Distribuição Geográfica e Atividades em Projetos Open Source: Um Estudo Exploratório

João Victor Arantes Cubel[1], **Hudson Silva Borges**[1]

[1]FACOM – Universidade Federal de Mato Grosso do Sul (UFMS)
Campo Grande – MS – Brasil

{joao.cubel,hudson.borges}@ufms.br

Abstract. *This work investigates the geographic distribution of contributors and activity types in the HyDE open source repository. Using data restored from RepoInsights into a local MongoDB instance, we analyzed seven interaction types and aggregated them by country. The results show that technical activities (commits and pull requests) are highly concentrated in a few countries, while social interactions (stargazers and reactions) are more globally distributed. Non-parametric statistical tests confirm significant differences between activity types. Weak but significant correlations were found between contribution volume, English proficiency, and Human Development Index. The study offers insights into geographic inequalities in OSS participation.*

Resumo. *Este trabalho investiga a distribuição geográfica dos desenvolvedores e os tipos de atividades realizadas no repositório open source HyDE. A partir de dados restaurados no MongoDB, analisamos sete categorias de interação e as agregamos por país. Os resultados mostram que atividades técnicas, como commits e pull requests, são concentradas em poucos países, enquanto interações sociais apresentam maior alcance global. Testes estatísticos confirmam diferenças significativas entre os tipos de atividade. Encontramos ainda correlações fracas, porém positivas, entre volume de contribuições, proficiência em inglês e IDH. O estudo evidencia desigualdades geográficas persistentes na participação em OSS.*

1. Introdução

O desenvolvimento de software passou por mudanças profundas nas últimas décadas, especialmente após a popularização dos sistemas de controle de versão distribuídos e das plataformas de hospedagem de código. Ferramentas como o GitHub transformaram a colaboração em software ao incorporar, além do controle de versão, mecanismos sociais que permitem acompanhar projetos, interagir com mantenedores e formar redes de colaboração entre desenvolvedores. Esse fenômeno, conhecido como *social coding* [Dabbish et al. 2012], ampliou significativamente o alcance dos projetos open source, tornando-os ecossistemas globais, dinâmicos e altamente interconectados.

Apesar desse caráter global, o desenvolvimento open source não ocorre de maneira uniforme entre regiões. Pesquisas anteriores mostram que a distribuição geográfica dos desenvolvedores é desigual, concentrando-se em países historicamente associados a avanços tecnológicos [Robles et al. 2008, Takhteyev and Hilts 2010,

Wachs et al. 2022, Rossi and Zacchiroli 2022]. Além disso, fatores como idioma, infraestrutura, renda e redes de contato influenciam quem consegue participar efetivamente de um projeto OSS. Assim, comunidades open source podem ser compreendidas não apenas como coletivos técnicos, mas também como comunidades moldadas por fatores sociais e geográficos.

Nesse contexto, conhecer a composição geográfica de um projeto torna-se fundamental para interpretar sua dinâmica colaborativa. Equipes de *Developer Relations*, por exemplo, dependem desse conhecimento para planejar estratégias de engajamento, traduzir documentação, organizar eventos e fortalecer comunidades locais. Embora existam estudos que analisam a localização dos desenvolvedores, a literatura ainda carece de investigações que relacionem essas informações às diversas atividades disponíveis na plataforma, como criação de *issues*, submissão de *pull requests*, participação em discussões ou simples interação social. Ou seja, sabe-se onde os desenvolvedores estão, mas ainda se conhece pouco sobre o que eles fazem de acordo com sua região.

Diante dessa lacuna, este trabalho busca responder a uma questão prática e conceitual: como a distribuição geográfica e o tipo de atividade realizada em um projeto open source se relacionam com fatores contextuais como proficiência em língua inglesa e condições socioeconômicas? Essa pergunta é relevante tanto para a academia, ao contribuir para a compreensão dos ecossistemas de software (*software ecosystems* ou ECOS), quanto para a indústria, especialmente para equipes de *Developer Relations* e mantenedores que precisam entender "quem faz o quê e a partir de onde" para planejar ações de engajamento, suporte e expansão da comunidade.

Especificamente, investigamos um ecossistema sociotécnico em torno de um repositório no GitHub, no qual desenvolvedores de diferentes países interagem por meio de múltiplos tipos de atividades (técnicas e sociais). Nosso objetivo é caracterizar como essas atividades se distribuem entre os países e em que medida essa distribuição está associada a fatores externos como o Índice de Desenvolvimento Humano (IDH) e a proficiência em língua inglesa.

Com base nesse contexto, este estudo é guiado pelas seguintes questões de pesquisa:

- **RQ1:** Como se distribuem geograficamente os desenvolvedores que participam do projeto open source analisado?
- **RQ2:** Há diferenças na distribuição geográfica dos desenvolvedores de acordo com os tipos de atividades realizadas (por exemplo, commits, issues, pull requests, reações, estrelas)?
- **RQ3:** Existe relação entre o nível de proficiência em língua inglesa de um país e sua intensidade de contribuição em projetos open source?
- **RQ4:** Há correlação entre indicadores socioeconômicos, como o Índice de Desenvolvimento Humano (IDH), e o volume de contribuições em projetos open source?

As próximas seções aprofundam esse arcabouço: a Seção 2 apresenta o contexto teórico e o conceito de ecossistemas de software; a Seção 3 discute estudos anteriores sobre distribuição geográfica de desenvolvedores; a Seção 4 descreve o processo de coleta, tratamento e análise dos dados; a Seção 5 apresenta os principais achados; a Seção 6 os

discute à luz das questões de pesquisa e das ameaças à validade; e, por fim, a Seção 7 traz as conclusões e trabalhos futuros.

2. Background

O desenvolvimento de software open source (OSS) é caracterizado por uma colaboração distribuída e descentralizada, na qual indivíduos de diferentes países e fusos horários contribuem para projetos comuns. Plataformas como o GitHub transformaram esse ecossistema ao incorporar, além do controle de versão, recursos sociais que permitem transparência, interação entre desenvolvedores e coordenação de trabalho [Dabbish et al. 2012]. Esses recursos reforçam a noção de *social coding*, na qual o desenvolvimento de software é mediado por mecanismos sociais que influenciam colaboração, reputação e engajamento [Tsay et al. 2014]. Esse paradigma evidencia que o desenvolvimento de software não é apenas um processo técnico, mas também social e geograficamente moldado.

Apesar de sua natureza global, as comunidades open source não são uniformemente distribuídas. A localização dos desenvolvedores influencia sua visibilidade, oportunidades de colaboração e até o reconhecimento dentro das comunidades. Estudos anteriores demonstram que fatores como idioma, infraestrutura tecnológica e proximidade com polos de inovação impactam diretamente a participação em projetos OSS [Takhteyev and Hilts 2010, Wachs et al. 2022]. Estudos longitudinais também apontam que, ao longo de cinco décadas de desenvolvimento de software livre, a produção de código evoluiu de uma concentração quase exclusiva na América do Norte e Europa para uma participação crescente de regiões da Ásia e América Latina, embora ainda marcada por desigualdades geográficas significativas [Rossi and Zacchiroli 2022]. Dessa forma, compreender como os desenvolvedores se distribuem geograficamente torna-se essencial para interpretar dinâmicas de poder, influência e contribuição nesses ambientes digitais.

Um exemplo emblemático dessa influência geográfica é o caso do projeto *Vue.js*, criado por Evan You em 2014[1]. O framework ganhou tração global após o autor dedicar uma documentação completa para a comunidade chinesa, traduzida e adaptada à sua realidade linguística e cultural. Esse esforço resultou em um crescimento exponencial do número de contribuidores e usuários da China, mostrando que a sensibilidade à geografia, mesmo em um ecossistema global, pode ser determinante para o sucesso de um projeto open source [You 2016]. Situações como essa evidenciam que o contexto local, a língua e as práticas regionais moldam o engajamento e a adoção de tecnologias.

Além da dimensão técnica e comunitária, o entendimento sobre a dispersão geográfica dos desenvolvedores possui aplicações práticas em áreas emergentes como *Developer Relations* (DevRel). Essa disciplina busca estreitar o relacionamento entre empresas e comunidades de desenvolvedores, criando estratégias para engajamento, suporte e fortalecimento de ecossistemas de software [Williams and Williams 2019]. Pesquisas sobre comunidades distribuídas mostram que práticas orientadas a engajamento impactam a formação de redes colaborativas e a sustentabilidade dos projetos [Teixeira and Robles 2015, Zagalsky et al. 2016]. Ao conhecer onde estão seus colaboradores mais ativos e quais atividades desempenham, organizações podem direcionar

[1] `https://github.com/vuejs/vue`

melhor esforços de comunicação, eventos, traduções e incentivos, promovendo comunidades mais diversas e equilibradas.

Nesse contexto, torna-se evidente a importância de compreender não apenas onde os desenvolvedores estão, mas também quais tipos de atividades desempenham. A localização geográfica pode estar associada a padrões distintos de contribuição, refletindo diferenças culturais, econômicas e educacionais. Por exemplo, certas regiões podem concentrar atividades de manutenção e correção de erros, enquanto outras se destacam em inovação e criação de novas funcionalidades. Conhecer essas nuances pode apoiar tanto a pesquisa acadêmica quanto estratégias de gestão de comunidades open source.

Além disso, este estudo busca investigar fatores externos que podem influenciar a participação dos países em projetos OSS. Entre eles, destacam-se a proficiência em língua inglesa, já que o inglês é a língua predominante em repositórios e documentação técnica, os indicadores socioeconômicos, como o Índice de Desenvolvimento Humano (IDH), e o nível de desenvolvimento tecnológico de cada país. A análise conjunta desses elementos permite compreender não apenas *onde* e *como* os desenvolvedores colaboram, mas também *por que* determinadas regiões se destacam em participação open source.

Em síntese, a literatura indica que projetos open source devem ser compreendidos como ecossistemas sociotécnicos, nos quais fatores técnicos, sociais e geográficos interagem de forma complexa. Embora estudos anteriores tenham demonstrado a persistência de desigualdades regionais na participação em OSS, ainda há espaço para investigações que relacionem a localização geográfica dos desenvolvedores aos diferentes tipos de atividade disponíveis nas plataformas de social coding. Esse arcabouço conceitual fundamenta a análise empírica apresentada nas próximas seções.

3. Trabalhos Relacionados

A investigação da distribuição geográfica de desenvolvedores de software open source (OSS) é um campo de pesquisa relativamente recente, que tem demonstrado que, mesmo em um contexto de colaboração digital e distribuída, a geografia continua exercendo papel relevante na formação e dinâmica das comunidades de software. No entanto, grande parte das pesquisas existentes se limita a mapear a localização dos desenvolvedores ou a examinar de forma geral como a distância física afeta a colaboração, sem explorar em profundidade os fatores que sustentam esses padrões. Esta seção revisa os principais estudos sobre o tema e posiciona o presente trabalho em relação à literatura existente.

Os primeiros esforços para compreender a geografia do desenvolvimento de software livre foram conduzidos por [Robles et al. 2008], que analisaram dados do *SourceForge* e listas de e-mail de grandes projetos de software. Por meio da inferência de localização baseada em domínios de e-mail e fusos horários, os autores identificaram uma forte concentração de desenvolvedores na Europa e na América do Norte, evidenciando a influência de fatores socioeconômicos e culturais na distribuição global de contribuidores. Posteriormente, [Takhteyev and Hilts 2010] realizaram o primeiro estudo de larga escala utilizando dados do GitHub, revelando que, embora a colaboração em OSS seja global e descentralizada, ela ainda apresenta um claro padrão de concentração regional e um notável "viés local", ou seja, desenvolvedores tendem a colaborar com outros que se encontram geograficamente próximos. Estudos mais recentes, como [Wachs et al. 2022] e [Rossi and Zacchiroli 2022], ampliaram essa análise com dados contemporâneos e séries

históricas mais extensas, mostrando que, embora a participação de países da Ásia e da América Latina tenha aumentado, a atividade OSS permanece fortemente concentrada em poucos polos regionais, principalmente na América do Norte e na Europa Ocidental. O foco comum dessa linha de pesquisa tem sido a quantificação e caracterização espacial dos desenvolvedores, em vez da análise da natureza de suas contribuições.

Adotando uma perspectiva nacional, [Filho et al. 2015] investigaram a distribuição de desenvolvedores de software no Brasil, buscando compreender como o prestígio de desenvolvedores influentes se relaciona com fatores socioeconômicos dos estados brasileiros. A metodologia envolveu a coleta de dados de 4.016 usuários ativos do GitHub e o uso de métricas de análise de redes sociais para mensurar o prestígio a partir das relações de seguidores (*follow relationships*). Os resultados confirmaram a hipótese de homofilia geográfica: desenvolvedores tendem a seguir outros localizados no mesmo estado, e aqueles em regiões mais urbanizadas e economicamente desenvolvidas, como São Paulo e Rio de Janeiro, apresentam maior prestígio. Esse estudo foi pioneiro ao conectar geografia e capital social no contexto do OSS, ainda que baseado em uma métrica indireta, como o número de seguidores. O presente trabalho avança nessa direção ao analisar diretamente as atividades de desenvolvimento (*commits*, *issues* e *pull requests*), buscando identificar se a própria prática de contribuição varia de acordo com a localização geográfica.

Sob uma perspectiva comparativa internacional, [Almarzouq and Alnahedh 2023] examinaram diferenças de engajamento em OSS entre 39 países, com o objetivo de identificar como os padrões de colaboração e os perfis de habilidade dos desenvolvedores variam conforme o contexto socioeconômico. Os autores coletaram dados via API do GitHub, considerando os 200 usuários e organizações mais influentes de cada país, e aplicaram técnicas de Processamento de Linguagem Natural (PLN) sobre as biografias dos perfis. As análises mostraram que, em países com alta atividade OSS, os perfis tendem a incluir termos como “Open Source”, “Founder” e “Creator”, enquanto em países com menor engajamento predominam descrições como “Student” e “Security Specialist”. Esses resultados sugerem que o envolvimento em OSS reflete características culturais e econômicas mais amplas, associadas ao empreendedorismo e à maturidade tecnológica local. Apesar de relevante, o estudo se baseia em autodeclarações textuais como proxy de atividade, sem examinar as ações de desenvolvimento propriamente ditas. O presente trabalho propõe avançar nessa lacuna ao observar diretamente os tipos de atividade realizados pelos contribuidores, oferecendo uma visão mais precisa da participação geográfica em projetos OSS.

Em síntese, a literatura existente traçou de forma sólida o panorama global da distribuição de desenvolvedores [Robles et al. 2008, Takhteyev and Hilts 2010, Wachs et al. 2022, Rossi and Zacchiroli 2022] e demonstrou a importância da geografia e das redes sociais na formação das comunidades de software [Filho et al. 2015]. Entretanto, a compreensão sobre como a localização geográfica influencia os diferentes tipos de atividade desempenhados pelos desenvolvedores ainda é incipiente. Trabalhos recentes [Almarzouq and Alnahedh 2023] abordaram o tema de forma indireta, baseando-se em perfis e metadados. Assim, este estudo busca preencher essa lacuna ao correlacionar diretamente a localização geográfica dos contribuidores com as atividades de desenvolvimento que executam em um projeto open source, permitindo compreender não apenas

onde os desenvolvedores estão, mas também como se colabora no ecossistema de software livre.

4. Metodologia

Este estudo segue uma abordagem quantitativa e exploratória, fundamentada na análise de dados públicos provenientes da plataforma GitHub. O objetivo é investigar a relação entre a distribuição geográfica de desenvolvedores, os tipos de atividades que realizam e fatores externos como proficiência em língua inglesa e indicadores socioeconômicos. A metodologia foi estruturada em quatro etapas principais: (i) seleção e coleta dos dados, (ii) tratamento e agregação, (iii) análise estatística e visualização e (iv) integração com fatores contextuais.

4.1. Seleção do Projeto

Este estudo adota um delineamento baseado em estudo de caso, em que o objetivo é analisar e compreender o fenômeno investigado a partir da observação aprofundada de um único projeto. Esse formato permite uma análise detalhada das dinâmicas de contribuição e da distribuição geográfica de seus participantes, servindo como base para estudos futuros que poderão ampliar a investigação para múltiplos repositórios, possibilitando validação estatística mais ampla dos achados.

O repositório analisado foi o *HyDE*, disponível em `https://github.com/HyDE-Project/HyDE`. A escolha foi motivada por critérios técnicos: (i) trata-se de um projeto ativo, com histórico contínuo de contribuições; (ii) possui múltiplos tipos de interação no GitHub (commits, pull requests, issues, discussions, reactions, estrelas e watchers), o que é essencial para comparar diferentes formas de participação; e (iii) apresenta diversidade geográfica de contribuidores, permitindo a análise exploratória da distribuição por país.

A familiaridade prévia do autor com o projeto facilitou o entendimento do contexto funcional e das principais rotinas de desenvolvimento, o que contribuiu para a interpretação dos resultados. No entanto, esse fator não foi o critério central de seleção, e sim um facilitador operacional.

4.2. Coleta dos Dados

Os dados utilizados neste estudo foram obtidos a partir de uma instância local do **MongoDB**, restaurada com arquivos exportados da plataforma *RepoInsights*[2]. As coleções analisadas incluem *Commit*, *PullRequest*, *Issue*, *Discussion*, *Reaction*, *Stargazer* e *Watcher*. Juntas, essas coleções representam um conjunto diversificado de atividades, abrangendo desde ações diretas de desenvolvimento até interações sociais e de engajamento.

Um desafio significativo na coleta de dados está relacionado à falta de padronização das informações de localização fornecidas pelos próprios usuários, já que esse dado é autodeclarado e inserido como texto livre no perfil. Isso resulta em variação semântica e sintática, além de valores ausentes em muitos casos. Por exemplo, um mesmo país pode ser representado como *"BR"*, *"Brazil"*, *"Brasil"*, *"Brazil-SP"*, *"São Paulo,*

[2] `https://repo-insights.hsborges.dev/`

Brazil" ou por abreviações como *"BRA"*. Situação semelhante ocorre com outras localidades, como *"USA"*, *"United States"*, *"California, US"* ou *"NY, USA"*; e *"España"*, *"Spain"* ou *"Madrid, ES"*.

Para lidar com essas inconsistências, foi conduzido um processo de normalização geográfica, mapeando expressões variantes para nomes padronizados conforme convenções ISO. Apenas registros com país identificável foram mantidos na análise, enquanto entradas sem qualquer informação de localização foram excluídas.

4.3. Tratamento e Agregação dos Dados

O pré-processamento foi conduzido utilizando o **MongoDB Compass**, por meio da construção de pipelines de agregação. Esses pipelines utilizam estágios específicos para integrar, filtrar e resumir os dados provenientes das diferentes coleções do banco.

Os estágios dos pipelines cumprem papéis específicos: `$lookup` é utilizado para juntar a coleção de eventos (por exemplo, *Commit*) com a coleção *Actor*, recuperando informações de perfil dos usuários; `$unwind` transforma listas (arrays) em múltiplos registros, permitindo analisar cada autor ou participante individualmente; `$match` filtra apenas os registros de interesse, como eventos do repositório HyDE com país identificado; `$group` agrega os dados por país e usuário, calculando contagens de eventos e o número de usuários únicos; e `$project` seleciona, renomeia e organiza os campos finais, produzindo uma tabela resumida por país, exportável em formato `.json`.

Cada pipeline consolidou estatísticas por país, contabilizando o número de usuários únicos e o volume de eventos conforme o tipo de recurso analisado. Como exemplos de agregações implementadas, temos:

- **Commit:** agrupamento por autor e país, calculando o total de commits por usuário e a soma por país;
- **PullRequest:** identificação de autores e responsáveis por merges, com contagem de usuários únicos por país;
- **Issue:** agregação de autores e editores, contabilizando usuários únicos envolvidos por país;
- **Discussion:** união entre autores e comentaristas utilizando `$unionWith`, seguida da contagem de usuários ativos por país;
- **Reaction, Stargazer e Watcher:** contagem dos usuários que realizaram cada ação, agregada por país.

Durante o pré-processamento, também foram realizadas a eliminação de duplicidades (deduplicação por identificador de usuário) e a padronização geográfica dos nomes dos países. Neste trabalho, o termo *normalização* refere-se exclusivamente à padronização de localização, mapeando variações como "Brasil", "Brazil", "BR" e "São Paulo, Brazil" para um único rótulo padronizado ("Brazil"), conforme o padrão ISO 3166 de códigos de países[3]. Não foi aplicada qualquer normalização estatística sobre as contagens de eventos.

Os resultados finais de cada pipeline foram exportados em formato `.json` para posterior análise estatística e visualização em Python.

[3] `https://www.iso.org/iso-3166-country-codes.html`

4.4. Análise Estatística e Visualização

A etapa de análise foi executada no **Google Colab**, utilizando `pandas` para manipulação de dados, `plotly.express` para visualizações interativas e `scipy.stats` para testes estatísticos. O fluxo de trabalho incluiu:

1. **Carga dos dados:** leitura dos arquivos `.json` exportados do MongoDB e conversão para *DataFrames*;
2. **Cálculo de métricas descritivas:** total de eventos, número de usuários únicos por país e porcentagem de perfis com país resolvido;
3. **Visualizações:** gráficos de barras (top N países), gráficos de pizza (distribuição percentual), mapas de calor (choropleth) e boxplots para análise de dispersão e outliers;
4. **Testes estatísticos:** aplicação de Mann–Whitney U para comparações pareadas, Kruskal–Wallis para comparar múltiplos grupos e correlação de Spearman para medir associações entre métricas por país.

Os arquivos resultantes (em formato `.csv`) foram utilizados para construção de tabelas e gráficos que compõem a análise exploratória.

4.5. Integração com Fatores Externos

Para investigar fatores que possam explicar diferenças na participação entre países, os dados consolidados foram integrados com indicadores externos:

- **Proficiência em inglês:** índice EF EPI 2024, que classifica países conforme o nível médio de fluência no idioma[4];
- **Índice de Desenvolvimento Humano (IDH):** dados do Programa das Nações Unidas para o Desenvolvimento (PNUD), considerando os valores mais recentes disponíveis[5].

Os dados foram unidos à base principal por nome de país, permitindo a realização de testes de correlação de Spearman entre o IDH, a proficiência em inglês e as métricas de contribuição, como número de usuários únicos e total de eventos em cada tipo de atividade.

4.6. Relação com as Questões de Pesquisa

Esta seção descreve como cada questão de pesquisa (RQ) foi abordada metodologicamente.

RQ1 Como se distribuem geograficamente os desenvolvedores que participam de projetos open source?
Para responder a esta questão, foram utilizadas todas as coleções agregadas (*Commit*, *PullRequest*, *Issue*, *Discussion*, *Reaction*, *Stargazer* e *Watcher*). As localizações dos usuários foram extraídas após a normalização, e os dados foram visualizados por meio de mapas de calor e rankings de países.

[4]EF English Proficiency Index (EF EPI) 2024, disponível em: `https://www.ef.com/wwen/epi/`.

[5]Relatórios de Desenvolvimento Humano (HDI) do PNUD, disponíveis em: `https://hdr.undp.org/`.

RQ2 Há diferenças na distribuição geográfica dos desenvolvedores de acordo com os tipos de atividades realizadas?
Para investigar essa questão, as métricas de usuários únicos por país foram comparadas entre os diferentes tipos de atividades. Foram aplicados testes não paramétricos (Mann–Whitney U e Kruskal–Wallis) para avaliar diferenças significativas entre distribuições.

RQ3 Existe relação entre o nível de proficiência em língua inglesa de um país e sua intensidade de contribuição em projetos open source?
Os dados agregados por país foram combinados com o índice de proficiência EF EPI 2024, e a correlação entre as métricas de contribuição e o nível de proficiência foi avaliada usando o coeficiente de Spearman.

RQ4 Há correlação entre indicadores socioeconômicos, como o IDH, e o volume de contribuições open source?
Para responder a esta questão, os dados consolidados de contribuições foram cruzados com o Índice de Desenvolvimento Humano (IDH), e foram calculadas correlações de Spearman entre as variáveis.

Encerradas as etapas de coleta, tratamento e análise estatística dos dados, a próxima seção apresenta os principais resultados obtidos, descrevendo os achados referentes à distribuição geográfica dos desenvolvedores, aos tipos de atividades desempenhadas e às correlações com fatores externos.

5. Resultados

Nesta seção apresentamos os resultados obtidos a partir da análise das atividades realizadas no repositório *HyDE*. Primeiramente, caracterizamos o conjunto de dados utilizado. Em seguida, discutimos os achados relativos a cada uma das questões de pesquisa (RQs), incluindo evidências empíricas, gráficos e resultados de testes estatísticos não paramétricos.

5.1. Caracterização do Dataset

O conjunto de dados analisado foi obtido a partir da plataforma RepoInsights e restaurado em uma instância local do MongoDB. Após os pipelines de agregação aplicados no MongoDB Compass e do posterior processamento no Google Colab, os dados consolidados contemplam sete tipos de atividades: *Commit*, *PullRequest*, *Issue*, *Discussion*, *Reaction*, *Stargazer* e *Watcher*.

Durante o processo de preparação, os dados passaram por etapas de normalização, limpeza e deduplicação. Esse tratamento incluiu a padronização das informações de localização dos usuários, permitindo que os países fossem identificados de forma consistente. Apenas registros com país reconhecido foram considerados nas análises.

A partir desse conjunto tratado, foi possível calcular, para cada país, o número de usuários únicos e o total de eventos em cada tipo de recurso. As análises subsequentes fazem uso desses dados consolidados para investigar diferenças entre atividades, padrões de distribuição geográfica e correlações com fatores externos.

Um aspecto importante da qualidade dos dados diz respeito à proporção de usuários com país identificável. Como parte do processo de normalização, apenas perfis nos quais foi possível extrair uma localização interpretável foram considerados como

“identificados”. O restante foi classificado como “não identificado”. A Tabela 1 apresenta a distribuição desses dois grupos para cada tipo de atividade analisada.

Table 1. Usuários identificados vs não identificados por recurso.

Recurso	Identificados	Não identificados	Total	% Identificados
Commit	50	79	129	38.76%
Discussion	50	116	166	30.12%
Issue	92	216	308	29.87%
PullRequest	27	58	85	31.76%
Reaction	114	229	343	33.24%
Stargazer	2205	3748	5953	37.05%
Watcher	13	29	42	30.95%

Observa-se que, em todos os recursos analisados, a proporção de usuários com país não identificado é superior a 60%. As atividades sociais (*Stargazer* e *Reaction*) concentram os maiores volumes absolutos de perfis não identificados, enquanto recursos técnicos como *Commit* e *PullRequest* apresentam menor quantidade de usuários, mas ainda assim com predominância de perfis sem localização definida. Esses valores reforçam a necessidade do processo de normalização geográfica e indicam que parte substancial da comunidade não disponibiliza informações geográficas em seu perfil, o que afeta diretamente a completude das análises.

A Figura 1 apresenta o ranking dos 10 países com maior volume total de atividades no repositório. Esse gráfico sintetiza, em uma única métrica agregada, a soma de todos os tipos de eventos realizados pelos usuários, permitindo visualizar a concentração das contribuições e identificar os países que desempenham papel mais relevante no projeto.

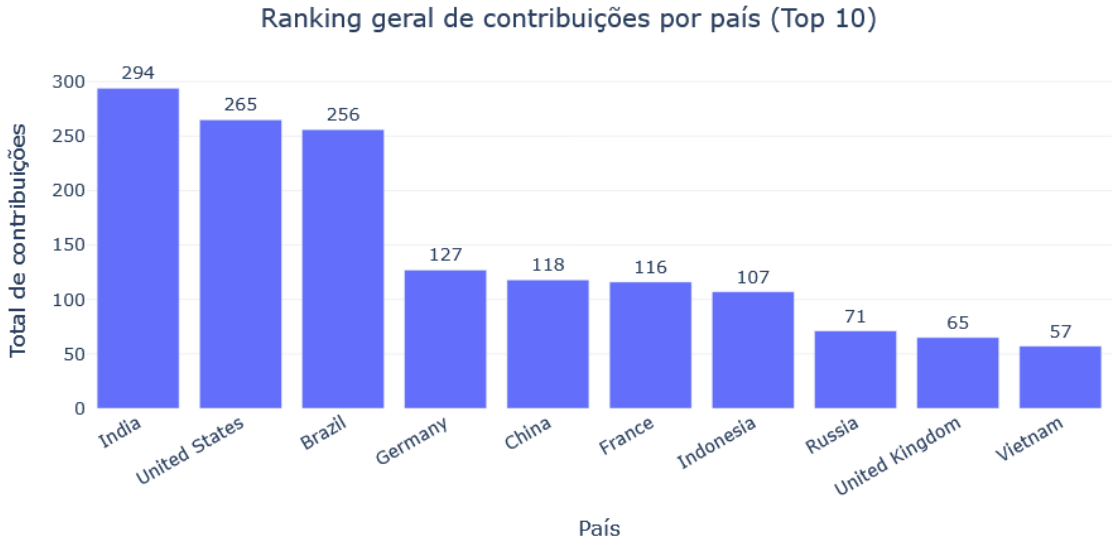

Figure 1. Ranking dos 10 países com maior volume total de atividades no repositório HyDE.

Observa-se uma forte concentração em poucos países, especialmente Índia, Es-

tados Unidos e Brasil, que juntos somam a maior parte das interações registradas. Em seguida, países europeus e asiáticos aparecem com menor, mas ainda relevante, volume de participação. Esse padrão confirma a existência de desigualdades geográficas já reportadas na literatura e fornece o panorama inicial sobre o qual as análises das questões de pesquisa são aprofundadas.

Embora o ranking apresente os países mais participativos, o mapa global da Figura 2 permite visualizar a distribuição espacial das atividades de forma mais intuitiva. A partir dele, é possível identificar regiões com maior densidade de contribuições e avaliar o alcance geográfico geral do projeto.

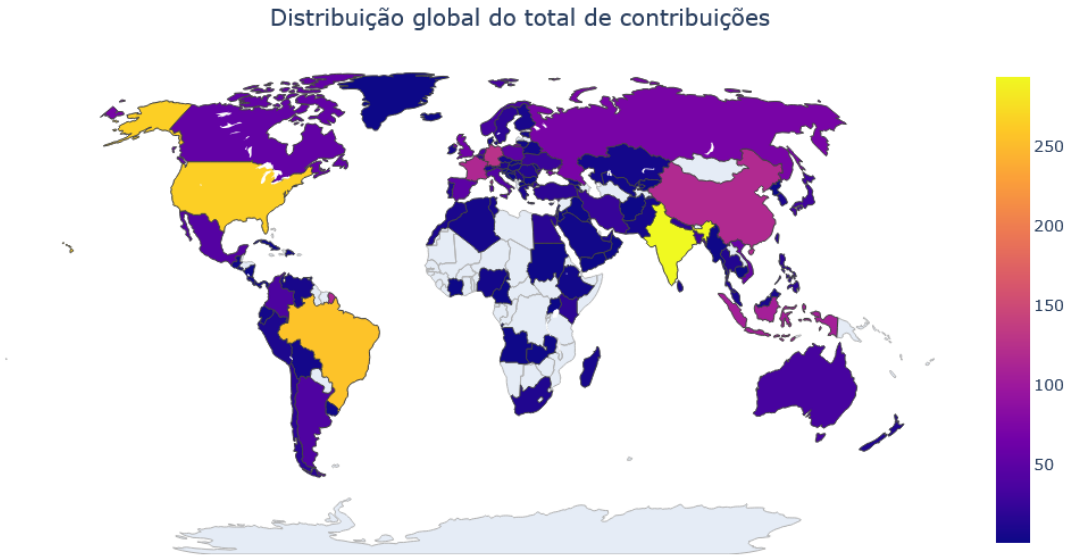

Figure 2. Mapa global do total de atividades agregadas por país.

Essa visualização reforça o padrão de concentração identificado no gráfico anterior, ao mesmo tempo em que destaca o caráter global, ainda que desigual, da comunidade envolvida no desenvolvimento do projeto.

5.2. RQ1 – Distribuição geográfica dos desenvolvedores

A RQ1 teve como objetivo identificar como se distribuem geograficamente os desenvolvedores que participam de diferentes atividades no projeto analisado. Para isso, foram considerados usuários únicos por país nas coleções *Commit*, *PullRequest*, *Issue*, *Discussion*, *Reaction*, *Stargazer* e *Watcher*.

Os resultados mostram que a participação global não é uniforme: um pequeno conjunto de países concentra a maior parte das atividades, enquanto a maioria dos demais apresenta participação reduzida. Países como Estados Unidos, Índia, Brasil, Alemanha e China aparecem de forma recorrente nas primeiras posições, indicando sua relevância no ecossistema do repositório. Esse padrão é semelhante ao relatado em estudos prévios sobre software livre, que destacam a existência de desigualdades regionais persistentes no envolvimento com projetos open source.

Além disso, observou-se uma diferença consistente entre atividades sociais e atividades técnicas. Recursos como *Stargazer* e *Reaction* possuem distribuição global mais

ampla e diversificada, enquanto atividades relacionadas ao código — como *Commit* e *PullRequest*, são significativamente mais concentradas. Esses contrastes sugerem que barreiras técnicas, culturais, linguísticas e de disponibilidade de tempo podem afetar o tipo de contribuição que usuários realizam.

Stargazers. A Figura 3 apresenta a distribuição geográfica dos usuários que marcaram estrela no repositório. Observa-se uma participação altamente distribuída, com forte presença na Índia, Brasil e Estados Unidos. Muitos países da Europa, Ásia e América Latina também aparecem com níveis consideráveis de atividade, evidenciando que o ato de favoritar um repositório constitui uma forma de engajamento de baixo custo, acessível a uma audiência global.

A Figura 4 mostra o ranking dos países com maior número de usuários que marcaram estrela. Os três primeiros — Índia, Brasil e Estados Unidos — concentram mais de 650 usuários combinados, destacando a ampla visibilidade internacional do repositório.

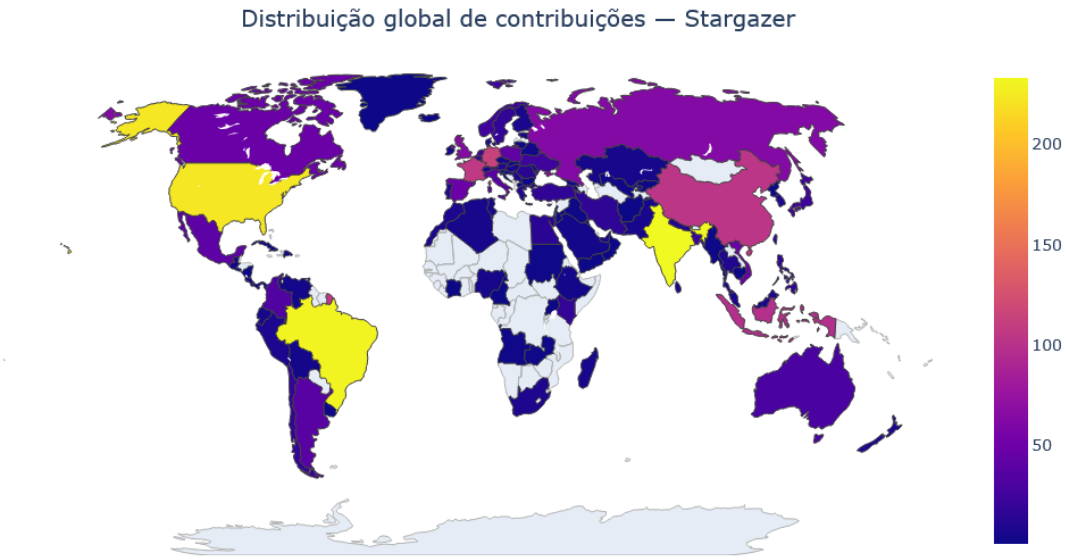

Figure 3. Distribuição geográfica dos usuários que marcaram estrela no repositório.

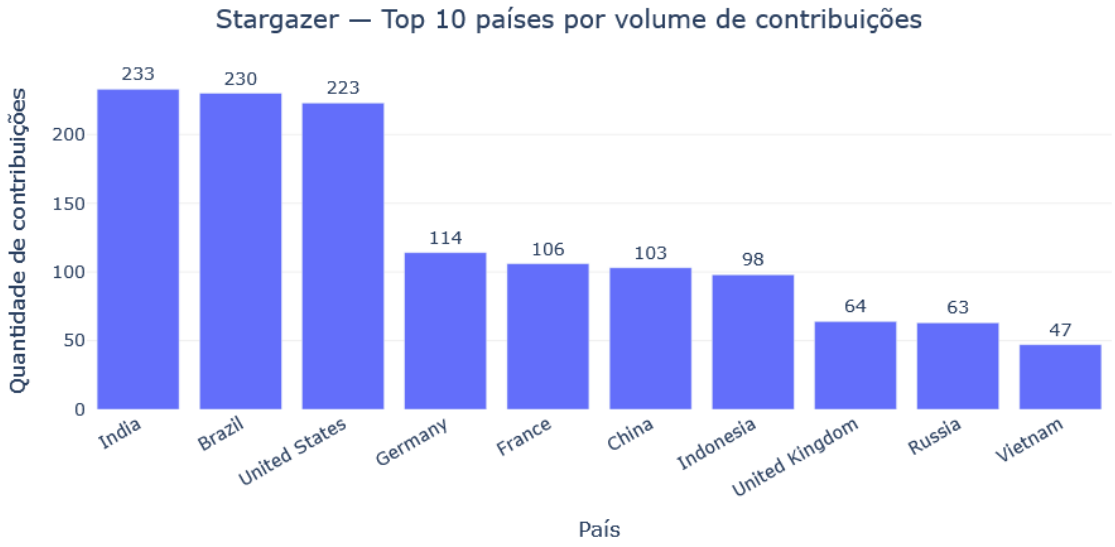

Figure 4. Top 10 países com maior número de usuários que marcaram estrela.

Reactions. A Figura 5 mostra o mapa de calor dos usuários que reagiram a conteúdos do repositório. Embora a distribuição global seja mais limitada que a dos *Stargazers*, ainda há boa diversidade geográfica, com destaque novamente para Índia, Estados Unidos e Brasil. Países europeus e do sudeste asiático aparecem em níveis intermediários.

A Figura 6 exibe o ranking das reações, que reforça a concentração das atividades sociais em poucos países líderes, mas ainda com participação dispersa em diversas regiões.

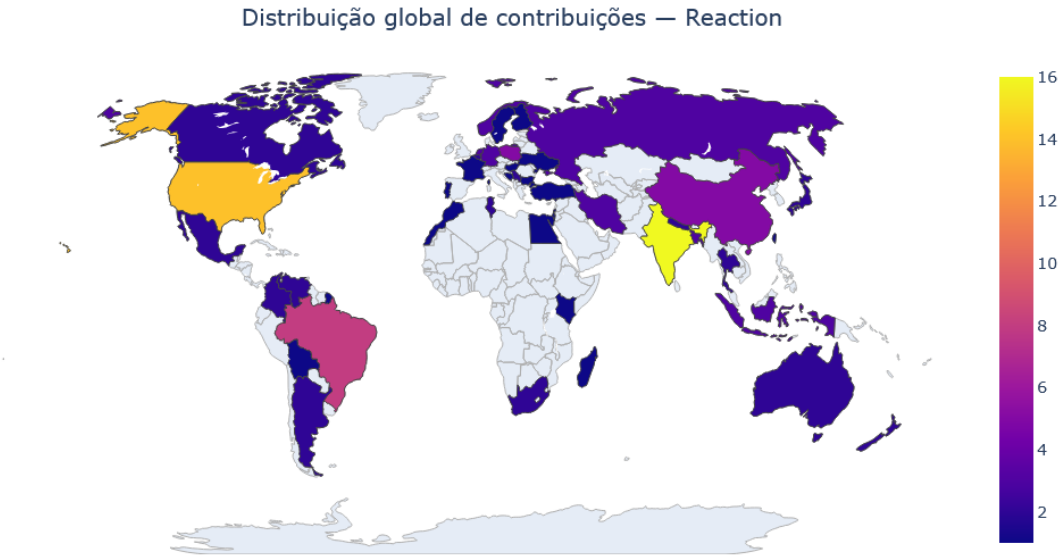

Figure 5. Distribuição geográfica de usuários que realizaram reações.

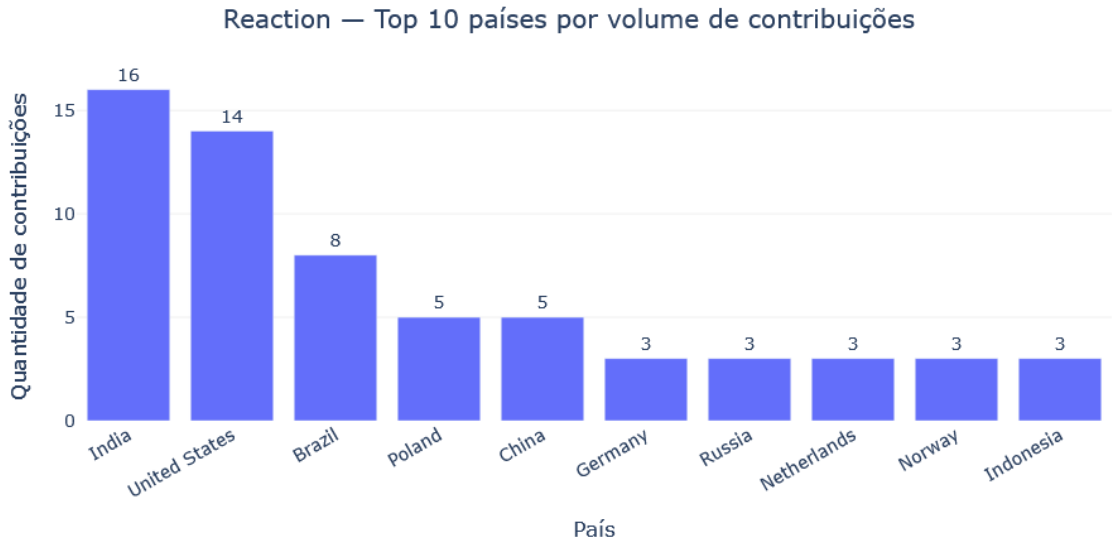

Figure 6. Top 10 países com maior número de usuários que realizaram reações.

Commits. A Figura 7 apresenta a distribuição dos autores de commits. Diferentemente das atividades sociais, observa-se aqui uma distribuição bastante restrita: apenas um conjunto limitado de países aparece com participação expressiva, sendo Estados Unidos e Índia os principais polos. A maior parte dos países do mapa não possui nenhum autor de commits identificado.

A Figura 8 reforça essa concentração, com um número muito menor de usuários por país em comparação às atividades sociais e com poucos países atingindo mais de três autores.

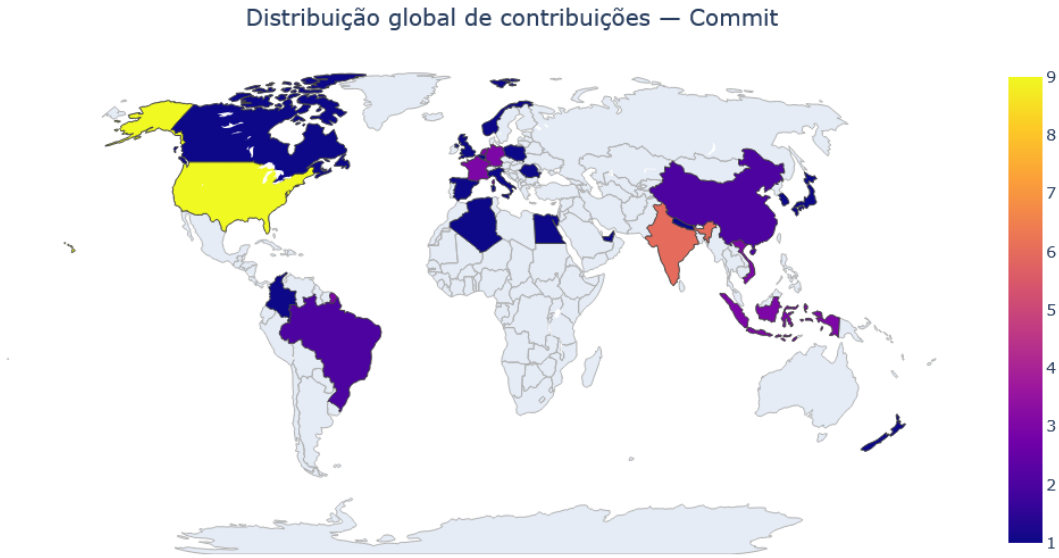

Figure 7. Distribuição geográfica dos autores de commits.

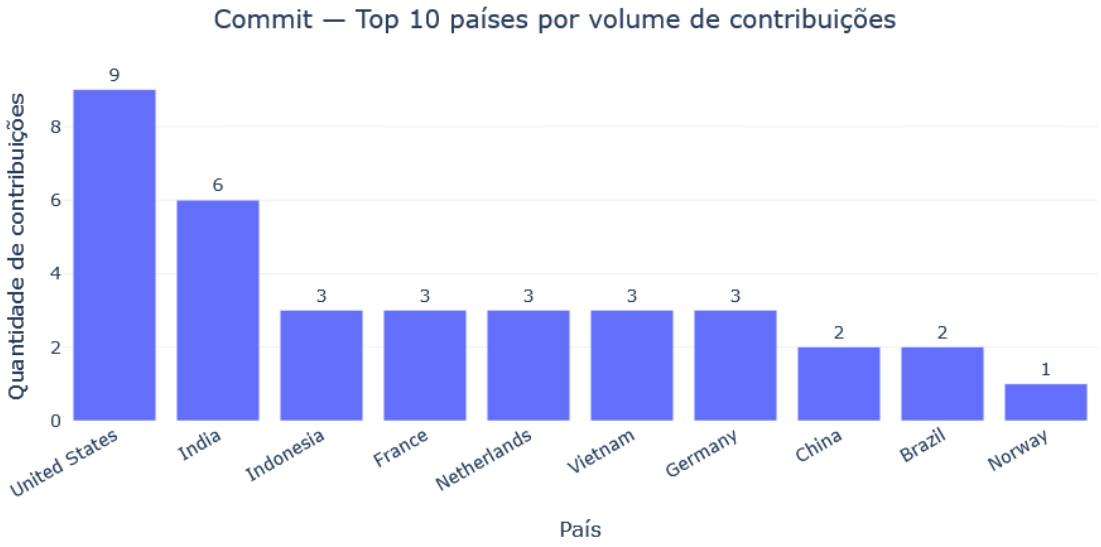

Figure 8. Top 10 países com maior número de autores de commits.

Pull Requests. A Figura 9 mostra a distribuição dos autores de pull requests. Assim como nos commits, observa-se forte concentração em poucos países, mas com ligeiramente mais diversidade geográfica. Estados Unidos e Índia aparecem novamente como os países mais relevantes, seguidos por Países Baixos, Noruega e Brasil com participações menores.

A Figura 10 apresenta o ranking correspondente, evidenciando que, mesmo entre os dez primeiros colocados, a quantidade total de autores por país permanece baixa, reforçando a maior exigência técnica e o maior esforço necessário para esse tipo de contribuição.

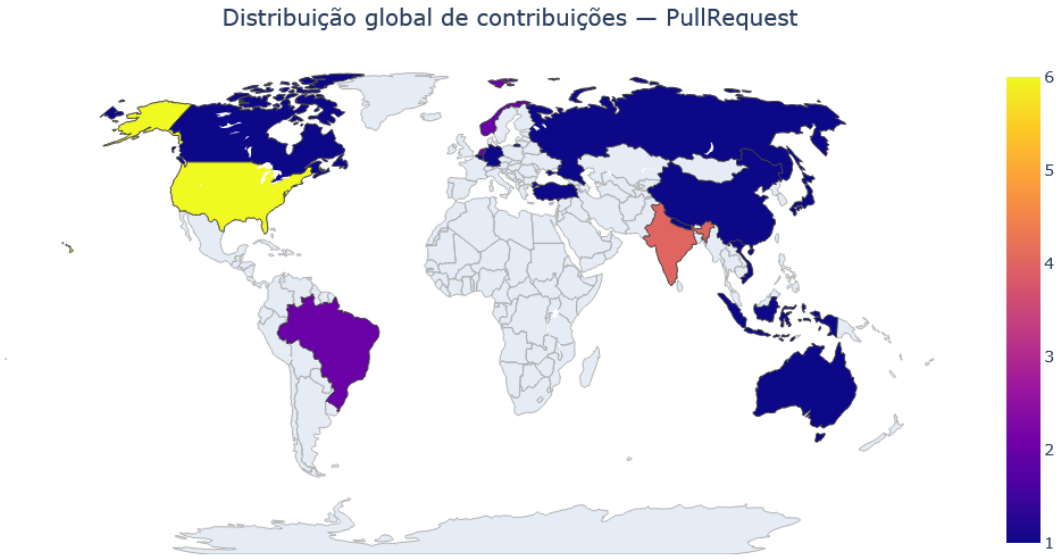

Figure 9. Distribuição geográfica dos autores de pull requests.

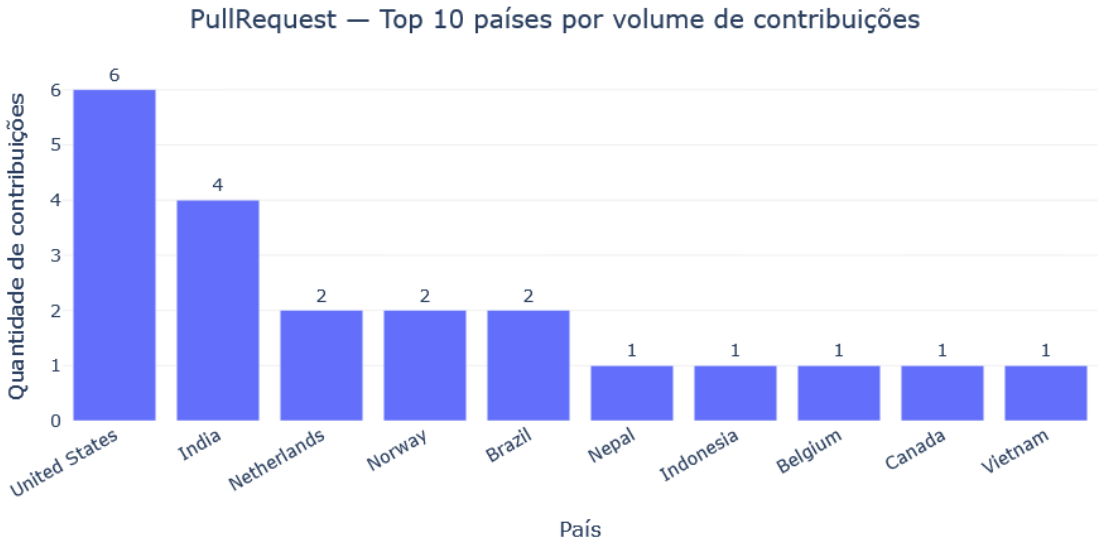

Figure 10. Top 10 países com maior número de autores de pull requests.

Síntese da RQ1. De forma geral, os resultados revelam dois padrões claros: (1) atividades sociais possuem ampla distribuição geográfica e engajam um número muito maior de usuários, enquanto (2) atividades técnicas apresentam forte concentração em poucos países, com barreiras mais evidentes à participação.

Esses achados estão alinhados com a literatura sobre software livre, que aponta para desigualdades regionais persistentes no desenvolvimento colaborativo e para a existência de diferentes perfis de engajamento conforme o tipo de atividade. Assim, a RQ1 evidencia que a participação global neste projeto é ampla em termos de visibilidade e interação social, mas permanece limitada quando se trata de contribuições diretamente relacionadas ao código.

5.3. RQ2 – Diferenças entre atividades realizadas pelos países

A RQ2 investigou se diferentes tipos de atividade apresentam distribuições geográficas distintas. Para responder a essa questão, foram aplicados testes estatísticos não paramétricos (Kruskal–Wallis e Mann–Whitney U), considerando a contagem de usuários únicos por país em cada recurso. O objetivo foi determinar se atividades sociais e atividades técnicas seguem padrões de dispersão similares ou se exibem comportamentos significativamente diferentes.

A Figura 11 apresenta a distribuição da quantidade de usuários únicos por país em cada recurso. Essa visualização permite observar diferenças iniciais entre as atividades antes da aplicação dos testes estatísticos.

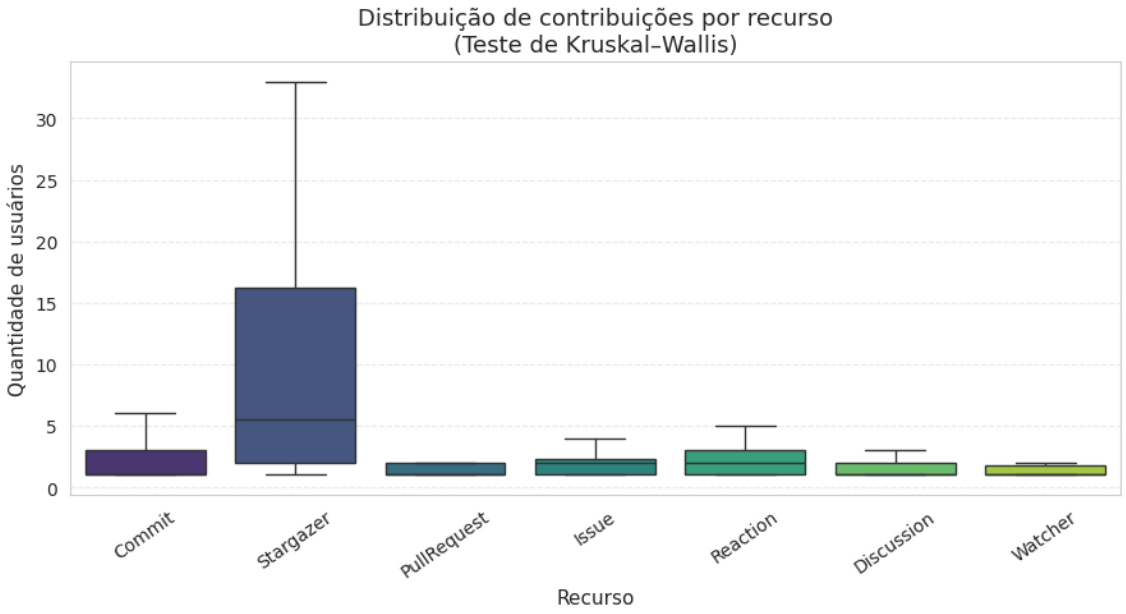

Figure 11. Distribuição da quantidade de usuários únicos por país em cada recurso (boxplot).

A Figura 11 mostra que a atividade *Stargazer* possui uma distribuição substancialmente mais ampla do que os demais recursos. Enquanto países apresentam dezenas ou até centenas de usuários marcando estrela, as atividades técnicas (Commit, PullRequest, Issue e Discussion) exibem valores muito baixos, com mediana próxima de 1 usuário por país. Além disso, a dispersão reduzida e a presença limitada de outliers nesses recursos técnicos indicam um padrão de forte concentração. Já *Stargazer*, por outro lado, apresenta elevada variabilidade e múltiplos outliers, caracterizando um alcance mais global. Esses contrastes visuais antecipam as diferenças estatísticas confirmadas nos testes subsequentes.

O teste de Kruskal–Wallis confirmou que as distribuições dos recursos não são equivalentes:

$$H = 73.36, \quad p = 8.35 \times 10^{-14}.$$

Esse resultado indica que pelo menos um dos recursos apresenta uma distribuição significativamente diferente das demais.

Para identificar quais pares de recursos diferem entre si, foram realizados testes Mann–Whitney U. A Figura 12 apresenta um mapa de calor com os valores de p obtidos em todas as comparações pareadas.

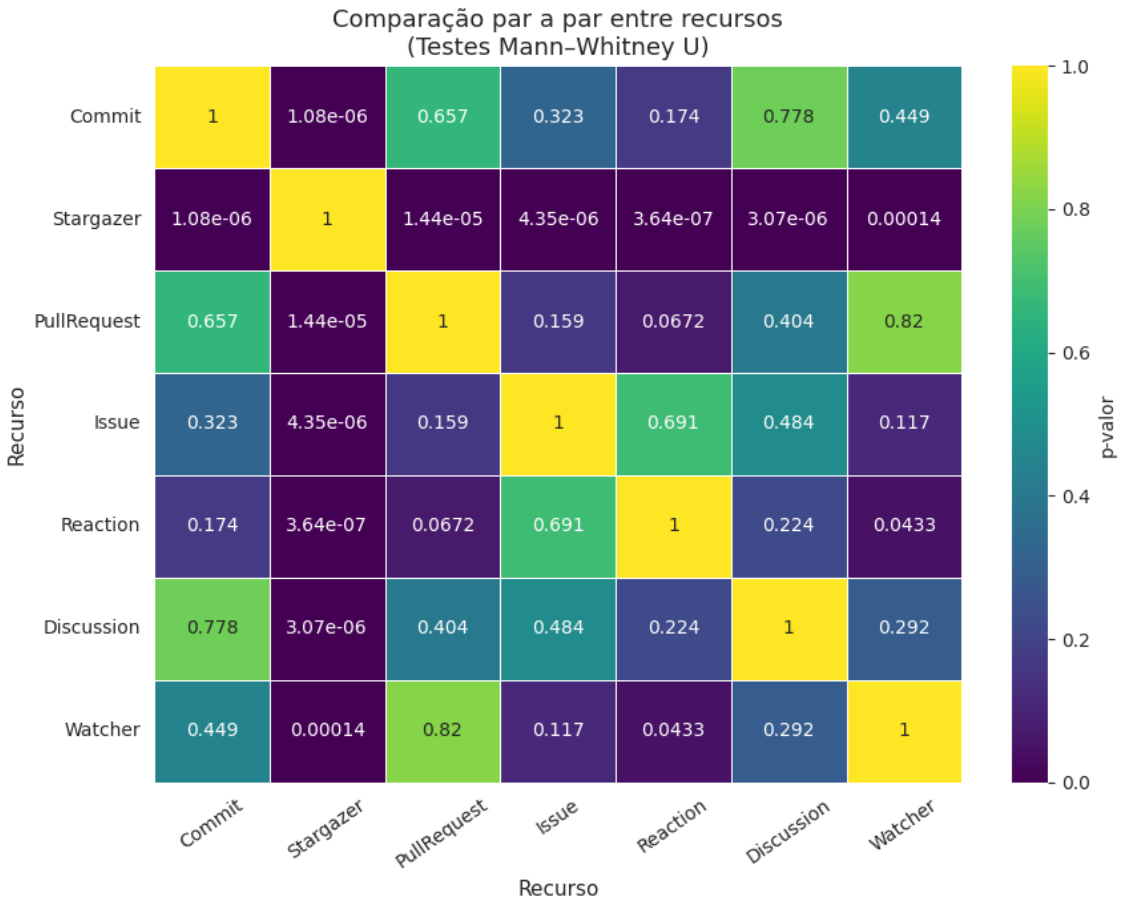

Figure 12. Mapa de calor com os valores de *p* nas comparações pareadas (Mann–Whitney U) entre recursos. Valores menores indicam diferenças estatisticamente significativas.

A Figura 12 permite visualizar claramente quais pares de recursos apresentam distribuições estatisticamente distintas. Observa-se que *Stargazer* difere significativamente de todos os demais recursos, com valores de $p < 0.001$ na maioria das comparações. Isso confirma quantitativamente a diferença já indicada no boxplot. Por outro lado, atividades técnicas como *Commit*, *PullRequest*, *Issue* e *Discussion* exibem valores de p elevados nas comparações entre si, indicando ausência de diferenças significativas e sugerindo padrões geográficos semelhantes. Apenas a comparação entre *Reaction* e *Watcher* apresenta diferença marginalmente significativa.

Como complemento visual às análises estatísticas, a Figura 13 apresenta a composição das contribuições nos dez países com maior volume total de atividades. Esse gráfico permite observar como diferentes tipos de recurso contribuem para o engajamento total de cada país.

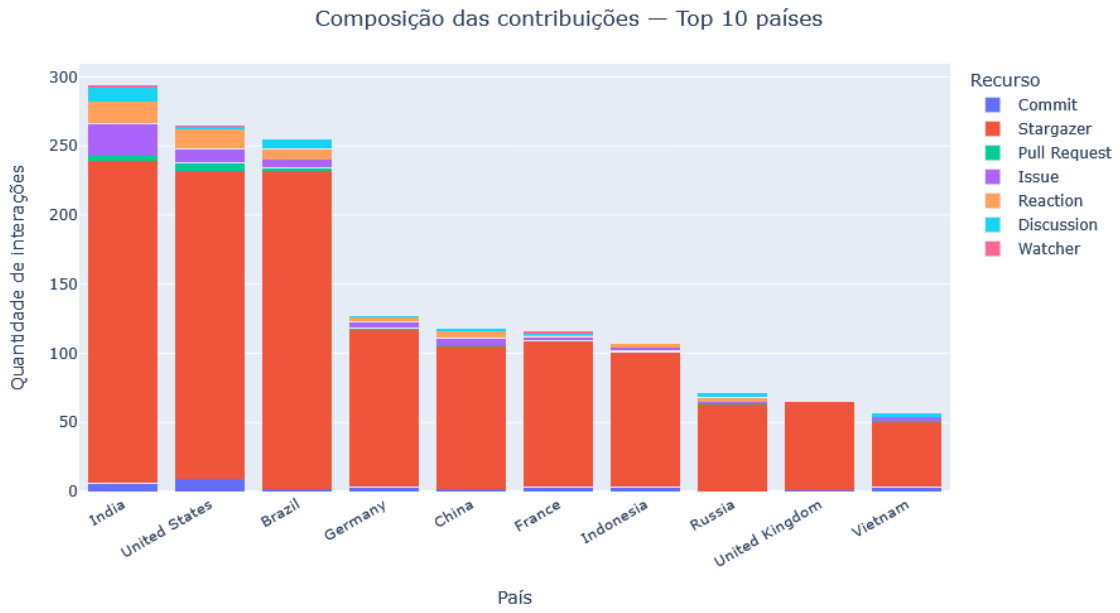

Figure 13. Composição das contribuições por recurso nos cinco países com maior volume total de atividades (barra empilhada).

A Figura 13 evidencia que, mesmo entre os países mais ativos, atividades sociais como *Stargazer* e *Reaction* dominam a maior parte do volume total de contribuições. Recursos técnicos como *Commit* e *PullRequest* aparecem em proporções muito menores, reforçando a ideia de que um número reduzido de países concentra a maior parte da produção técnica. Essa diferença estrutural entre recursos sociais e técnicos é coerente com os padrões identificados nos testes de Kruskal–Wallis e Mann–Whitney.

Em síntese, os resultados da RQ2 confirmam a existência de diferentes perfis de engajamento geográfico entre os tipos de atividade. Enquanto as interações sociais apresentam uma distribuição muito mais ampla e heterogênea entre os países, as atividades técnicas são caracterizadas por forte concentração em poucos polos. Essa distinção reforça a hipótese de que fatores como proficiência técnica, barreiras linguísticas e disponibilidade de tempo influenciam diretamente o tipo de participação realizado no ecossistema open source.

5.4. Análise Complementar: Relação entre Issues e Pull Requests

Além das análises realizadas nas RQs principais, conduzimos uma investigação complementar sobre os padrões geográficos relacionados à abertura e resolução de *issues* e *pull requests* (PRs). Essas duas formas de participação representam etapas centrais do fluxo de desenvolvimento colaborativo: a identificação de problemas e sugestões, seguida da contribuição efetiva em código para solucioná-los. Essa análise permite observar diferenças de engajamento entre países e identificar perfis distintos de contribuição.

Issues. A Figura 14 apresenta a quantidade de *issues* criadas e resolvidas por país. Observa-se que diversos países contribuem com a abertura de *issues*, indicando participação ampla na identificação de problemas. Entretanto, a resolução dessas *issues* é

altamente concentrada: poucos países aparecem com valores expressivos de *issues* resolvidas. Os Estados Unidos, Índia e China se destacam tanto na criação quanto na resolução, enquanto a maior parte dos demais países contribui majoritariamente apenas com a abertura.

Esse padrão revela um fenômeno importante: muitos países atuam como consumidores ou testadores do projeto, reportando problemas e sugerindo melhorias, enquanto um conjunto reduzido de países desempenha o papel de mantenedor, responsável por solucionar essas demandas. A discrepância entre issues criadas e resolvidas reflete diferenças de capacidade técnica, disponibilidade de tempo e envolvimento com o ecossistema do projeto.

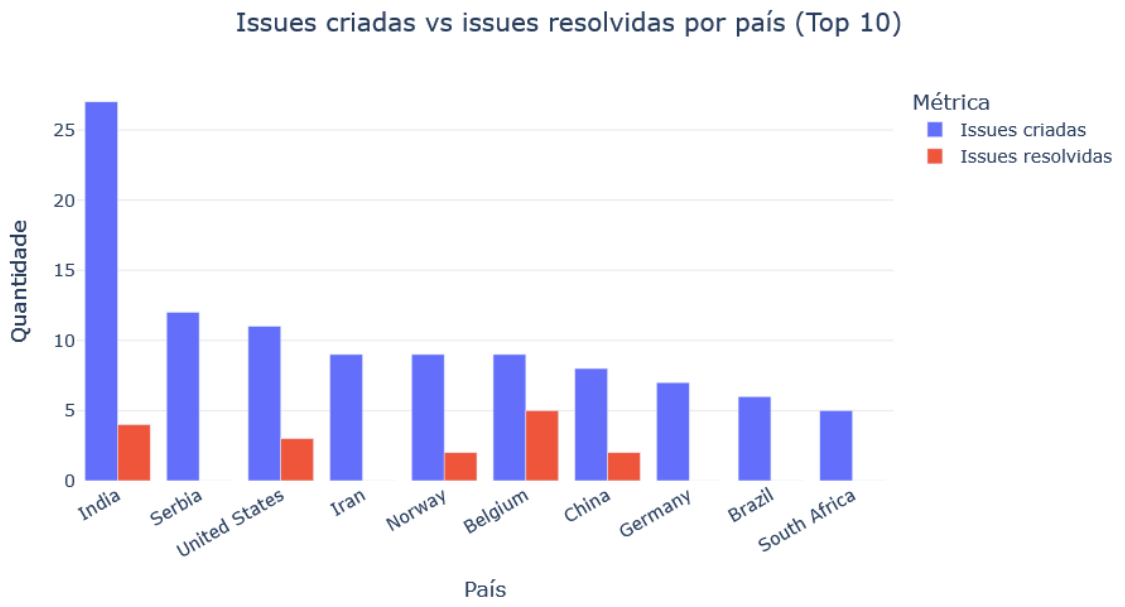

Figure 14. Issues criadas e resolvidas por país.

Pull Requests. A Figura 15 apresenta a quantidade de PRs criadas e mergeadas por país. O padrão observado é semelhante ao das *issues*, porém ainda mais concentrado. Poucos países criam PRs, e um número ainda menor tem PRs aceitas. Países como Estados Unidos e Nepal aparecem como os únicos com volume significativo de PRs mergeadas, enquanto diversos países apresentam PRs criadas, mas nenhuma aceita.

Essa diferença entre criação e merge confirma a existência de barreiras à entrada no processo de contribuição técnica. Criar uma PR exige conhecimentos de código, ferramentas de versionamento e entendimento da base do projeto; ter a PR aceita exige adicionalmente alinhamento com práticas do repositório, padrões de qualidade e interação com os mantenedores. Assim, novamente, apenas um pequeno núcleo de países participa de forma consistente nessas etapas.

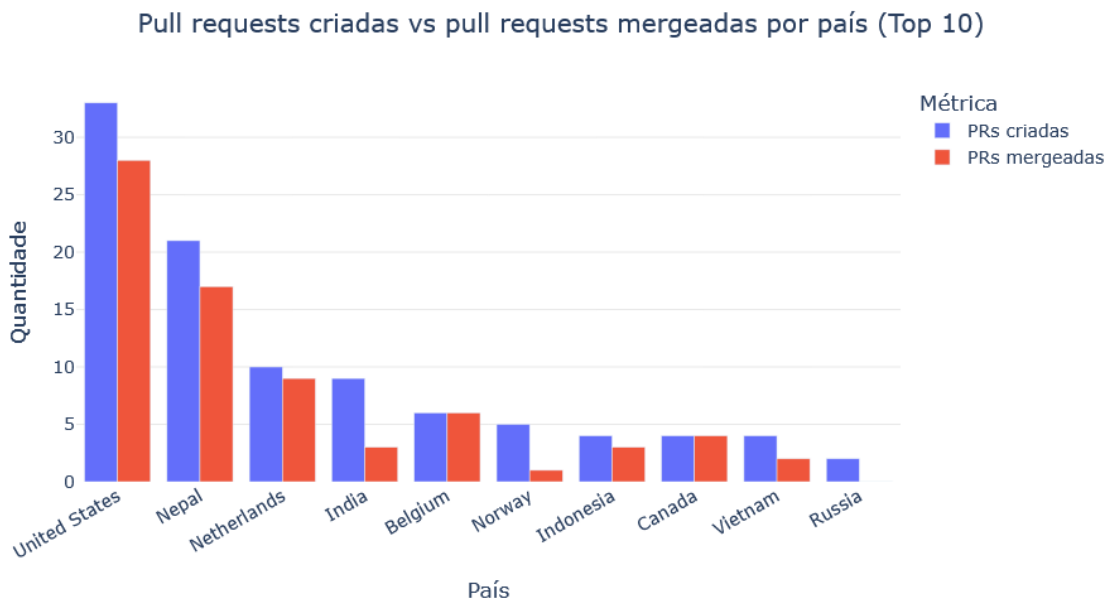

Figure 15. Pull requests criadas e mergeadas por país.

Comparação geral. A Figura 16 sintetiza a relação entre *issues* e PRs ao apresentar, lado a lado, as quantidades criadas e resolvidas de cada recurso. Essa comparação evidencia que muitos países participam ativamente da abertura de *issues*, mas poucos avançam para a criação de PRs, e menos ainda para o *merge*. Os Estados Unidos são o único país com alto volume em todas as etapas (issues criadas, resolvidas, PRs criadas e mergeadas), reforçando seu papel central no processo de desenvolvimento do projeto. Outros países, como Índia e Holanda, apresentam alguma atividade tanto em issues quanto em PRs, mas em menor magnitude.

Essa análise permite identificar três perfis distintos de países:

- **Repórteres**: países que majoritariamente abrem *issues*, mas não contribuem com PRs (por exemplo, Brasil, Alemanha e Vietnam).
- **Contribuidores técnicos**: países que criam PRs, mas com baixa taxa de aceitação.
- **Mantenedores**: países que criam e têm PRs aceitas com frequência, tendo papel central na evolução do projeto (Estados Unidos, Nepal e em menor escala Índia e Holanda).

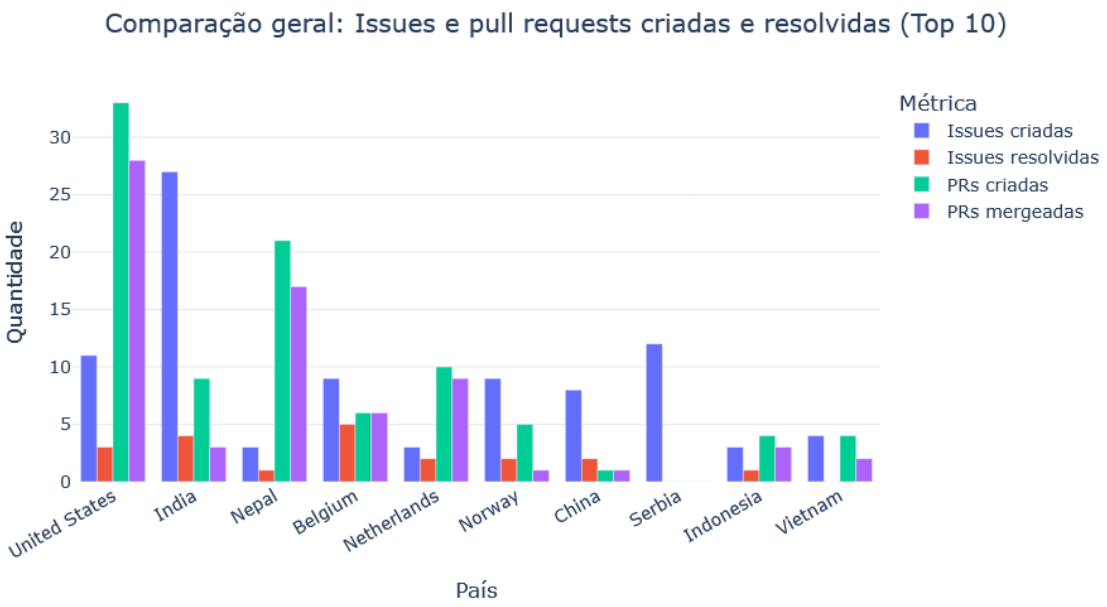

Figure 16. Comparação geral entre issues e pull requests criadas e resolvidas.

Síntese. No conjunto, os resultados mostram que:

- A abertura de *issues* é relativamente distribuída entre países, sugerindo ampla base de usuários e diversidade de feedback.
- A resolução de *issues* e a aceitação de PRs são altamente concentradas, indicando que o trabalho de manutenção é realizado por um grupo muito reduzido de países.
- Países exibem perfis distintos de participação, variando entre repórteres, contribuidores técnicos e mantenedores.
- A forte centralização das atividades técnicas reforça os padrões identificados nas RQs 1 e 2, evidenciando barreiras estruturais à participação técnica global.

Assim, a análise complementar de *issues* e PRs confirma que, embora a participação social e periférica seja global, a contribuição técnica permanece amplamente centralizada em poucos atores, reforçando desigualdades de engajamento no desenvolvimento colaborativo.

5.5. RQ3 – Relação entre proficiência em inglês e contribuições

A RQ3 investigou se existe associação entre o nível médio de proficiência em língua inglesa de um país (EF EPI 2024) e seu volume de contribuições no projeto analisado. Para isso, foi calculado o coeficiente de Spearman, apropriado para avaliar relações monotônicas sem pressupor linearidade ou normalidade na distribuição dos dados.

A Figura 17 apresenta o gráfico de dispersão entre o índice EF EPI e o total de contribuições, incluindo uma linha de tendência por nível de proficiência. Observa-se que a relação é difusa, com países de baixo e médio nível de inglês exibindo tanto valores baixos quanto altos de contribuições.

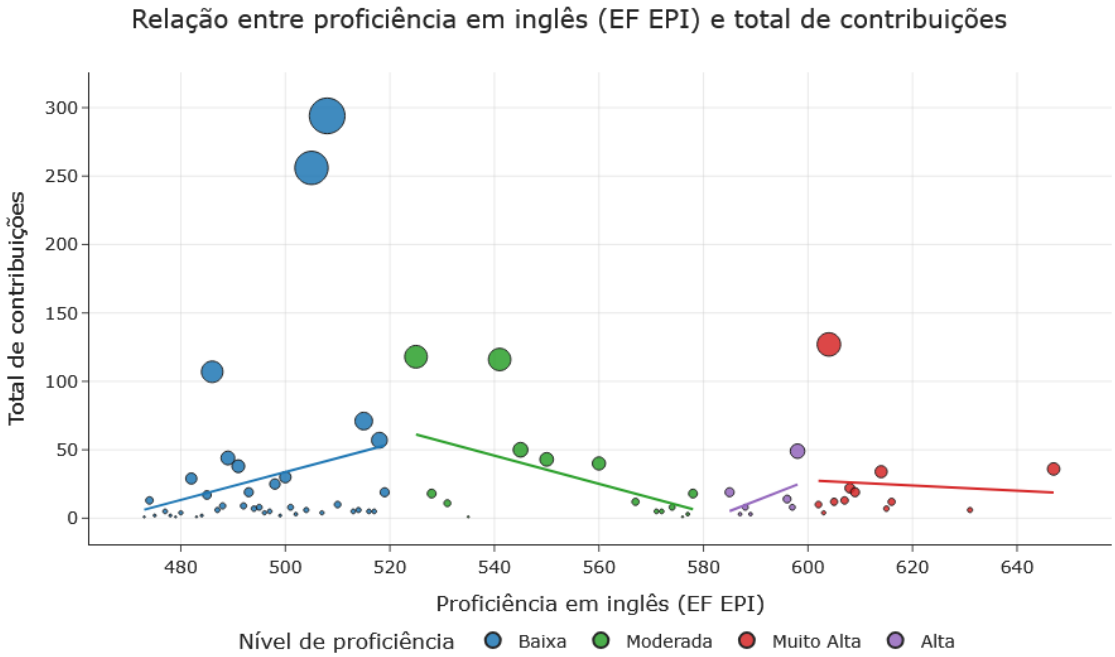

Figure 17. Relação entre proficiência em inglês (EF EPI) e total de contribuições. O tamanho dos pontos representa o volume de contribuições e as cores indicam o nível de proficiência.

O coeficiente de Spearman obtido foi:

$$\rho = 0.231, \quad p = 0.049.$$

Esse resultado indica uma correlação fraca e positiva, sugerindo que países com maior proficiência média em inglês tendem a apresentar níveis mais altos de contribuição, ainda que essa relação seja modesta. A proficiência em inglês, apesar de relevante no ecossistema técnico, não explica sozinha as diferenças observadas entre países.

5.6. RQ4 – Relação entre IDH e volume de contribuições

A RQ4 analisou a relação entre o Índice de Desenvolvimento Humano (IDH) dos países e o volume total de contribuições. Assim como na RQ3, foi utilizado o coeficiente de Spearman, adequado para avaliar relações monotônicas.

A Figura 18 apresenta o gráfico de dispersão entre IDH e total de contribuições. O padrão da nuvem de pontos mostra forte assimetria: a maioria dos países possui baixo volume de contribuições, independentemente do valor de IDH, mas alguns países com IDH elevado concentram picos de atividade.

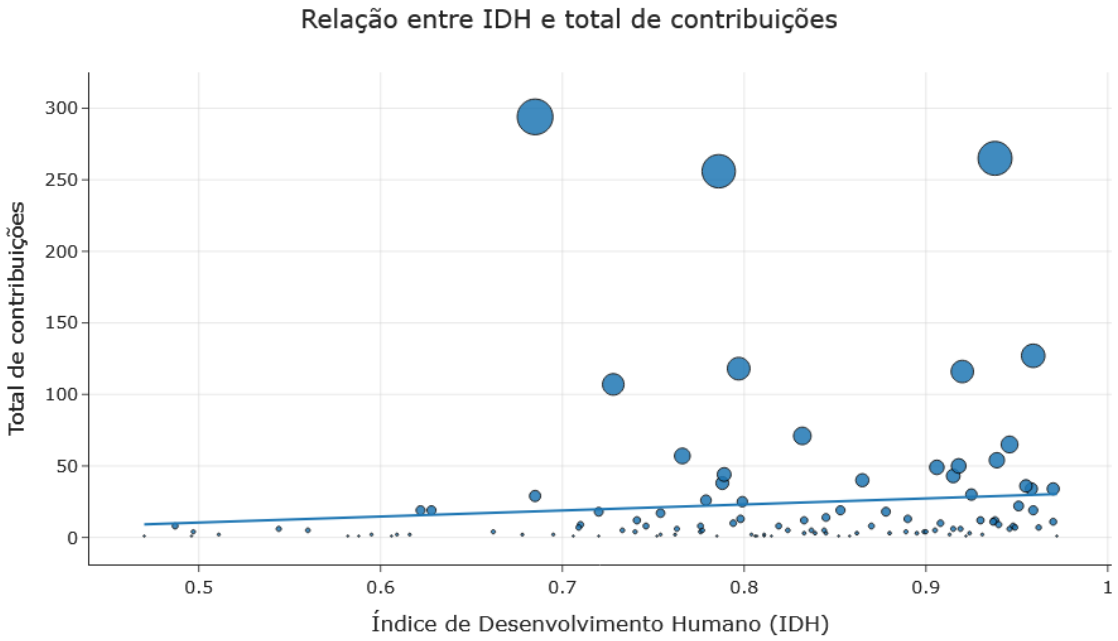

Figure 18. Relação entre IDH e total de contribuições. O tamanho dos pontos representa o volume de contribuições totais por país.

O coeficiente de Spearman encontrado foi:

$$\rho = 0.309, \quad p = 0.001.$$

Trata-se de uma correlação fraca e positiva, mas estatisticamente significativa, indicando que países com maior desenvolvimento socioeconômico tendem a apresentar maior engajamento em atividades open source. Entretanto, a grande dispersão dos pontos mostra que esse fator, isoladamente, não explica toda a variabilidade da participação.

6. Discussões

Os resultados apresentados nas seções anteriores revelam um conjunto consistente de evidências sobre como a localização geográfica, os tipos de atividades realizadas no GitHub e fatores externos, como proficiência em inglês e indicadores socioeconômicos, influenciam a participação em projetos open source. Sob a perspectiva de ecossistemas de software (*software ecosystems*), o repositório analisado pode ser compreendido como um ecossistema sociotécnico, no qual práticas técnicas, mecanismos sociais e características geográficas interagem para moldar padrões de engajamento. Essa visão permite interpretar o projeto não apenas como um repositório de código, mas como uma rede de atores, artefatos e interações distribuídas globalmente. Nesta seção, discutimos essas evidências de forma integrada, conectando as questões de pesquisa (RQs), explorando possíveis explicações para os padrões observados e apontando implicações e direções para estudos futuros.

A análise da RQ1 confirma um cenário amplamente discutido na literatura: a participação em projetos OSS permanece concentrada em poucos países. Estados Unidos,

Índia, Alemanha e Brasil aparecem como polos relevantes em praticamente todos os tipos de atividade analisados. Essa concentração sugere que, apesar da natureza global das plataformas de social coding, o ecossistema open source ainda reflete desigualdades históricas relacionadas a capital técnico, acesso à infraestrutura digital e domínio da língua inglesa. Observa-se, ainda, que atividades de entrada mais simples, como *Stargazer* e *Watcher*, apresentam distribuição geográfica mais ampla, enquanto ações diretamente relacionadas ao desenvolvimento de código, como *Commit* e *PullRequest*, permanecem restritas a um conjunto reduzido de países.

Esse padrão conecta-se diretamente aos achados da RQ2. Os testes estatísticos indicaram diferenças significativas entre as distribuições geográficas associadas aos diferentes tipos de atividade, com um contraste marcante entre recursos de caráter social e recursos técnicos. Esses resultados sugerem a existência de camadas distintas de engajamento no ecossistema. Atividades sociais, em especial aquelas de baixa exigência técnica, tendem a engajar uma base mais ampla e global de usuários. Em contrapartida, atividades técnicas apresentam maior fricção de entrada, exigindo familiaridade com ferramentas de controle de versão, entendimento da base de código, alinhamento com as práticas do projeto e, frequentemente, comunicação em inglês. Quanto maior essa fricção técnica e cognitiva, menor tende a ser o número de participantes e mais concentrada se torna a distribuição geográfica dos contribuidores.

É importante reconhecer que as atividades classificadas como sociais não são homogêneas. Enquanto *Discussions* e *Reactions* envolvem participação ativa e interação entre usuários, a ação de marcar uma estrela possui, em muitos casos, caráter passivo, funcionando como um mecanismo de favoritação ou acompanhamento do projeto. Ainda assim, mesmo ao considerar apenas interações sociais ativas, observa-se uma dispersão geográfica maior do que aquela associada às contribuições técnicas. Essa diferença reforça a robustez da distinção entre camadas de engajamento e sugere que a acessibilidade das ações desempenha papel central na ampliação ou restrição da participação global.

As RQs 3 e 4 reforçam essa interpretação ao analisar fatores contextuais externos. A correlação fraca e positiva entre proficiência média em língua inglesa e volume de contribuições sugere que o idioma exerce influência no engajamento, ainda que não seja um fator determinante por si só. Uma explicação plausível é que a documentação, as discussões técnicas e grande parte da comunicação em projetos open source ocorrem em inglês, o que pode elevar a barreira de entrada para desenvolvedores de países com menor nível de fluência. Assim, países com maior proficiência em inglês tendem a apresentar maior densidade de contribuidores em atividades técnicas, embora exceções relevantes persistam.

De forma semelhante, a associação observada entre o Índice de Desenvolvimento Humano (IDH) e o volume de contribuições, embora também fraca, aponta para a influência de fatores estruturais mais amplos. Países com maior IDH geralmente dispõem de melhores condições educacionais, infraestrutura tecnológica mais robusta e maior acesso a oportunidades de formação em computação, o que pode favorecer a participação em atividades de desenvolvimento, manutenção e revisão de código. No entanto, a dispersão dos dados indica que IDH e proficiência em inglês explicam apenas parte da variabilidade observada, sugerindo que fatores culturais, institucionais, organizacionais e históricos também desempenham papel relevante na configuração da participação em OSS. Essas

interpretações devem ser consideradas à luz das limitações metodológicas discutidas na Seção 6.1, especialmente no que se refere à alta proporção de usuários sem localização identificável e à ausência de normalização per capita.

A integração das quatro RQs permite delinear um cenário no qual múltiplos fatores se sobrepõem. Barreiras linguísticas, técnicas e socioeconômicas condicionam quem participa e em que tipo de atividade; ações de baixo atrito apresentam maior alcance global, enquanto contribuições técnicas permanecem concentradas em países com maior capital tecnológico. Esses padrões sugerem a existência de desigualdades estruturais na formação e na dinâmica das comunidades open source, que vão além das plataformas e refletem desigualdades mais amplas do setor de tecnologia em escala global.

Os resultados também apontam caminhos claros para pesquisas futuras. Estudos comparativos envolvendo múltiplos repositórios podem avaliar a generalização dos padrões observados. A classificação semântica de *issues*, distinguindo, por exemplo, dúvidas de uso, relatos técnicos de bugs e propostas de novas funcionalidades, pode revelar perfis de participação mais finos entre países. A incorporação de métricas normalizadas por população total ou por estimativas da força de trabalho em tecnologia pode permitir comparações mais equilibradas entre contextos nacionais. Por fim, abordagens qualitativas, análises de redes sociais ou estudos sobre interações entre núcleos de mantenedores e a periferia global do ecossistema podem aprofundar a compreensão das dinâmicas de colaboração em projetos open source.

Em síntese, os achados deste estudo indicam que a geografia continua a exercer papel central na dinâmica de participação em software open source, influenciando tanto a distribuição quanto a natureza das contribuições. Embora plataformas como o GitHub reduzam barreiras de comunicação e coordenação, desigualdades estruturais persistem, moldando quem participa, como participa e a partir de onde se participa nos ecossistemas de software contemporâneos.

6.1. Ameaças à validade e limitações

Este estudo apresenta algumas ameaças importantes à validade, que devem ser consideradas na interpretação dos resultados.

Usuários sem país identificado. Como mostrado na Tabela 1, entre 60% e 70% dos usuários não tiveram sua localização geográfica identificada. Há, portanto, o risco de viés sistemático: é possível que desenvolvedores de regiões com maior preocupação regulatória (como países europeus sujeitos ao GDPR) ou com culturas de maior anonimato estejam sub-representados no conjunto "com país identificado". Se esse for o caso, os rankings podem superestimar a participação de países cujos desenvolvedores tendem a preencher a localização no perfil. Como não temos acesso aos países dos usuários não identificados, não é possível corrigir diretamente esse viés, o que limita a validade externa das conclusões sobre concentração geográfica.

Ausência de normalização per capita. As análises deste trabalho utilizam contagens absolutas de usuários e eventos por país. Esse tipo de métrica favorece países muito populosos, como Índia, Estados Unidos, Brasil e China, que tendem a aparecer entre os primeiros colocados em rankings agregados. Não foi aplicada qualquer normalização por população total ou por população estimada de desenvolvedores, o que seria desejável para comparar a *intensidade relativa* de participação entre países. A inclusão de métricas per

capita é um caminho promissor para trabalhos futuros, desde que se obtenham estimativas confiáveis de base populacional de profissionais de TI.

Natureza das *issues*. Na análise complementar, observamos que muitos países abrem *issues*, mas poucos têm *pull requests* mergeadas. Entretanto, não realizamos uma classificação semântica das *issues* em categorias como "dúvida/suporte", "relato técnico de bug" ou "proposta de melhoria". Sem essa distinção, não é possível afirmar se certos países atuam predominantemente como usuários finais que buscam suporte ou como engenheiros que relatam problemas técnicos. A aplicação de técnicas de análise textual para categorizar *issues* é uma extensão natural deste trabalho.

Bots e contas automatizadas. Este estudo não aplicou uma filtragem explícita para identificação ou exclusão de contas automatizadas (*bots*). No entanto, parte dessas contas foi indiretamente excluída do conjunto analisado, uma vez que o processo de normalização geográfica considerou apenas usuários com país identificável. Como muitos bots não possuem campo de localização preenchido ou utilizam descrições genéricas, eles acabam sendo classificados como "não identificados" e, portanto, removidos das análises por país. Ainda assim, não é possível garantir que todas as contas automatizadas tenham sido excluídas, já que alguns bots associados a organizações ou contas institucionais podem possuir localização definida. Assim, eventos gerados por bots podem permanecer na base e inflar artificialmente métricas técnicas de determinados países, configurando uma ameaça à validade interna dos resultados.

Classificação de atividades sociais. Neste trabalho, agrupamos *stargazers*, *reactions* e *discussions* sob o rótulo de atividades "sociais". No entanto, a ação de marcar uma estrela pode ter, em muitos casos, um caráter predominantemente passivo, funcionando como um bookmark ou sinal de interesse, sem necessariamente envolver interação direta entre usuários. Mesmo assim, os resultados mostram que *reactions* e *discussions* também apresentam distribuição geográfica mais ampla do que *commits* e *pull requests*, o que sustenta a conclusão de que atividades de menor fricção tendem a ser mais globais. Ainda assim, reconhecemos que agrupar ações passivas e ativas como "sociais" simplifica um espectro de comportamentos mais diverso.

Apesar dessas limitações, os padrões observados oferecem evidências úteis sobre como a participação se organiza no ecossistema analisado, desde que interpretados com cautela e como ponto de partida para estudos posteriores.

7. Conclusões

Este estudo investigou a relação entre a distribuição geográfica dos desenvolvedores, os tipos de atividades realizadas em um projeto open source e fatores externos como proficiência em inglês e indicadores socioeconômicos. A análise conduzida sobre os dados do repositório *HyDE* permitiu caracterizar o engajamento global em diferentes camadas de participação, especialmente ao integrar métricas derivadas de múltiplos tipos de eventos da plataforma GitHub.

Os resultados referentes à RQ1 mostram que a participação em OSS permanece geograficamente concentrada, com Estados Unidos, Índia, Alemanha e Brasil entre os países mais representativos. Essa concentração aparece de forma consistente em praticamente todos os recursos analisados, indicando que, embora a plataforma permita

colaboração global, a contribuição efetiva continua vinculada a países com histórico de maior desenvolvimento tecnológico.

A RQ2 revelou que diferentes atividades apresentam distribuições geográficas distintas, com diferenças estatisticamente significativas entre alguns dos recursos analisados. O contraste mais evidente está entre atividades sociais, como *Stargazer*, e atividades técnicas, como *Commit* e *PullRequest*. Esses resultados sugerem a existência de níveis de engajamento distintos, indicando que ações de baixo custo, como marcar estrelas ou seguir o projeto, apresentam grande alcance global, enquanto atividades técnicas permanecem mais restritas e concentradas.

As análises referentes às RQs 3 e 4 indicam que tanto a proficiência em inglês quanto o Índice de Desenvolvimento Humano possuem correlação fraca, porém positiva, com o volume de contribuições. Esses resultados mostram que tais fatores influenciam o engajamento, embora não o expliquem completamente. A fluência em inglês pode facilitar a compreensão da documentação, a participação em discussões técnicas e a navegação geral pela plataforma, enquanto o IDH pode refletir acesso à educação, infraestrutura tecnológica e oportunidades de formação. No entanto, a baixa intensidade dessas correlações sugere que outros fatores culturais, institucionais ou comunitários também desempenham papel relevante na participação em OSS.

Integrando as quatro questões de pesquisa, este trabalho evidenciou que a dinâmica da colaboração em software livre é influenciada por múltiplos fatores superpostos: características técnicas, sociais, geográficas e socioeconômicas. Embora as plataformas reduzam barreiras, desigualdades históricas continuam moldando quem colabora, de que forma e em quais tipos de atividade. Compreender essas camadas é fundamental tanto para pesquisadores quanto para equipes de *Developer Relations*, que podem utilizar essas informações para planejar estratégias mais inclusivas e efetivas.

Como direções futuras, recomenda-se a replicação deste estudo em múltiplos repositórios, ampliando a capacidade de generalização dos resultados. Investigando-se repositórios de domínios diferentes, como linguagens de programação, ferramentas de produtividade ou projetos de grande escala institucional, seria possível verificar se os padrões observados se repetem em outros contextos. Abordagens qualitativas também podem complementar as análises quantitativas, ajudando a identificar barreiras sociais, linguísticas e culturais percebidas por contribuidores ao redor do mundo. Por fim, variáveis externas adicionais, como investimentos governamentais em tecnologia, políticas educacionais ou indicadores de conectividade, podem enriquecer análises futuras sobre os determinantes da participação global em software open source.

Em síntese, este trabalho contribui para o avanço da compreensão sobre como localização, atividade e contexto socioeconômico se relacionam na colaboração em projetos de software livre, revelando padrões que podem apoiar tanto pesquisas acadêmicas quanto práticas de gestão e engajamento em comunidades open source.

References

[Almarzouq and Alnahedh 2023] Almarzouq, M. and Alnahedh, M. (2023). An investigation of cross-country differences in open source software activity. In *Proceedings of the 3rd International Conference on Computing and Information Technology (ICCIT)*, pages 611–618.

[Dabbish et al. 2012] Dabbish, L., Stuart, C., Tsay, J., and Herbsleb, J. (2012). Social coding in github: Transparency and collaboration in an open software repository. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*, pages 1277–1286.

[Filho et al. 2015] Filho, F. F., Santos, G. N., Rodrigues, R. S., and Rocha, A. R. (2015). A study on the geographical distribution of brazil's prestigious software developers. In *Proceedings of the 11th Brazilian Symposium on Software Components, Architectures, and Reuse (SBCARS)*, pages 11–20.

[Robles et al. 2008] Robles, G., Gonzalez-Barahona, J. M., and Izquierdo-Cortazar, D. (2008). Geographic origin of libre software developers. In *Proceedings of the 2008 International Conference on Open Source Systems*, pages 199–206.

[Rossi and Zacchiroli 2022] Rossi, A. and Zacchiroli, S. (2022). Geographic diversity in public code contributions: An exploratory large-scale study over 50 years. *Empirical Software Engineering*, 27(3):66.

[Takhteyev and Hilts 2010] Takhteyev, Y. and Hilts, A. (2010). Investigating the geography of open source software through github. In *Proceedings of the 2010 Annual Conference on Human Factors in Computing Systems*, pages 1–10.

[Teixeira and Robles 2015] Teixeira, J. and Robles, G. (2015). Lessons learned from applying social network analysis on open source projects. In *Proceedings of the 2015 48th Hawaii International Conference on System Sciences*, pages 5297–5306.

[Tsay et al. 2014] Tsay, J., Dabbish, L., and Herbsleb, J. (2014). Influence of social and technical factors on success of pull requests in github. In *Proceedings of the 36th International Conference on Software Engineering*, pages 356–366.

[Wachs et al. 2022] Wachs, J., Nitecki, M., Schueller, W., and Polleres, A. (2022). The geography of open source software: Evidence from github. *Technological Forecasting & Social Change*, 176:121478.

[Williams and Williams 2019] Williams, C. M. and Williams, J. R. (2019). Developer relations: How companies foster and sustain external communities. In *Proceedings of the 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, pages 43–52.

[You 2016] You, E. (2016). Between the wires: An interview with vue.js creator evan you. `https://medium.freecodecamp.org/between-the-wires-an-interview-with-vue-js-creator-evan-you-e383cbf5` Acessado em: 07 nov. 2025.

[Zagalsky et al. 2016] Zagalsky, A., Lopes, C. V., Sarma, A., Colussi, D., and Storey, M. A. (2016). How the r community creates and curates knowledge: An extended study of stackoverflow and irc. In *Proceedings of the 2016 13th International Conference on Mining Software Repositories (MSR)*, pages 441–451.