



10° EBBC

Encontro Brasileiro de
Bibliometria e Cientometria



DO ARXIV AO PERIÓDICO: fatores associados ao registro de *preprints* como publicações formais

Joel Perozo-Vasquez

<https://orcid.org/0000-0002-1223-8927>

ufscjoel@gmail.com

Universidade Federal de Santa Catarina
(UFSC)

Marcos André Braz Vaz

<https://orcid.org/0000-0002-9135-7763>

braz.vaz@ufsc.br

Universidade Federal de Santa Catarina
(UFSC)

Thiago Magela Rodrigues Dias

<https://orcid.org/0000-0001-5057-9936>

thiagomagela@gmail.com

Universidade Federal de Santa Catarina
(UFSC)

Resumo:

Objetivo: Analisar possíveis determinantes associados à conversão de preprints depositados no arXiv (2010-2025) em artigos publicados nas subáreas “Bibliotecas Digitais” (cs.DL) e “Redes Sociais e Informação” (cs.SI). Metodologia: Estudo cientométrico inferencial (23.136 registros), empregando correlação de Spearman e regressão logística. Resultados: Os resultados sugerem associação positiva entre o número de revisões e a probabilidade de registro de publicação formal, estimando um aumento de 18,3% (cs.DL) e 18,7% (cs.SI) na probabilidade de publicação para cada nova versão depositada. Conclusão: O esforço de revisão iterativo apresentou-se como a variável de maior associação com o sucesso editorial, indicando o repositório como espaço ativo de maturação do conhecimento.

Palavras-Chave: *Preprints*. Comunicação Científica. Cientometria. Sucesso Editorial. Avaliação por Pares.

EIXO TEMÁTICO:

Acesso Aberto, Ciência Aberta e Dados Abertos

MODALIDADE:

Resumo expandido

DOI: 10.22477/x.ebbc.1085

COMO CITAR ESTE ARTIGO:

PEROZO-VASQUEZ, Joel; VAZ, Marcos André Braz; DIAS, Thiago Magela Rodrigues. Do ArXiv ao Periódico: fatores associados ao registro de preprints como publicações formais. In: EBBC - Encontro Brasileiro de Bibliometria e Cientometria. 10. **Anais [...]**. Curitiba, 2026. DOI: 10.22477/x.ebbc.1085. Disponível em: <https://doi.org/10.22477/x.ebbc.1085>.



1 INTRODUÇÃO

No ecossistema da Ciência Aberta (*Open Science*), os *preprints* aceleram o ciclo da pesquisa ao permitir o compartilhamento imediato de resultados e a validação comunitária prévia à revisão por pares formal (Waltman *et al.*, 2021). Repositórios de *preprints* não substituem os periódicos, mas estabelecem uma “simbiose editorial” (Larivière *et al.*, 2014): o repositório garante visibilidade e precedência, enquanto o periódico ou anais de evento provê a certificação formal via Identificador de Objeto Digital (DOI). Essa dinâmica expõe o manuscrito ao escrutínio público, permitindo seu amadurecimento inicial (Nielsen, 2012).

A disseminação de *preprints* tem impactado a dinâmica da comunicação científica, no entanto, essa adoção apresenta fortes assimetrias disciplinares (Wang *et al.*, 2020). Enquanto campos como a física e subáreas da ciência da computação possuem uma tradição consolidada de uso de repositórios como o arXiv, outras áreas do conhecimento apresentam resistências, variações institucionais e ritmos distintos de aceitação. Nesse contexto, compreender os fatores que se associam à transição de *preprints* para publicações formais torna-se relevante para os estudos de cientometria, desde que se observem as especificidades dos diferentes regimes de validação científica.

Neste cenário, a conversão de um *preprint* em publicação formal representa o que convencionamos chamar de “sucesso editorial”. O problema central desta pesquisa reside em compreender a complexidade dos fatores que se correlacionam com essa transição, utilizando a plataforma arXiv como ambiente de observação.

Portanto, este artigo objetiva analisar empiricamente associações estatísticas entre características dos *preprints* e sua posterior vinculação a publicações científicas.

cas formais. A pesquisa foi delimitada às subáreas de Bibliotecas Digitais (*Digital Libraries* - cs.DL) e Redes Sociais e Informação (*Social Networks and Information* - cs.SI), devido ao seu forte alinhamento epistêmico com a Ciência da Informação e à maturidade histórica dessas comunidades na adoção do compartilhamento prévio. Mediante modelagem estatística inferencial, testamos a hipótese de que o Esforço de Revisão, mensurado pelas múltiplas versões do documento depositadas na plataforma, apresenta maior força de associação do que atributos estruturais (métricas textuais) e de rede (coautoria e interdisciplinaridade) com o desfecho de publicação.

Estudos cientométricos têm buscado compreender como diferentes características influenciam a circulação do conhecimento científico. Variáveis como tamanho das equipes, características textuais e interdisciplinaridade têm sido investigadas. No caso específico dos *preprints*, o processo de revisão iterativa, evidenciado pelas múltiplas versões, pode representar um indicador de adaptação às expectativas editoriais (Lin *et al.*, 2020), embora essa dinâmica possa embutir forte influência de processos externos, como as exigências da revisão por pares tradicional (*peer review*).

2 MATERIAIS E MÉTODOS

Este é um estudo cientométrico inferencial de natureza quantitativa, baseado em metadados completos do *snapshot* público do *arXiv dataset* (Cornell University). Todo o processo de coleta, tratamento e análise dos dados foi realizado mediante codificação em *notebook jupyter* na linguagem *python*, utilizando as bibliotecas *pandas*, *numpy*, *scipy* e *statsmodels*.

O corpus de pesquisa compreende 23.136 registros bibliográficos, extraídos, em 31 de janeiro de 2026, cobrindo submissões entre os anos de 2010 a 2025 nas

áreas cs.DL e cs.SI. Embora os metadados do *dataset* sejam extensos quanto aos registros internos da plataforma, reconhece-se a diferença entre a completude técnica do banco de dados e a completude do fenômeno, uma vez que a conversão editorial envolve trajetórias não capturadas integralmente pelos sistemas.

A variável dependente (VD), denominada “sucesso editorial”, foi modelada de forma dicotômica: atribuiu-se valor 1 aos *preprints* que registraram atualização com um DOI formal de publicação, e 0 aos demais. Assume-se, portanto, que a VD atua como uma proxy imperfeita, refletindo não apenas o aceite editorial, mas também o comportamento de atualização pelo autor. As variáveis independentes (VIs) operacionalizadas foram: (i) esforço de revisão (número de versões depositadas do manuscrito); (ii) colaboração (contagem de coautores); (iii) extensão textual do título (número de caracteres no título); (iv) extensão textual do resumo (número de palavras no resumo); (v) interdisciplinaridade (quantidade de categorias temáticas secundárias indexadas); e (vi) ano de submissão.

A análise dos dados foi conduzida em quatro etapas principais: (1) Inicialmente, realizamos descrição estatística das variáveis; (2) aplicamos o coeficiente de correlação de *Spearman* e o diagnóstico de Fator de Inflação da Variância (VIF) para verificar associações e atestar a ausência de multicolinearidade severa ($VIF < 5$) entre as VIs; (3) na sequência, empregamos a Regressão Logística Binária para estimar as razões de chance (*Odds Ratio* - OR) condicionais para as subáreas cs.DL e cs.SI. O nível de significância estatística adotado foi $\alpha = 0,05$.

Inicialmente estimou-se o modelo completo com todas as covariáveis, e em seguida foi adotado o protocolo de seleção de variáveis pelo método *stepwise*, orientado pelo critério de informação de *Akaike* (AIC), para obter o modelo mais parcimonioso. O modelo foi avaliado por múltiplos critérios. O ajuste global foi avaliado

pelo pseudo- R^2 (*Mc Fadden*, *Cox & Snell* e *Nagelkerke*), enquanto a capacidade discriminatória foi avaliada sob a curva ROC (AUC). Adicionalmente, foram conduzidas análises de diagnóstico para verificar a adequação do modelo, incluindo a avaliação dos resíduos por meio de envelope simulado, a identificação de observações influentes via distância de *Cook* e a verificação de possíveis problemas de especificação.

É importante ressaltar que o modelo logístico permite identificar associações probabilísticas, controlando os efeitos simultâneos das variáveis, sem, contudo, aferir relações de causalidade estrita.

3 RESULTADOS

Nesta seção apresentam-se os achados da pesquisa mediante a estatística descritiva, análise de correlação e modelagem logística.

3.1 ESTATÍSTICA DESCRITIVA

A subárea cs.DL apresentou um volume absoluto de submissões expressivamente maior (19.035 *preprints* contra 4.101 de cs.SI) e uma taxa de presença de DOI levemente maior (17,04% versus 16,35%). Historicamente, o pico de conversões registradas ocorreu na coorte de 2011 (cs.DL: 27,14%; cs.SI: 24,48%), apresentando um declínio nas submissões mais recentes (2024), atingindo cerca de 9,9% em ambas.

Quanto à coautoria, ambas as áreas possuem mediana de 3 autores. Contudo, cs.SI apresenta equipes mais densas (Terceiro Quartil [Q3] = 5 autores; Média = 3,7 autores), enquanto cs.DL (Q3 = 4 autores) concentra outliers extremos, atingindo até 90 coautores (Desvio Padrão = 3,3 autores), contra um máximo de 43 autores em cs.SI. O esforço de revisão é mediano e homogêneo (1 versão para

ambas), embora cs.SI registre maior propensão a atualizações múltiplas (Média = 1,6 versões; Máx. = 31 versões) em relação a cs.DL (Média = 1,4 versões; Máx. = 10 versões).

As métricas textuais revelam convenções opostas: *preprints* de cs.DL possuem títulos mais longos (Mediana = 81 caracteres) comparados aos de cs.SI (Mediana = 72 caracteres). Inversamente, os resumos em cs.SI são mais extensos (Mediana = 174 palavras) que em cs.DL (Mediana = 164 palavras). Quanto à interdisciplinaridade, cs.SI atua de forma levemente mais transversal (Média = 2,5 categorias; Q3 = 3 categorias), do que cs.DL (Média = 1,9 categorias; Q3 = 2 categorias).

3.2 ANÁLISE DE CORRELAÇÃO E MULTICOLINEARIDADE

A estimação do Coeficiente de Spearman (ρ) revelou que, em cs.DL, as associações bivariadas com o sucesso editorial são fracas, sendo as maiores com o Esforço de Revisão ($\rho = 0,08$) e Extensão do Resumo ($\rho = 0,06$). Em cs.SI, o panorama é similar, com correlação fraca no Esforço de Revisão ($\rho = 0,13$) e na Interdisciplinaridade ($\rho = 0,09$). O tamanho da equipe, por sua vez, apresentou correlação próxima de nula ($\rho = -0,01$). O diagnóstico de multicolinearidade confirmou a adequação das variáveis selecionadas para a etapa regressiva, com todos os valores de VIF muito abaixo do limiar crítico de 5, indicando variância independente satisfatória entre os preditores.

3.3 REGRESSÃO LOGÍSTICA

A modelagem logística confirmou o Esforço de Revisão (número de versões no depositadas no *arXiv*) possui a associação estatística condicional mais consistente com a vinculação a um DOI. Mantendo as demais variáveis constantes, cada nova versão registrada se associa a um incremento nas chances de atualização formal

em 18,3% para cs.DL (OR = 1,183; $p < 0,001$) e em 18,7% para cs.SI (OR = 1,187; $p < 0,001$).

Também foram observados resultados estatisticamente significativos para as demais covariáveis nos modelos mais parcimoniosos ajustados separadamente por subárea. Para a categoria cs.DL, o procedimento de seleção indicou a permanência das covariáveis relacionadas à extensão textual do resumo e ao ano de submissão. Já para a categoria cs.SI, foram selecionadas as covariáveis associadas à colaboração, à interdisciplinaridade e ao ano de submissão.

Em termos de magnitude dos efeitos para cs.DL, observou-se que a extensão textual do resumo apresenta associação positiva com o desfecho, com OR de aproximadamente 1,003, enquanto o ano de submissão apresenta associação negativa (OR $\approx 0,98$). Esses resultados indicam que, embora exista associação estatística entre as variáveis analisadas e a presença de DOI, os efeitos observados são de pequena magnitude.

Para o modelo ajustado na categoria cs.SI, os resultados apresentam padrão distinto. A variável associada à colaboração apresentou OR de aproximadamente 1,02, indicando um efeito positivo, porém modesto. A interdisciplinaridade apresentou associação positiva mais pronunciada, com OR de aproximadamente 1,13. Por sua vez, o ano de submissão apresentou associação negativa com o desfecho (OR $\approx 0,91$). De maneira geral, assim como observado para cs.DL, os efeitos estimados são estatisticamente significativos, porém de magnitude relativamente reduzida.

Essa característica também se reflete nas medidas de qualidade global dos modelos ajustados por subárea. Para a categoria cs.DL, o pseudo- R^2 de McFadden foi de aproximadamente 0,008, enquanto os coeficientes de *Cox & Snell* e *Nagelkerke*

apresentaram valores próximos de 0,01 e 0,015, respectivamente. Já para a categoria cs.SI, observou-se melhora relativa no ajuste, com pseudo- R^2 de McFadden em torno de 0,032 e valores de *Cox & Snell* e *Nagelkerke* de aproximadamente 0,032 e 0,055. Apesar dessa diferença entre as subáreas, os valores permanecem baixos em termos absolutos, indicando que, embora as variáveis consideradas apresentem associação estatisticamente significativa com o desfecho, elas explicam apenas uma parcela limitada da variabilidade observada. Esses resultados reforçam a distinção entre significância estatística e relevância substantiva, sugerindo que os efeitos identificados possuem baixa capacidade explicativa no contexto analisado.

As análises de diagnóstico indicaram adequação do ajuste do modelo. O envelope de resíduos simulado não evidenciou desvios sistemáticos relevantes, e a análise da distância de *Cook* não indicou a presença de observações altamente influentes, sugerindo estabilidade das estimativas. Assim, os resultados apontam que o modelo está adequadamente especificado sob a distribuição assumida, ainda que apresente baixo poder explicativo.

4 CONCLUSÕES

Este estudo evidenciou associações estatísticas entre as características dos registros de *preprints* no arXiv, nas subáreas cs.DL e cs.SI, e a presença de identificadores formais de publicação (DOI). O número de versões do manuscrito destacou-se como a variável de maior correlação com este desfecho. Observou-se que *preprints* com maior número de versões atualizadas no repositório tendem a apresentar maior probabilidade estatística de registro final.

Ressalta-se, contudo, a forte possibilidade de endogeneidade desta métrica. A relação observada não permite inferir causalidade direta, visto que o incremento nas versões pode ser consequência, e não causa, do processo editorial externo.

O depósito de uma nova versão reflete frequentemente a resposta iterativa do autor às exigências (“*Revise & Resubmit*”) de um periódico ou conferência durante o processo tradicional de revisão por pares (*peer review*).

Devem ser destacadas as limitações metodológicas e interpretativas desta pesquisa. Primeiramente, o uso do DOI como *proxy* de sucesso editorial é falho, uma vez que depende do comportamento proativo de atualização dos autores no *arXiv* e negligencia publicações em veículos (como anais de eventos locais ou revistas de nicho) que não adotam este identificador, gerando inevitáveis falsos negativos. Além disso, as conclusões restringem-se às duas subáreas analisadas, não cabendo generalizações amplas para campos disciplinares com dinâmicas epistêmicas distintas.

Outra limitação crítica reside no viés da janela temporal. O modelo logístico aplicado trata o desfecho de forma estática e não incorpora a censura à direita. Como o corpus abrange depósitos até 2025, os *preprints* mais recentes tiveram muito menos tempo de exposição ao “risco” de publicação do que os de 2010, comprometendo parcialmente a comparabilidade e influenciando o declínio aparente nas taxas descritivas recentes.

Trabalhos futuros devem adotar modelos de Análise de Sobrevida (ex: regressão de Cox) para controlar de forma robusta o tempo até o evento. Ainda assim, conclui-se que o repositório atua de fato como um ambiente dinâmico, onde a persistência do esforço de atualização do manuscrito está estatisticamente acoplada à trajetória de validação formal em áreas fronteiriças da Ciência da Informação.



REFERÊNCIAS

LARIVIÈRE, V.; SUGIMOTO, C. R.; MACALUSO, B.; MILOJEVIĆ, S.; CRONIN, B.; THELWALL, M. arXiv e-prints and the journal of record: An analysis of roles and relationships. **Journal of the Association for Information Science and Technology**, v. 65, n. 6, p. 1157-1169, 2014. DOI: [10.1002/asi.23044](https://doi.org/10.1002/asi.23044). Acesso em 10 jul. 2026.

LIN, J.; YU, Y.; ZHOU, Y.; ZHOU, Z.; SHI, X. How many preprints have actually been printed and why: a case study of computer science preprints on arXiv. **Scientometrics**, [S. l.], v. 124, n. 1, p. 555-574, 1 jul. 2020. DOI: [10.1007/s11192-020-03430-8](https://doi.org/10.1007/s11192-020-03430-8). Acesso em 10 jul. 2026.

NIELSEN, M. **Reinventing Discovery**: The New Era of Networked Science. Princeton: Princeton University Press, 2012. 272 p. ISBN: 9780691160191.

WALTMAN, L.; PINFIELD, S.; RZAYEVA, N.; OLIVEIRA HENRIQUES, S.; FANG, Z.; BRUMBERG, J.; GREAVES, S.; HURST, P.; COLLINGS, A.; HEINRICHS, A.; LINDSAY, N.; MACCALLUM, C. J.; MORGAN, D.; SANSONE, S.-A.; SWAMINATHAN, S. Scholarly Communication in Times of Crisis The response of the scholarly communication system to the COVID-19 pandemic. **Research on Research Institute**. Report, 2021. DOI: [10.6084/m9.figshare.17125394.v1](https://doi.org/10.6084/m9.figshare.17125394.v1). Acesso em 10 jul. 2026.

WANG, Z.; CHEN, Y.; GLÄNZEL, W. Preprints as accelerator of scholarly communication: An empirical analysis in Mathematics. **Journal of Informetrics**, [S. l.], v. 14, n. 4, p. 101097, 1 nov. 2020. DOI: [10.1016/j.joi.2020.101097](https://doi.org/10.1016/j.joi.2020.101097). Acesso em 10 jul. 2026.