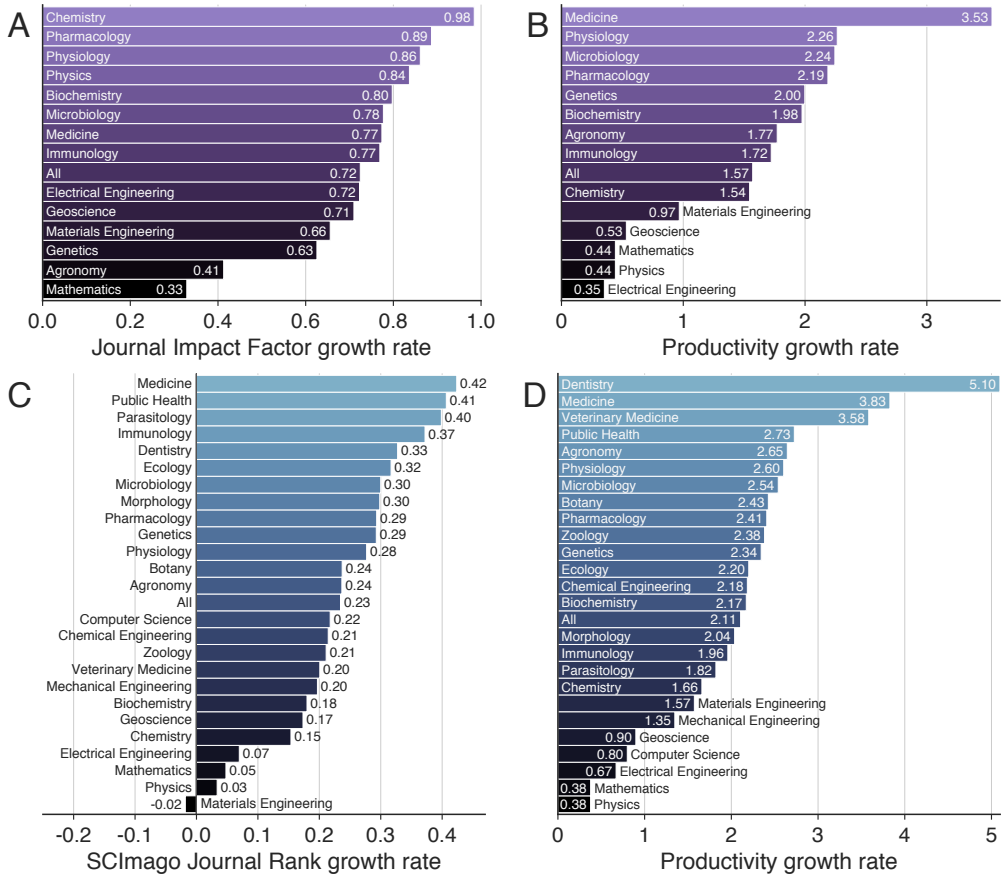

Figura A.43: Taxas de crescimento por década do prestígio médio dos jornais e da produtividade. Os painéis **(A)** e **(B)** mostram as taxas de crescimento por década do prestígio médio dos jornais e da produtividade estimados do conjunto de dados JIF. Painéis **(C)** e **(D)** representam o mesmo para o conjunto de dados SJR. Estimamos as taxas de crescimento ajustando um modelo linear à evolução temporal reportada na Figura A.42 para cada disciplina em cada conjunto de dados. Além disso, estimamos a taxa de crescimento agregando os dados de todas as disciplinas (indicado por "*all*" nos gráficos de barra).

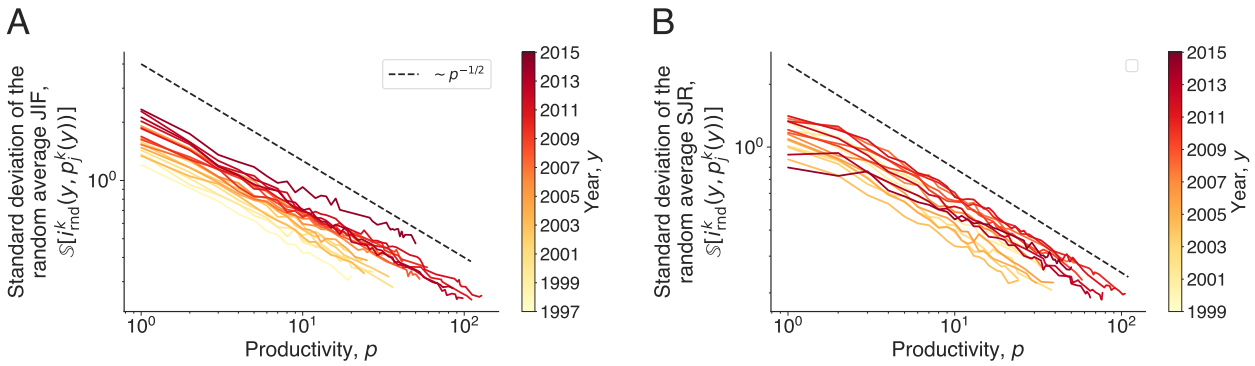

Figura A.44: Efeito do tamanho da produtividade na dispersão do prestígio médio dos jornais. (A) Desvio padrão ($\mathbb{S}[i^k_{\text{rnd}}(y, p^k_j(y))]$) do valor médio do fator de impacto do jornal (JIF) para 1000 amostras aleatórias de p publicações de pesquisadores da física como uma função de 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. O desvio padrão diminui com p, confirmando que baixa produtividade está associada com grande variabilidade. Por outro lado, alta produtividade está associada com baixa variabilidade no prestígio médio dos jornais. O painel **(B)** mostra os mesmos resultados considerando o ranque de jornais SCImago (SJR) como indicador de prestígio de jornal. Um comportamento similar é observado para todos os anos e disciplinas em ambos os conjuntos de dados.

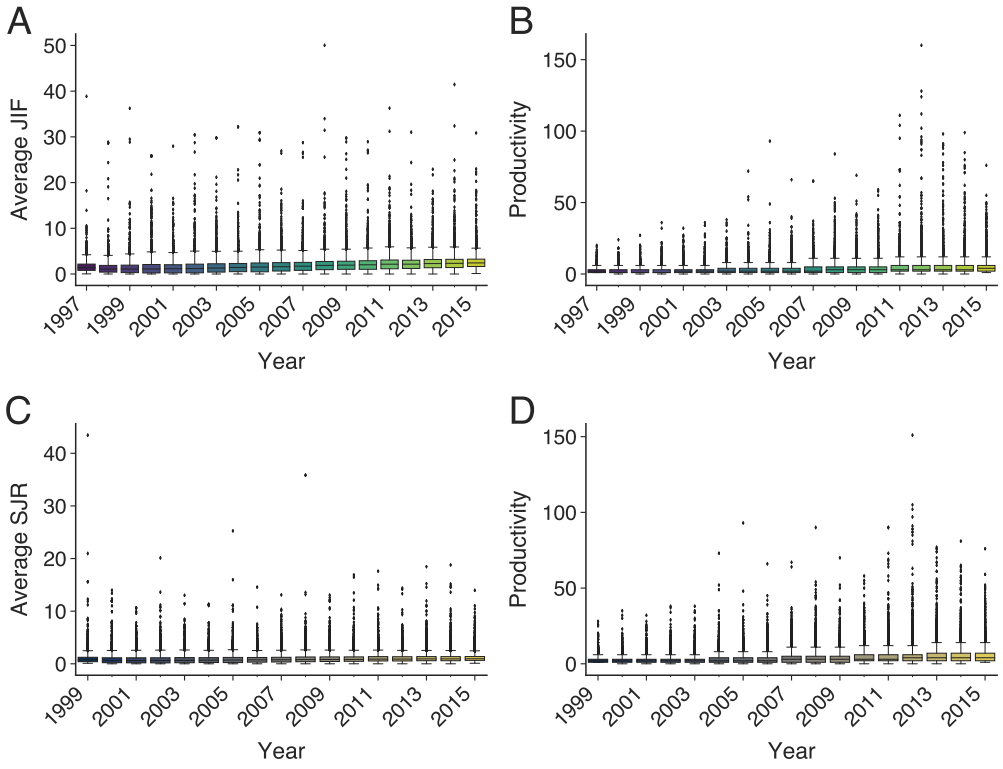

Figura A.45: Valores *outliers* do prestígio médio dos jornais e da produtividade. Os diagramas de caixa retratam o grau de dispersão do **(A)** prestígio médio dos jornais (JIF) e da **(B)** produtividade dos pesquisadores no conjunto de dados JIF em cada ano. Os painéis **(C)** e **(D)** mostram resultados análogos para o conjunto de dados SJR. 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).

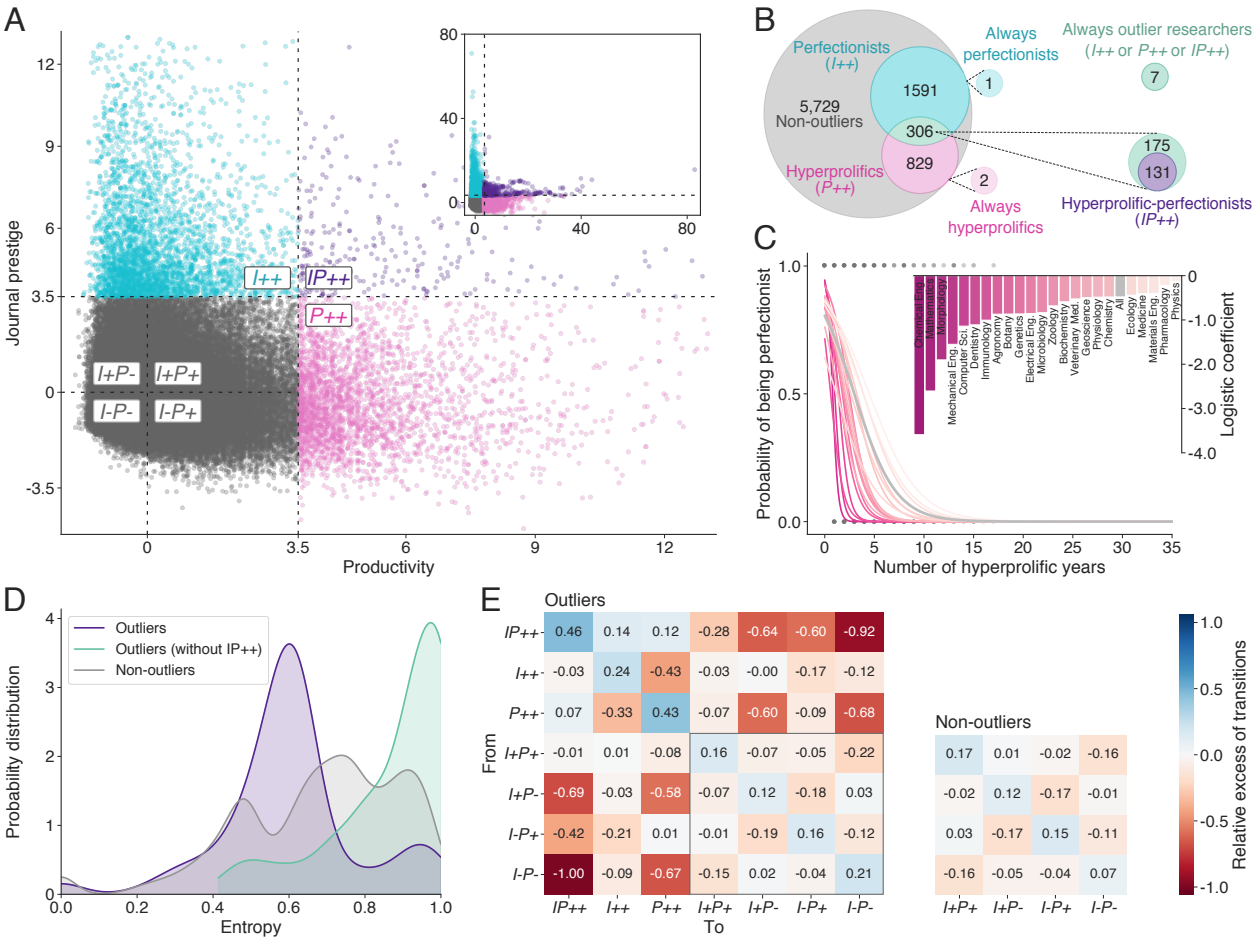

Figura A.46: Prestígio de jornal *versus* produtividade considerando o conjunto de dados SJR. (A) Relação entre impacto médio dos jornais e produtividade em unidades padronizadas (a inserção mostra o intervalo completo do plano). Os marcadores representam anos da carreira de pesquisadores de 25 disciplinas em nosso estudo. **(B)** Diagrama de Venn mostrando o conjunto de relações entre as quatro categorias de pesquisadores. **(C)** Probabilidade de ser um pesquisador perfeccionista tendo um dado 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 agregado para 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 com a ocupação dos setores do plano para as carreiras individuais dos pesquisadores. A curva em roxo mostra os resultados da 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 para pesquisadores *outliers* (esquerda) e não *outliers* (direita). Cada célula representa o excesso relativo de transições entre dois setores comparado com o modelo nulo, que corresponde às versões embaralhadas das carreiras dos pesquisadores para 10.000 realizações.

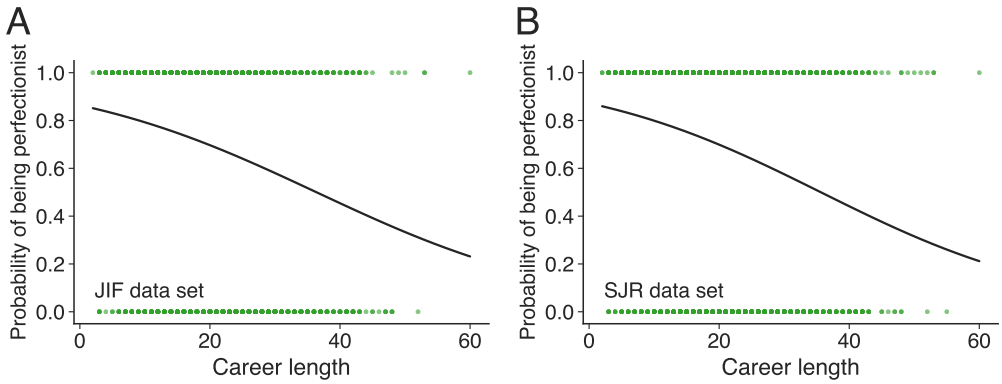

Figura A.47: Efeito do comprimento da carreira na probabilidade de ser perfeccionista. Estimamos a probabilidade de ser perfeccionista como uma função do comprimento da carreira do pesquisador via modelo logístico. O painel **(A)** mostra essa probabilidade para o conjunto de dados JIF e o painel **(B)** mostra a mesma análise para o conjunto de dados SJR. Para o conjunto de dados JIF, a probabilidade de ser perfeccionista decresce de 79% a 58% quando o comprimento de carreira cresce de 10 para 30 anos. Para o conjunto de dados SJR, a probabilidade de ser perfeccionista decresce de 80% a 58% para a mesma variação no comprimento de carreira.

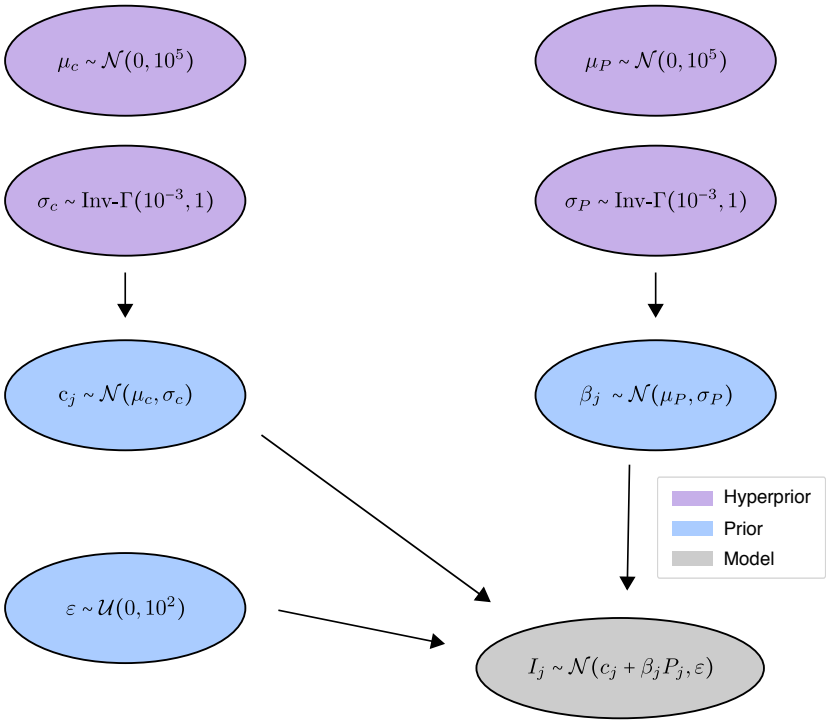

Figura A.48: Representação visual do modelo hierárquico bayesiano definido pela Eq. (3.1). Descrição esquemática do modelo hierárquico bayesiano da Eq. (3.1) usado para estimar o efeito da produtividade no prestígio médio dos jornais para pesquisadores não *outliers*. Formas em roxo representam distribuições a *hiperpriori*, formas em azul representam distribuições a *priori* e a forma em cinza representa a estrutura geral do modelo hierárquico.

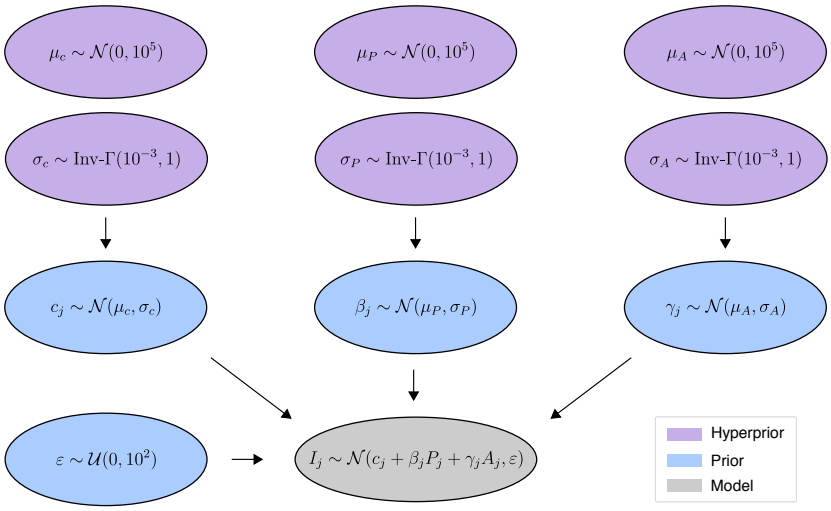

Figura A.49: Representação visual do modelo hierárquico bayesiano com a variável independente de ano da carreira definido pela Eq. (3.3). Descrição esquemática do modelo hierárquico bayesiano da Eq. (3.3) usado para estimar o efeito da produtividade e do ano da carreira no prestígio médio dos jornais 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.

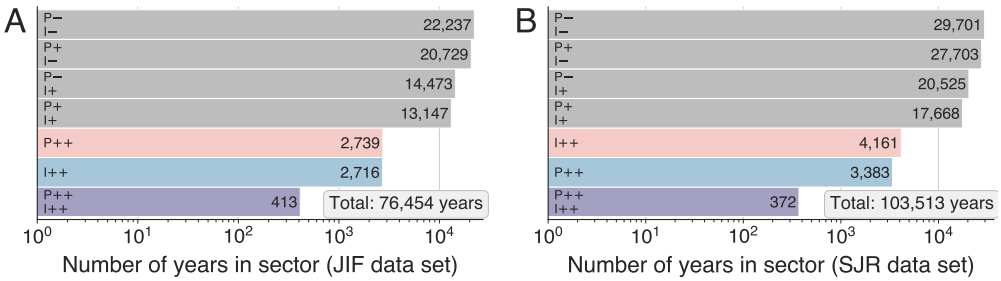

Figura A.50: Demografia do plano prestígio de jornal *versus* produtividade. Os gráficos de barra mostram o número de anos de carreira em cada setor do plano prestígio médio dos jornais *versus* produtividade. O painel **(A)** refere-se ao conjunto de dados JIF e o painel **(B)** refere-se ao conjunto de dados SJR. Notamos que setores não *outliers* são mais povoados do que setores *outliers*. Além disso, o setor $I{-}P{-}$ é o setor mais povoado para ambos os conjunto de dados, enquanto o setor $IP{+}{+}$ é o mais subpovoado.

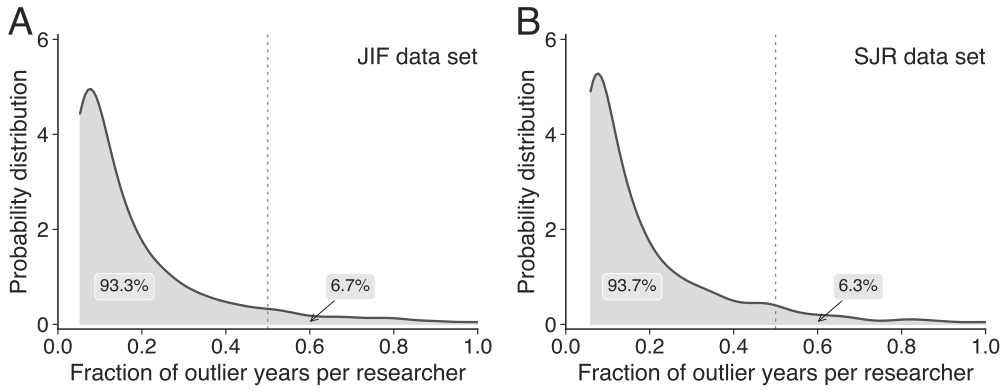

Figura A.51: Anos *outliers* em carreiras científicas. Distribuição de probabilidade da fração de anos *outliers* na carreira de pesquisadores para o **(A)** conjunto de dados JIF e o **(B)** conjunto de dados SJR. Apenas 6,7% dos pesquisadores *outliers* têm mais do que 50% dos anos de carreira dentro de setores *outliers* no conjunto de dados JIF. Para o conjunto de dados SJR, apenas 6,3% dos pesquisadores *outliers* têm mais do que 50% dos anos de carreira dentro de setores *outliers*. Verificamos também que mais do que 47,6% dos pesquisadores são *outliers* em apenas um ano para o conjunto de dados JIF e 48,8% dos pesquisadores para o conjunto de dados SJR. Dessa forma, anos *outliers* são raros em carreiras científicas até mesmo para acadêmicos *outliers*.

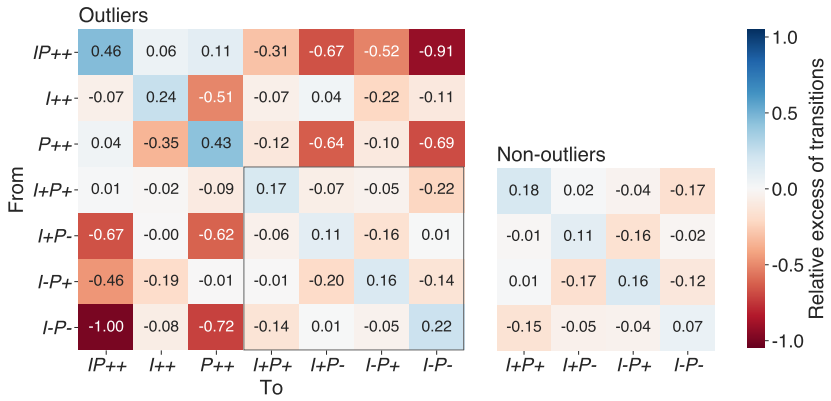

Figura A.52: Matriz de transição entre os setores do plano para o conjunto de dados SJR considerando apenas o conjunto de disciplinas presentes no conjunto de dados JIF. Cada célula representa o excesso relativo de transições entre dois setores comparado com o modelo nulo, que corresponde às versões embaralhadas das carreiras dos pesquisadores para 10.000 realizações. Notamos que os padrões de transições mostrados nesta figura são muito similares àquelas reportadas na Figura A.46E.

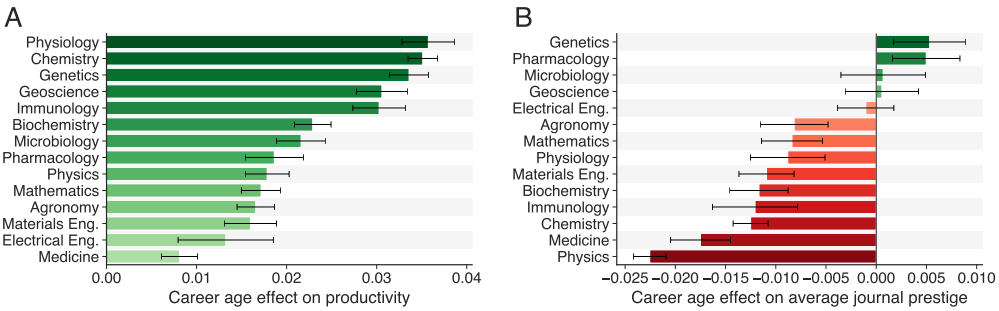

Figura A.53: Efeito da idade da carreira na produtividade e no prestígio médio dos jornais para diferentes disciplinas. Os gráficos de barra mostram o efeito da idade da carreira na **(A)** produtividade e no **(B)** prestígio médio dos jornais para cada disciplina no conjunto de dados JIF. Estimamos os valores por meio de um modelo linear da associação média entre idade da carreira e produtividade e, também, da associação média entre idade da carreira e prestígio médio dos jornais (Figura 3.2) para cada disciplina. As barras de erro indicam o erro padrão dos coeficientes lineares. Observamos uma tendência crescente da produtividade com a progressão da carreira para todas as disciplinas e uma tendência decrescente do prestígio médio dos jornais para a maioria das disciplinas.

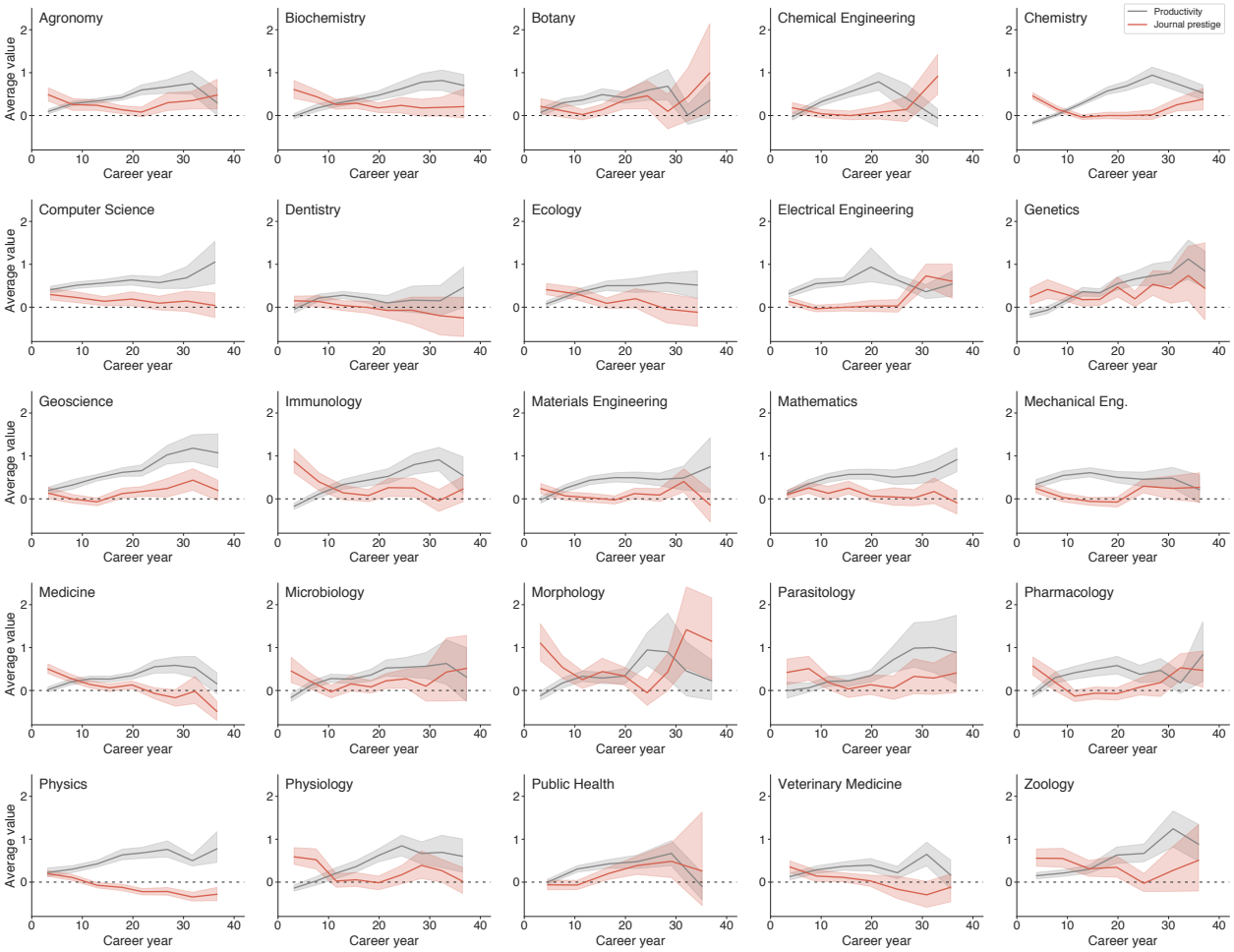

Figura A.54: Valores médios da produtividade e do impacto médio dos jornais ao longo da carreira dos pesquisadores para diferentes disciplinas. Essas visualizações mostram os valores médios da produtividade (curva em cinza) e do prestígio médio dos jornais (curva em vermelho) calculados a partir de médias 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*. A produtividade média aumenta com a progressão da carreira para todas as disciplinas (Figura A.55A) e mostra um platô ou pequeno decréscimo em estágios posteriores da carreira para a maioria das disciplinas. Apesar de algumas disciplinas apresentarem padrões mais complexos, o valor médio do prestígio médio dos jornais mostra uma tendência decrescente sutil e é usualmente maior nos estágios iniciais da carreira para a maioria das disciplinas (Figura A.55B).

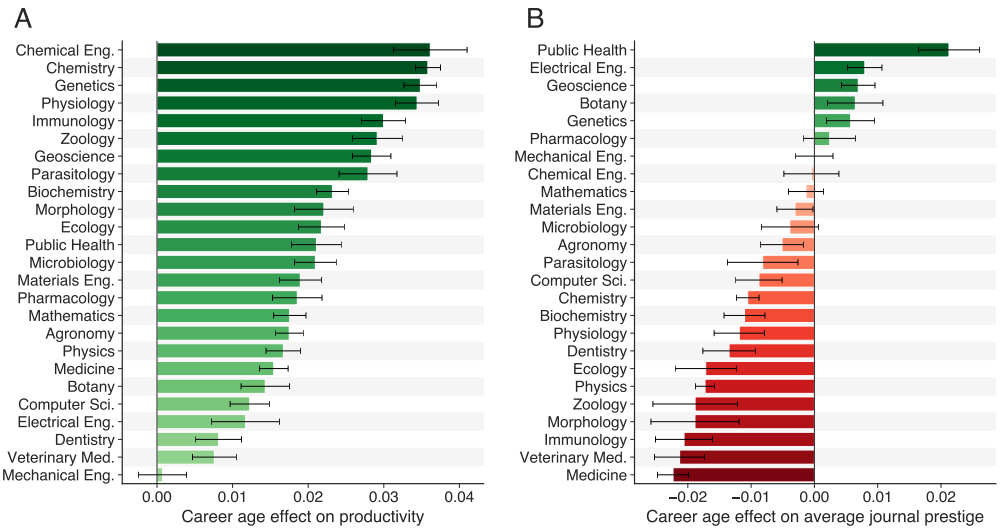

Figura A.55: Efeito da idade da carreira na produtividade e no prestígio médio dos jornais para diferentes disciplinas considerando o conjunto de dados SJR. Os gráficos de barra mostram o efeito da idade da carreira na **(A)** produtividade e no **(B)** prestígio médio dos jornais para cada disciplina no conjunto de dados SJR. Estimamos os valores por meio de um modelo linear da associação média entre idade da carreira e produtividade e, também, da associação média entre idade da carreira e prestígio médio dos jornais (Figura A.54) para cada disciplina. As barras de erro indicam o erro padrão dos coeficientes lineares. Observamos uma tendência crescente da produtividade com a progressão da carreira para todas as disciplinas e uma tendência decrescente do prestígio médio dos jornais para a maioria das disciplinas.

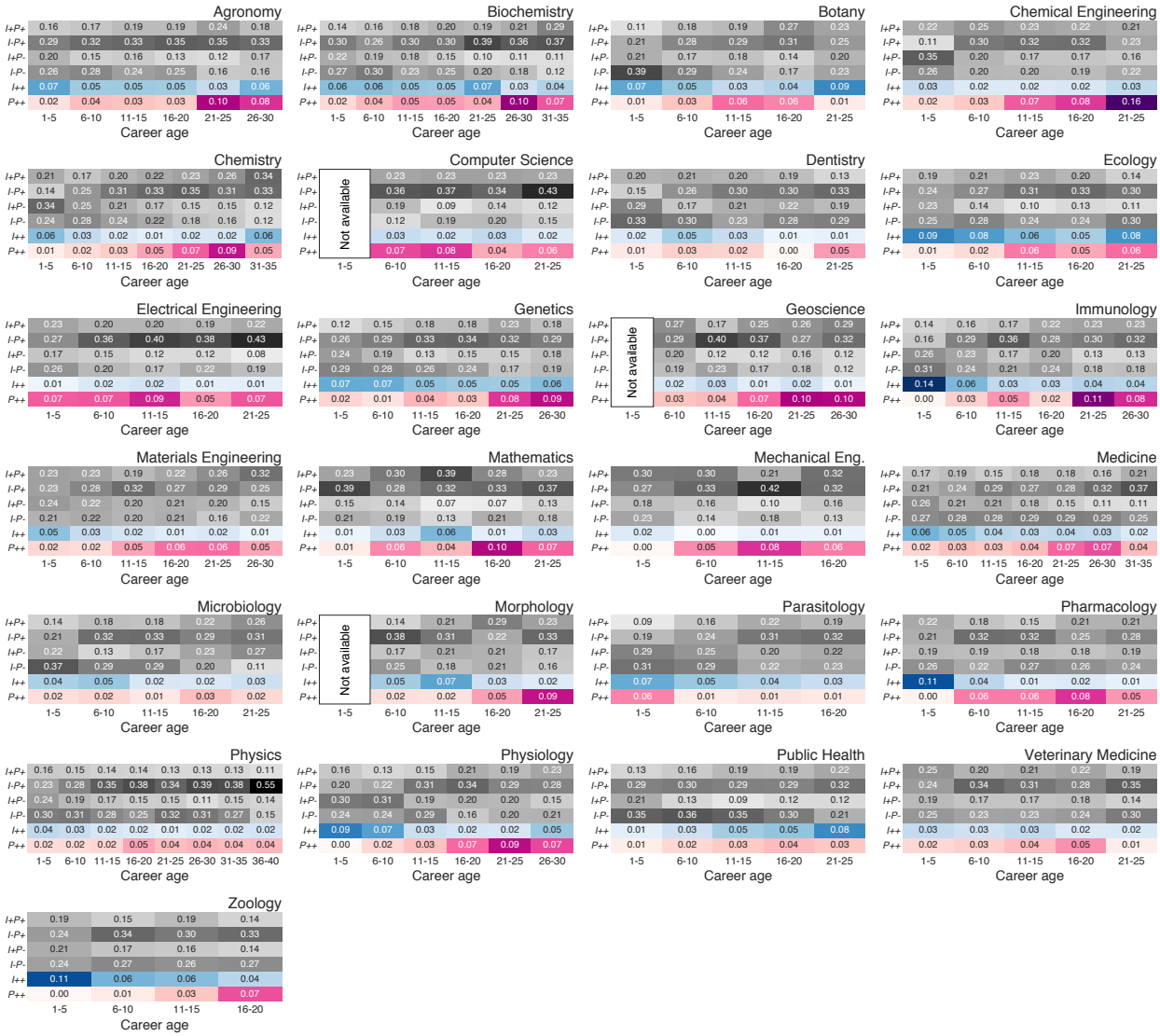

Figura A.56: Tendências de ocupação no plano prestígio de jornal *versus* produtividade ao longo das carreiras dos pesquisadores considerando o conjunto de dados SJR. Os painéis mostram a fração dos anos das carreiras em cada setor não *outlier* e nos setores *outliers* $I{+}{+}$ e $P{+}{+}$ como uma função da idade 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. Os setores de baixa produtividade são mais povoados durante os anos iniciais da carreira. Além disso, há uma tendência de mudança para setores de alta produtividade em estágios posteriores da carreira para a maioria das disciplinas. Apenas intervalos de 5 anos com pelo menos 20 pesquisadores são mostrados nessas visualizações.

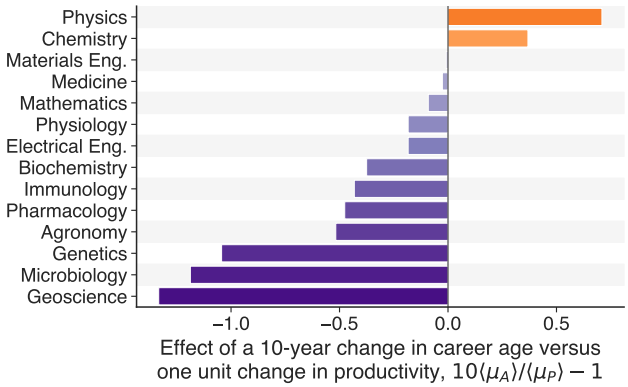

Figura A.57: Comparação entre os efeitos da idade da carreira e produtividade no prestígio médio dos jornais. Gráficos de barra que comparam o efeito de uma progressão de 10 anos na carreira com o efeito de aumentar uma unidade da produtividade (z-score) para um pesquisador típico de cada disciplina no conjunto de dados JIF. Esses valores representam a fração de quão maior ou menor é o efeito da idade 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). Assim, frações ao redor de zero indicam que um aumento de 10 anos na idade da carreira afeta o impacto médio dos jornais 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 médio dos jornais. Para o conjunto de dados JIF, uma progressão de 10 anos na carreira tem efeito maior apenas para química e física.

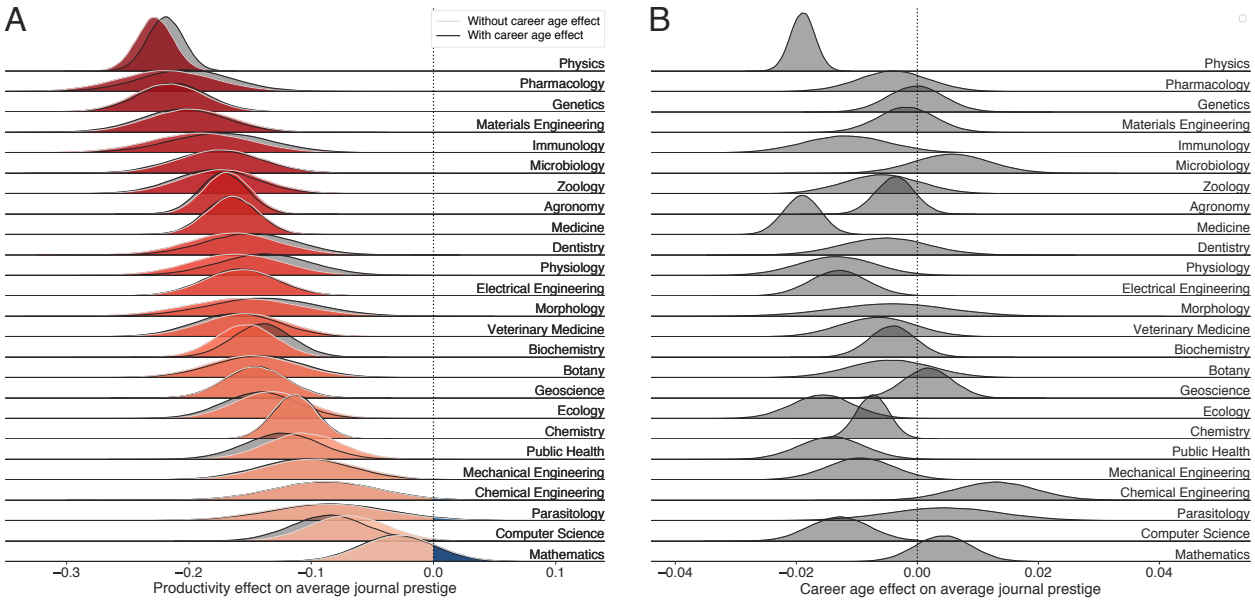

Figura A.58: Efeito da produtividade no prestígio de jornal para pesquisadores não *outliers* considerando o conjunto de dados SJR. (A) Distribuições de probabilidade a *posteriori* do valor médio do coeficiente linear (μ_P) ao considerar a associação entre produtividade e impacto de jornal para pesquisadores não *outliers* de cada disciplina. As curvas preenchidas coloridas representam os resultados sem levar em consideração os efeitos da idade da carreira, enquanto as curvas preenchidas em cinza mostram as distribuições de μ_P após incluir a idade 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 da idade da carreira no impacto médio dos jornais para pesquisadores não *outliers* de cada disciplina.

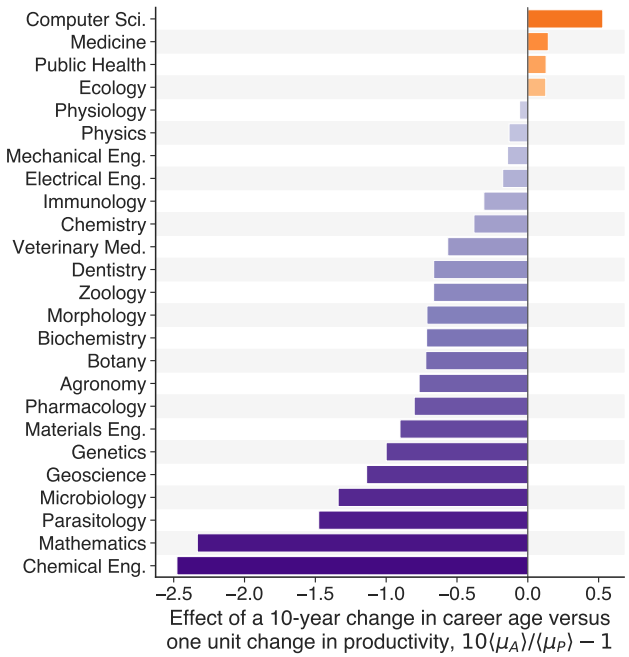

Figura A.59: Comparação entre os efeitos da idade da carreira e produtividade no prestígio médio dos jornais considerando o conjunto de dados SJR. Gráficos de barra que comparam o efeito de uma progressão de 10 anos na carreira com o efeito de aumentar uma unidade da produtividade (z-score) para um pesquisador típico de cada disciplina no conjunto de dados SJR. Esses valores representam a fração de quão maior ou menor é o efeito da idade 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). Assim, frações ao redor de zero indicam que um aumento de 10 anos na idade da carreira afeta o impacto médio dos jornais 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 médio dos jornais. Para o conjunto de dados SJR, uma progressão de 10 anos na carreira tem efeito maior apenas para ciência da computação, ecologia, medicina e saúde coletiva.

APÊNDICE B

Tabelas adicionais

Neste apêndice, apresentamos todas as tabelas adicionais ao texto principal.

Tabela B.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	4523
Bioquímica	258	3482
Engenharia Elétrica	232	2302
Engenharia de Materiais	210	2496
Farmacologia	147	2003
Fisiologia	136	1757
Física	686	9348
Genética	210	2709
Geociências	229	2195
Imunologia	109	1415
Matemática	212	2128
Medicina	357	4765
Microbiologia	131	1670
Química	577	7701

Tabela B.2: **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	4391
Bioquímica	239	3123
Botânica	124	1359
Ciência da Computação	230	2036
Ecologia	160	1821
Engenharia Elétrica	239	2297
Engenharia Mecânica	187	1921
Engenharia Química	124	1536
Engenharia dos Materiais	204	2535
Farmacologia	142	1878
Fisiologia	133	1672
Física	670	8474
Genética	188	2409
Geociências	273	2725
Imunologia	102	1299
Matemática	215	2147
Medicina	361	4983
Medicina Veterinária	178	2138
Microbiologia	131	1698
Morfologia	71	956
Odontologia	151	1937
Parasitologia	72	956
Química	566	7314
Saúde Coletiva	144	1734
Zoologia	126	1418

Referências Bibliográficas

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

[2] Max Roser, H. R. & Ortiz-Ospina, E. Internet. *Our World in Data* (2015). Https://ourworldindata.org/internet.

[3] Uso de internet, televisão e celular no Brasil. Instituto Brasileiro de Geografia e Estatística (IBGE). Available: https://educa.ibge.gov.br/jovens/materias-especiais/20787-uso-de-internet-televisao-e-celular-no-brasil.html. Accessed: 10 May 2022.

[4] Storage. European Council for Nuclear Research (CERN). Available: https://home.cern/science/computing/storage. Accessed: 10 May 2022.

[5] Mitchell, M. *Complexity: A Guided Tour* (Oxford University Press, 2009).

[6] Flake, G. W. *The Computational Beauty of Nature: Computer Explorations of Fractals, Chaos, Complex Systems, and Adaptation* (MIT Press, 1998).

[7] Levin, S. A. Ecosystems and the biosphere as complex adaptive systems. *Ecosystems* **1**, 431–436 (1998).

[8] Ribeiro, H. V., Alves, L. G., Martins, A. F., Lenzi, E. K. & Perc, M. The dynamical structure of political corruption networks. *Journal of Complex Networks* **6**, 989–1003 (2018).

[9] Martins, A. F., da Cunha, B. R., Hanley, Q. S., Gonçalves, S., Perc, M. & Ribeiro, H. V. Universality of political corruption networks. *Scientific reports* **12**, 1–10 (2022).

[10] Sigaki, H. Y., de Souza, R., de Souza, R., Zola, R. & Ribeiro, H. Estimating physical properties from liquid crystal textures via machine learning and complexity-entropy methods. *Physical Review E* **99**, 013311 (2019).

[11] Sigaki, H. Y., Lenzi, E. K., Zola, R. S., Perc, M. & Ribeiro, H. V. Learning physical properties of liquid crystals with deep convolutional neural networks. *Scientific reports* **10**, 1–10 (2020).

[12] Pessa, A. A., Zola, R. S., Perc, M. & Ribeiro, H. V. Determining liquid crystal properties with ordinal networks and machine learning. *Chaos, Solitons & Fractals* **154**, 111607 (2022).

[13] Hanley, Q. S., Lewis, D. & Ribeiro, H. V. Rural to urban population density scaling of crime and property transactions in english and welsh parliamentary constituencies. *PloS one* **11**, e0149546 (2016).

[14] Ribeiro, H. V., Rybski, D. & Kropp, J. P. Effects of changing population or density on urban carbon dioxide emissions. *Nature communications* **10**, 1–9 (2019).

[15] Sutton, J., Shahtahmassebi, G., Ribeiro, H. V. & Hanley, Q. S. Rural–urban scaling of age, mortality, crime and property reveals a loss of expected self-similar behaviour. *Scientific reports* **10**, 1–13 (2020).

[16] Alves, L. G., Rybski, D. & Ribeiro, H. V. Commuting network effect on urban wealth scaling. *Scientific reports* **11**, 1–10 (2021).

[17] Ribeiro, H. V., Oehlers, M., Moreno-Monroy, A. I., Kropp, J. P. & Rybski, D. Association between population distribution and urban gdp scaling. *Plos one* **16**, e0245771 (2021).

[18] Sigaki, H. Y., Perc, M. & Ribeiro, H. V. Clustering patterns in efficiency and the coming-of-age of the cryptocurrency market. *Scientific reports* **9**, 1–9 (2019).

[19] Alves, L. G., Sigaki, H. Y., Perc, M. & Ribeiro, H. V. Collective dynamics of stock market efficiency. *Scientific reports* **10**, 1–10 (2020).

[20] Sigaki, H. Y., Perc, M. & Ribeiro, H. V. History of art paintings through the lens of entropy and complexity. *Proceedings of the National Academy of Sciences* **115**, E8585–E8594 (2018).

[21] Vieira, D. S., Picoli, S. & Mendes, R. S. Robustness of sentence length measures in written texts. *Physica A: Statistical mechanics and its applications* **506**, 749–754 (2018).

[22] 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).

[23] Vieira, D. S., Riveros, J. M., Jauregui, M. & Mendes, R. S. Anomalous diffusion behavior in parliamentary presence. *Physical Review E* **99**, 042141 (2019).

[24] Cardoso, M., Mendes, R., Souza, J. & Ribeiro, H. Gender difference in candidature processes for brazilian elections. *Physica A: Statistical Mechanics and its Applications* **537**, 122525 (2020).

[25] Vieira, D. S., García-Girón, J., Heino, J., Toivanen, M., Helm, A. & Alahuhta, J. Little evidence of range size conservatism in freshwater plants across two continents. *Journal of Biogeography* **48**, 1200–1212 (2021).

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

[27] Petrobras descobre novo poço de petróleo em área de pré-sal no Rio. Available: https://www.cnnbrasil.com.br/business/petrobras-descobre-novo-poco-de-petroleo-em-area-de-pre-sal-no-rio/. Accessed: 10 May 2022.

[28] A pandemia em dados: a Covid-19 decifrada para a sociedade. Conexão Ciência. UEM. Available: https://conexaociencia.com.br/em-meio-ao-caos-fisicos-encontram-uma-area-de-afinidade/#. Accessed: 10 May 2022.

[29] Nishiura, H., Linton, N. M. & Akhmetzhanov, A. R. Serial interval of novel coronavirus (covid-19) infections. *International journal of infectious diseases* **93**, 284–286 (2020).

[30] Cori, A., Ferguson, N. M., Fraser, C. & Cauchemez, S. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American journal of epidemiology* **178**, 1505–1512 (2013).

[31] 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).

[32] 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).

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

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

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

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

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

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

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

[40] 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).

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

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

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

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

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

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

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

[48] 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).

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

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

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

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

[53] 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).

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

[55] 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).

[56] Andrieu, C. & Thoms, J. A tutorial on adaptive mcmc. *Statistics and computing* **18**, 343–373 (2008).

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

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

[59] PyMC3 – Plots. Available: https://pymc3-testing.readthedocs.io/en/rtd-docs/api/plots.html. Accessed: 7 June 2022.

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

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

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

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

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

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

[66] Hogg, R. V., McKean, J. W. & Craig, A. T. *Introduction to Mathematical Statistics* (Pearson, 2005).

[67] United Nations, World Urbanization Prospects: Urban population (% of total). Available: http://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS. Accessed: 27 May 2020.

[68] Jiang, L. & O'Neill, B. C. Global urbanization projections for the Shared Socioeconomic Pathways. *Global Environ. Chang.* **42**, 193–199 (2017).

[69] International Civil Aviation Organization, Civil Aviation Statistics of the World and ICAO staff estimates. Available: https://data.worldbank.org/indicator/IS.AIR.PSGR. Accessed: 27 May 2020.

[70] Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L. & Daszak, P. Global trends in emerging infectious diseases. *Nature* **451**, 990–993 (2008).

[71] Wolfe, N. D., Dunavan, C. P. & Diamond, J. Origins of major human infectious diseases. *Nature* **447**, 279–283 (2007).

[72] Xiao, K. *et al.* Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature* **583**, 286–289 (2020).

[73] World Health Organization, Coronavirus disease (COVID-19) Situation Report - 209. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200816-covid-19-sitrep-209.pdf?sfvrsn=5dde1ca2_2. Accessed: 19 Aug 2020.

[74] Maier, B. F. & Brockmann, D. Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. *Science* **368**, 742–746 (2020).

[75] Vasconcelos, G. L., Macêdo, A. M. S., Ospina, R., Almeida, F. A. G., Duarte-Filho, G. C. & Souza, I. C. L. Modelling fatality curves of COVID-19 and the effectiveness of intervention strategies. *medRxiv* (2020).

[76] Moghadas, S. M. *et al.* Projecting hospital utilization during the COVID-19 outbreaks in the United States. *Proceedings of the National Academy of Sciences* **117**, 9122–9126 (2020).

[77] Gatto, M., Bertuzzo, E., Mari, L., Miccoli, S., Carraro, L., Casagrandi, R. & Rinaldo, A. Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. *Proceedings of the National Academy of Sciences* **117**, 10484–10491 (2020).

[78] Dowd, J. B., Andriano, L., Brazel, D. M., Rotondi, V., Block, P., Ding, X., Liu, Y. & Mills, M. C. Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proceedings of the National Academy of Sciences* **117**, 9696–9698 (2020).

[79] Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400 (2020).

[80] West, R., Michie, S., Rubin, G. J. & Amlôt, R. Applying principles of behaviour change to reduce SARS-CoV-2 transmission. *Nature Human Behaviour* **4**, 451–459 (2020).

[81] Walker, P. G. *et al.* The impact of COVID-19 and strategies for mitigation and suppression in low-and middle-income countries. *Science* **369**, 413–422 (2020).

[82] Flaxman, S. *et al.* Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020).

[83] Block, P., Hoffman, M., Raabe, I. J., Dowd, J. B., Rahal, C., Kashyap, R. & Mills, M. C. Social network-based distancing strategies to flatten the COVID-19 curve in a post-lockdown world. *Nature Human Behaviour* **4**, 588–596 (2020).

[84] Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C. & West, G. B. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences* **104**, 7301–7306 (2007).

[85] Bettencourt, L. M. The origins of scaling in cities. *Science* **340**, 1438–1441 (2013).

[86] Batty, M. *The New Science of Cities* (MIT Press, Cambridge, MA, 2013).

[87] West, G. B. *Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies* (Penguin, New York, 2017).

[88] Youn, H., Bettencourt, L. M., Lobo, J., Strumsky, D., Samaniego, H. & West, G. B. Scaling and universality in urban economic diversification. *Journal of The Royal Society Interface* **13**, 20150937 (2016).

[89] Gao, J., Zhang, Y.-C. & Zhou, T. Computational socioeconomics. *Physics Reports* **817**, 1–104 (2019).

[90] Acuna-Soto, R., Viboud, C. & Chowell, G. Influenza and pneumonia mortality in 66 large cities in the United States in years surrounding the 1918 pandemic. *PLoS One* **6**, e23467 (2011).

[91] Antonio, F. J., de Picoli Jr, S., Teixeira, J. J. V. & dos Santos Mendes, R. Growth patterns and scaling laws governing AIDS epidemic in Brazilian cities. *PLoS ONE* **9**, e111015 (2014).

[92] Melo, H. P. M., Moreira, A. A., Batista, É., Makse, H. A. & Andrade, J. S. Statistical signs of social influence on suicides. *Scientific Reports* **4**, 1–6 (2014).

[93] Rocha, L. E., Thorson, A. E. & Lambiotte, R. The non-linear health consequences of living in larger cities. *Journal of Urban Health* **92**, 785–799 (2015).

[94] Schläpfer, M., Bettencourt, L. M., Grauwin, S., Raschke, M., Claxton, R., Smoreda, Z., West, G. B. & Ratti, C. The scaling of human interactions with city size. *Journal of the Royal Society Interface* **11**, 20130789 (2014).

[95] Stier, A. J., Berman, M. G. & Bettencourt, L. COVID-19 attack rate increases with city size. *arXiv:2003.10376* (2020).

[96] Cardoso, B.-H. F. & Gonçalves, S. Urban scaling of COVID-19 epidemics. *arXiv:2005.07791* (2020).

[97] Delatorre, E., Mir, D., Graf, T. & Bello, G. Tracking the onset date of the community spread of SARS-CoV-2 in Western Countries. *medRxiv* (2020).

[98] Brasil.io – Boletins informativos e casos do coronavírus por município por dia. Available: https://brasil.io/dataset/covid19/caso/. Accessed: 27 May 2020.

[99] Brazil's Public healthcare System (SUS), Department of Data Processing (DATASUS). Available: http://datasus.saude.gov.br. Accessed: 27 May 2020.

[100] Leitao, J. C., Miotto, J. M., Gerlach, M. & Altmann, E. G. Is this scaling nonlinear? *Royal Society Open Science* **3**, 150649 (2016).

[101] Souch, J. M. & Cossman, J. S. A commentary on rural-urban disparities in COVID-19 testing rates per 100,000 and risk factors. *The Journal of Rural Health* (2020).

[102] Hsiang, S. *et al.* The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature* **584**, 262–267 (2020).

[103] Gao, J., Yin, Y., Jones, B. F. & Wang, D. Quantifying policy responses to a global emergency: Insights from the COVID-19 pandemic. *arXiv:2006.13853* (2020).

[104] Brandtner, C., Bettencourt, L., Stier, A. & Berman, M. G. Creatures of the state? Metropolitan counties compensated for state inaction in initial US response to COVID-19 pandemic. *Mansueto Institute for Urban Innovation Research Paper* (2020).

[105] Grasselli, G., Pesenti, A. & Cecconi, M. Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: Early experience and forecast during an emergency response. *JAMA* **323**, 1545–1546 (2020).

[106] Arabi, Y. M., Murthy, S. & Webb, S. COVID-19: A novel coronavirus and a novel challenge for critical care. *Intensive Care Medicine* **46**, 833–836 (2020).

[107] Castro, M. C. *et al.* Brazil's unified health system: the first 30 years and prospects for the future. *The Lancet* **394**, 345–356 (2019).

[108] Verity, R. *et al.* Estimates of the severity of coronavirus disease 2019: A model-based analysis. *The Lancet Infectious Diseases* **20**, 669–677 (2020).

[109] Heroy, S. Metropolitan-scale COVID-19 outbreaks: How similar are they? *arXiv:2004.01248* (2020).

[110] 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).

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

[112] 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).

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

[114] 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).

[115] Wilsdon, J. *The Metric Tide: Independent review of the role of metrics in research assessment and management* (Sage, 2016).

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

[117] 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).

[118] 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).

[119] 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).

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

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

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

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

[124] 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).

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

[126] 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).

[127] 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).

[128] San Francisco Declaration on Research Assessment (2012). Available at https://sfdora.org/read/. Accessed August 2020.

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

[130] Nuffield Council on Bioethics. The findings of a series of engagement activities exploring the culture of scientific research in the UK (2014). Available at https://www.nuffieldbioethics.org/assets/pdfs/The-culture-of-scientific-research-report.pdf. Accessed September 2020.

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

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

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

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

[135] 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).

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

[137] 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).

[138] 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).

[139] 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).

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

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

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

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

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

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

[146] 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).

[147] Larivière, V. & Sugimoto, C. R. *The Journal Impact Factor: A Brief History, Critique, and Discussion of Adverse Effects*, 3–24 (Springer, 2019).

[148] 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).

[149] RESOLUÇÃO N.º 058/2020-CAD - Universidade Estadual de Maringá. http://www.drh.uem.br/res/Resolu%C3%A7%C3%A3o-058-2020-CAD.pdf. [Último acesso 10 de junho de 2022].

[150] Chamada CNPq N° 06/2019 - Bolsas de Produtividade em Pesquisa. http://memoria.cnpq.br/chamadas-publicas?p_p_id=resultadosportlet_WAR_resultadoscnpqportlet_INSTANCE_0ZaM&filtro=encerradas&detalha=chamadaDivulgada&idDivulgacao=8722. [Último acesso 12 de dezembro de 2019].

[151] 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).

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

[153] 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).

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

[155] Waltman, L. & Traag, V. A. Use of the journal impact factor for assessing individual articles need not be statistically wrong. *F1000Research* **9**, 366 (2020).

[156] 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).

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

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

[159] 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).

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

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

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

[163] statsmodels (2020). Available at https://www.statsmodels.org. Accessed August 2020.

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

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

[166] Price, M. Some scientists publish more than 70 papers a year. Here's how – and why – they do it (2018). Accessed September 2020.