

# How Successful Are Open Source Contributions From Countries With Different Levels of Human Development?

Leonardo B. Furtado, Federal University of Pará

Bruno Cartaxo, Federal Institute of Pernambuco

Christoph Treude, University of Adelaide

Gustavo Pinto, Federal University of Pará

*// In this article we studied whether developers' locations relate to the outcome of a pull request (PR). Our results suggest that developers from countries with low human development indexes perform a small fraction of the overall PRs and are the ones that face rejection the most. //*



**DEVELOPERS IN OPEN** source software (OSS) projects must make decisions about contributions made by other community members, such as whether or not to accept a pull request (PR). Previous studies have shown that factors, such as gender and community status, may influence the chances of contributions being accepted.<sup>1</sup>

In this article, we studied whether developers based in countries with low human development are less likely to succeed in contributing to OSS projects. We used the Human Development Index (HDI) to measure the human development of a country. HDI measures three dimensions of human development: health, education, and income per capita. According to the United Nations (UN) Development Programme (<http://hdr.undp.org/en/content/human-development-index-hdi>), “the health dimension is assessed by life expectancy at birth, the education dimension is measured by mean of years of schooling for adults aged 25 years and more and expected years of schooling for children of school entering age. The standard of living dimension is measured by Gross Domestic Product (GDP) per capita.” HDI is also widely used by the UN<sup>2</sup> and many other international organizations.

To conduct this work, we analyzed 44,630 PRs performed by 14,133 contributors into 20 well-known and well-studied OSS projects. Our investigation suggests that, indeed, developers based in low HDI locations perform fewer PRs and, proportionally, are the ones with the highest rejection rates.

## Method

To conduct our work, we chose OSS projects that are 1) long-lived (that is, more than two years of historical

records), 2) popular (that is, more than 5,000 stars on GitHub), 3) well-studied (that is, studied in other research works), 4) diverse (that is, in terms of their domain), and 5) active (that is, more than 1,000 PRs submitted). We then manually selected a few OSS projects that met these criteria: atom/atom, d3/d3, php/php-src, Microsoft/vscode, django/django, mongodb/mongo, ionic-team/ionic, python/cpython, facebook/react, mozilla-mobile/firefox-ios, apple/swift, Homebrew/brew, scikit-learn/scikit-learn, laravel/laravel, angular/angular, zulip/zulip, facebook/react-native, spyder-ide/spyder, tensorflow/tensorflow, and vuejs/vue.

For each OSS project, we crawled contributors' (for example, names, GitHub handles, location, and so on) and contributions' (for example, PRs performed, PR status, and so forth) data. Overall, we obtained data from 16,836 contributors and 96,592 contributions. We applied the following criteria for analyzing PR data:

- First, we excluded PRs that were integrated by the submitters themselves, thus excluding 22,356 PRs.
- Second, we identified contributors with organizational email addresses and we excluded their PRs. We did this because these developers can work for companies that support these projects and have a large stake in sending PRs, most of which are more likely to be accepted. This excluded 5,544 PRs.
- Third, we excluded PRs from contributors who are part of the project organization or who are part of some organization that funds this project. To find the names of these organizations, we inspected project pages and

looked for backers or funder pages. This led to the exclusion of 18,823 PRs.

After these procedures, we were left with 49,869 PRs from 15,654 contributors.

Since location is not a mandatory field on GitHub, we observed that not all contributors had filled it. We discarded contributors that did not provide their location. Moreover, on GitHub, the location field is a free text form; therefore, GitHub users can fill it with any information. We created a tool that matches the textual information provided in the location field with a location database curated by [simplemaps.com](https://simplemaps.com) (Simplemaps for simplicity). Simplemaps is a database that provides the name of cities, states, countries, and other geographical information. According to its website, they "built it from the ground up using authoritative sources, such as the NGIA, U.S. Geological Survey, U.S. Census Bureau, and NASA" (<https://simplemaps.com/data/world-cities>). Although Simplemaps provides a comprehensive database, some adjustments were still needed. For instance, since we perceived that some GitHub users fill their locations with well-known acronyms (for example, developers often use NY and NYC to mean New York City), we had to enrich the database with them. Using this approach, we were able to categorize 14,133 (90%) contributors that filled the location field. We discarded the 1,521 GitHub contributors whose location we could not infer.

Regarding the contributions, we focused only on closed PRs, due to our interest in analyzing the relationship between the contributors' location and the acceptance/

rejection of the PR. Therefore, we had to rely on PRs that have already passed through the code review process. A total of 44,630 PRs were then selected for analysis; they were submitted between September 2010 and September 2019 (when we collected data).

Regarding the countries' populations and HDI, we used the UN database (<http://hdr.undp.org/en/data>). We adopted the same four-level HDI stratification (very high, high, medium, and low) that the UN traditionally uses in their reports.<sup>2</sup> We considered the year 2018 as the most recent data available.

Our data and tools are available at <https://github.com/LeonardoFurtado/github-user-informations-collector>.

## Results

### Contributions Based on the Location

Overall, developers from the United States, the United Kingdom, and Germany performed the highest number of contributions (20,731 of 44,630 analyzed PRs) whether the contribution was accepted or not. In particular, contributors based in the United States are by far the most active in this regard, performing 14,795 PRs (33% of the total contributions). Table 1 summarizes the top 20 locations of the developers who contributed the most to our studied projects. If we consider countries' HDIs, only four (20%) of the top 20 are not at the very high HDI level (0.800–1.000), namely China, India, Brazil, and Ukraine. This lack of representativeness for lower HDI countries becomes even more evident when we consider the top 20 countries by the number of PRs per country population. All of these countries have very high HDI levels. This

shows that, although countries like China, India, and Brazil are in the top 20 when considering the absolute number of PRs, it is probably due to their large populations. In terms of individual work, Canada is the location that has the highest ratio of PRs per contributor (4.15 PRs/contributor), followed by France (3.77 PRs/contributor), and the United States (3.72 PRs/contributor). On the other hand, Latin America-based and Africa-based developers are significantly less

active than their peers from North America, Europe, and Oceania. Latin American developers performed only 1,183 (2%) of the total contributions in our data set.

#### Acceptance Based on the Location

On average, 32% of the PRs from developers of all countries were accepted. This number rises to 41% when we consider just the top 20 countries in Table 1. Japan-based developers are the ones with the

highest acceptance ratio (51%) followed by United Kingdom-based developers (48%) and Australia-based developers (46%). When we look at the how the PR acceptance ratio relates to countries' HDIs, we can see that the higher HDI levels tend to have a higher PR acceptance ratio on average, as seen in Table 2. An exception are the countries grouped at the low HDI level, which have a PR acceptance ratio similar to the countries grouped under the very high

Table 1. The number of PRs performed per developers' locations.

Country	Number of PRs	Number of PRs/Pop.M.	Number of contr.	Number of PRs/contr.	Acceptance (%)	HDI
United States	14,795	12.91	4,223	3.5	44.45	0.92
United Kingdom	3,179	13.22	887	3.58	48.13	0.92
Germany	2,757	11.01	915	3.01	44.98	0.939
India	2,590	0.51	696	3.72	35.71	0.647
Canada	2,301	14.93	554	4.15	35.77	0.922
France	2,053	8.38	545	3.77	45.93	0.891
China	1,860	0.53	758	2.45	39.09	0.758
Australia	1,212	14.62	364	3.33	46.7	0.938
Japan	1,171	2.96	377	3.11	51.24	0.915
The Netherlands	942	21.64	370	2.55	32.91	0.933
Russia	846	2.23	325	2.6	39.01	0.824
Brazil	780	1.62	339	2.3	32.56	0.761
Poland	650	5.73	217	3	25.69	0.872
New Zealand	576	33.4	157	3.67	39.93	0.921
Sweden	544	19.7	197	2.76	40.07	0.937
Switzerland	432	19.06	162	2.67	45.83	0.946
Taiwan	362	4.77	113	3.2	44.2	0.911
Spain	352	3.38	158	2.23	27.27	0.893
Ukraine	338	3.44	152	2.22	33.43	0.75
South Korea	325	2.4	123	2.64	59.69	0.906

Pop.M, country population (in millions); Contr, contributors; Acceptance, acceptance ratio.

HDI level, on average. However, this discrepancy may occur due to the difference between the number of contributors in countries at the low HDI level (58) and in countries at the very high HDI level (11,344). This is also observable looking at the number of PRs, which is 110 summing up all low HDI level countries, whereas the PR number is 36,972 for the very high HDI level. Consequently, there are countries like Syria, Rwanda, and Senegal with an acceptance rate of 50% but with only two PRs. Developers in Africa, for instance, had 52% of their PRs accepted (although they have performed only 389). South Africa-based developers, in particular, contributed with 117 PRs (with a 59% acceptance rate), although the country has a high HDI (0.705). South American developers faced an even smaller ratio (only 34% of 682 contributions were accepted). Moreover, we noted that 8,794 contributors (19% of the total) performed just one PR (the so-called drive-by-commits or casual contributors,<sup>3</sup> that is, contributors that perform at most one contribution and leave the project). We found that casual contributors are more frequently based in the United States and United Kingdom (43% of developers based in these two countries performed just one PR). In a manual inspection of these casual contributions, we found that a significant number of them are related to improving the documentation (for example, PR 18353 on apple/swift; <https://github.com/apple/swift/pull/18353>), although more complex contributions exist, such as the one from a Poland-based contributor who fixed a bug that occurred during the installation of the atom/atom project on Ubuntu Linux (<https://github.com/atom/atom/pull/3773>).

### Rejection Based on the Location

In terms of rejection, it seems that, regardless of location, having a PR rejected is commonplace. In particular, 59% of the overall PRs were rejected. Interestingly, developers from 29 locations had 100% rejection. These contributors, however, made very few contributions (that is, developers from locations, such as Paraguay, Ethiopia, and Burma, performed at most seven PRs). When manually inspecting these rejected PRs, we noted that some contributors may not have mastered how to use Git/GitHub. For instance, PR 12336 (<https://github.com/scikit-learn/scikit-learn/pull/12336>) on scikit-learn/scikit-learn does not change a single line of code and has a misleading commit message. Moreover, developers from other low HDI locations have contributed more frequently, but they still face a high rejection rate. For example, developers based in Indonesia submitted 107 PRs, with 78 rejected (72%). Bangladesh-based developers submitted 65 PRs, with 87% rejected. When taking into account only the developers' locations with more than 250 PRs, we found that Poland-based developers were the ones that faced the most PR rejections (74%), followed by Spain-based developers (72%), and Brazil-based developers (67%) (Figure 1).

### Related Work

Von Engelhardt and colleagues<sup>4</sup> employed some heuristics on SourceForge (for example, email headers, time zone, and Internet Protocol address) to infer the location of contributors. Bird and Nachiappan<sup>5</sup> employed heuristics, such as the email domain, social networks, and even commit history, to determine the location of the top contributors of Firefox and Eclipse. Spinellis<sup>6</sup> analyzed the FreeBSD operating system by investigating the impact of geographical location on code quality. Vasilescu and colleagues<sup>7</sup> used the GitHub location to infer the presence of female developers on OSS projects. Bjørn and Boulus-Rødje<sup>8</sup> studied the role that infrastructural accessibility plays on the success of a tech start-up in Palestine.

To the best of our knowledge, the work of Rastogi et al.<sup>9</sup> is the closest to our work. However, their work focuses on the developers' location that contributed the most. In our study, however, we shed additional light on developers from low HDI locations that happen to contribute the least or were rejected the most.

### Implications

Our findings indicate that contributors from low HDI countries might

**Table 2. The HDI versus PR acceptance ratio.**

HDI	Acceptance ratio		
	Median (%)	Mean (%)	Standard deviation (%)
Very high (0.8–1)	39.18	35.46	15.79
High (0.7–0.799)	28.57	29.02	24.24
Medium (0.55–0.699)	20.29	30.17	33.63
Low (<0.549)	36.36	32.62	26.48

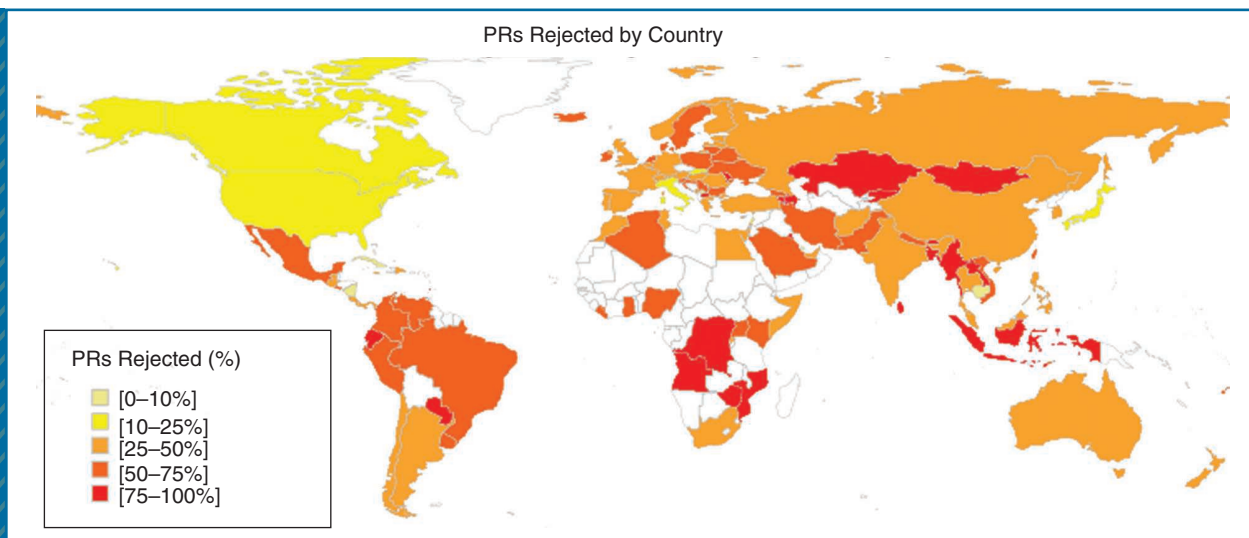


FIGURE 1. PRs rejected per developers' location.

find it hard to contribute to an open source. Given this observation, open source communities might want to promote sprints, hackathons, and other onboarding programs in these locations. Similarly, companies that fund open source communities might also want to fund mentors in low HDI locations. These local activities might help to foster an open source culture in other less wealthy locations.


## Limitations

First, our study is restricted to GitHub; although GitHub is the largest software development platform, we acknowledge that particular countries might have preferences for other platforms. Similarly, developers in some countries may have low participation in certain popular projects because they do not align with the goals of that country or software developers in that country.

Our study is also limited to the number of PRs studied, which clearly does not represent all possible forms of contributions available in OSS projects. Another limitation

is that the location field on GitHub is a free form (that is, it accepts any information). Although we employed some additional steps to make sure that the location exists, we still may have considered developers with inaccurate locations (for example, outdated ones). Finally, there are many other factors that may influence the PR decision-making process. Our work focused on one factor, the location. Therefore, it is unclear how other factors correlate to ours, which we left for future work.

In this article, we studied whether developers' locations have any correlation to PR decision making. We mined data from 44,000 PRs performed in 20 popular OSS projects. We report three main findings. First, developers based in high HDI locations, such as the United States, the United Kingdom, and Germany, are the ones that contribute the most. Second, in terms of acceptance, again, developers based in high HDI locations, such as Japan

and the United Kingdom, have the highest acceptance ratios. Third, in terms of rejection, however, developers based in low HDI locations, such as Ethiopia, Burma, and Paraguay, never had any contribution accepted. High rejection rates were also common in other low HDI locations. 

## References

1. J. Tsay, L. Dabbish, and J. Herbsleb, "Influence of social and technical factors for evaluating contribution in GitHub," in *Proc. Int. Conf. Software Engineering*, 2014, pp. 356–366. doi: 10.1145/2568225.2568315.
2. "Human development report 2019: Beyond income, beyond averages, beyond today: Inequalities in human development in the 21st century," United Nations, New York. [Online]. Available: <http://hdr.undp.org/sites/default/files/hdr2019.pdf>
3. G. Pinto, I. Steinmacher, and M. A. Gerosa, "More common than you think: An in-depth study of casual contributors," in *Proc. Int. Conf. Software Analysis, Evolution and Reengineering*, 2016, pp. 112–123.





**LEONARDO B. FURTADO** is an undergraduate student at the Federal University of Pará, Belém, Pará, 66075-110, Brazil. His research interests include empirical software engineering. Further information about him can be found at [leonardofurtado.com](http://leonardofurtado.com). Contact him at [srleonardofurtado@gmail.com](mailto:srleonardofurtado@gmail.com).



**CHRISTOPH TREUDE** is a senior lecturer in the School of Computer Science at the University of Adelaide, Adelaide, SA 5005, Australia. His research interests include advancing collaborative software engineering through empirical studies and the innovation of tools and processes that taking into account the wide variety of artifacts available in software repositories. He received his Ph.D. in computer science from the University of Victoria, Canada, in 2012. He is an Australian Research Council Discovery Early Career Research Award Fellow. Further information about him can be found at <https://ctreude.ca>. Contact him at [christoph.treude@adelaide.edu.au](mailto:christoph.treude@adelaide.edu.au).



**BRUNO CARTAXO** is an associate professor at the Federal Institute for Education, Science, and Technology of Pernambuco, Pernambuco, Paulista/PE CEP: 53.441-600, Brazil. He received his Ph.D. in computer science from the Center of Informatics at the Federal University of Pernambuco in 2018. His research interests include conducting pure and applied research in the broad area of software engineering and technology transfer. Further information about him can be found at <http://brunocartaxo.com>. Contact him at [email@brunocartaxo.com](mailto:email@brunocartaxo.com).



**GUSTAVO PINTO** is an assistant professor of computer science at the Federal University of Pará, Belém, Pará, 66075-110, Brazil. His research interests focus on the interactions between people and code, spanning the areas of software engineering and programming languages. He received his Ph.D. from the Federal University of Pernambuco, Brazil, in 2015. He currently serves as the coeditor in chief of *Journal of Software Engineering Research and Development*. Further information about him can be found at <https://gustavopinto.org/>. Contact him at [gpinto@ufpa.br](mailto:gpinto@ufpa.br).

4. S. von Engelhardt, A. Freytag, and C. Schulz, "On the geographic allocation of open source software activities," *Int. J. Innovation Digital Economy (IJIDE)*, vol. 4, no. 2, pp. 25–39, 2013.
5. C. Bird and N. Nagappan, "Who? Where? What? Examining distributed development in two large open source projects," in *Proc. Int. Conf. Mining Software Repositories*, 2012, pp. 237–246.
6. D. Spinellis, "Global software development in the freeBSD project," in *Proc. Int. Workshop Global Softw. Dev. Practitioner (GSD '06)*, 2006, pp. 73–79. 2006.
7. B. Vasilescu et al., "Gender and tenure diversity in GitHub teams," in *Proc. CHI*, 2015, pp. 3789–3798. doi: 10.1145/2702123.2702549.
8. P. Bjørn and N. Boulus-Rødje, "Infrastructural inaccessibility: Tech entrepreneurs in occupied Palestine," *ACM Trans. Comput.-Hum. Interact.*, vol. 25, no. 5, Oct. 2018 Art. no. 26. doi: 10.1145/3219777.
9. A. Rastogi, N. Nagappan, G. Gousios, and A. van der Hoek, "Relationship between geographical location and evaluation of developer contributions in GitHub," in *Proc. Int. Symp. Empirical Software Engineering and Measurement*, 2018, pp. 22:1–22:8. doi: 10.1145/3239235.3240504.