

Relationship between Geographical Location and Evaluation of Developer Contributions in GitHub

地理的位置とGitHubでの開発者の貢献の評価の関係

著者

Ayushi Rastogi (UCI, Irvine)

Nachiappan Nagappan (Microsoft Research, Redmond)

Georgios Gousios (TU Delft, Netherlands)

André van der Hoek (UCI, Irvine)

論文情報：

ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM) (ESEM '18), October 11-12, 2018, Oulu, Finland. ACM, New York, NY, USA, 8 pages.

発表者：堀口 日向

1. 背景

さまざまな研究 (後ほど説明) により，PRの受け入れには

技術的要因

社会的要因

が影響する

例

- ・ PR投稿者とレビュワーが親しいほど，受け入れられやすい[26]
- ・ コミュニティ内で地位の高い開発者のPRは受け入れられやすい[26]

OSS開発では，異なる国の開発者からオンラインで共同作業

住んでる国の影響を受けるかは研究されていない

2. 既存研究

PRの受け入れに影響のある要因まとめ

1. 開発者自身の性質

技術的および社会的評判[7] [14] [19] [26]
ファイルの変更サイズ，フォロワー数など

← 住んでる国
はここに該当

2. プロジェクト自体の性質

成熟度，人気など[14] [26]

3. PR自体の性質

変更のサイズ，品質，目的など[14] [26]] [30] [24]

3. 目的

OSS開発において
地理的位置がPRの受け入れに
影響があるかどうかを分析



PRの受け入れに関する
知識体系に貢献する

4. 方法 | ロジスティック回帰分析 (1/2)

Characteristic	Measure			
	Tsay et. al [26]	Gousios et. al [14]	Our study	Variable name
Project characteristics				
Maturity	#months in existence	-	#months in existence	proj_months_existence
Team size	#contributors	#active core team members	#active core team members	-
Popularity	#watchers	#watchers	#watchers	proj_watchers
Size of code	-	#non-comment LOC	#non-comment LOC	proj_ncloc
Openness to external contributions	-	%external contributions	%external contributions	proj_external_contribs
Code quality	-	#test LOC per 1,000 LOC	#test LOC per 100,000 LOC	proj_test_loc_per_lloc
	-	#test cases per 1,000 LOC	#test cases per 100,000 LOC	-
	-	#asserts per 1,000 LOC	#asserts per 100,000 LOC	-
Developer characteristics				
Status in community	#followers	#followers	#followers	dev_followers
Status in project	direct commit access (yes=1; no=0)	-	direct commit access (yes=1; no=0)	dev_commit_access
Social closeness	followed integrator prior (yes=1; no=0)	-	followed integrator prior (yes=1; no=0)	dev_followed_integrator
	watched project prior (yes=1; no=0)	-	watched project prior (yes=1; no=0)	dev_watched_project
Experience	-	#previous pull requests on project	#previous pull requests on project	dev_prev_pull_requests
	-	previous pull request success rate	previous pull request success rate	dev_success_rate
	-	-	#months of project participation	dev_months_participation
Pull request characteristics				
Uncertainty of pull request	#comments	#comments	#comments	pr_comments
Size of change	#changed LOC	#changed LOC	#changed LOC	pr_changed_loc
	#files changed	#files changed	#files changed	pr_changed_files
Quality	inclusion of tests (yes=1; no=0)	#changed LOC in tests	inclusion of tests (yes=1; no=0)	pr_test_inclusion
Geographical location				
Country of submitter			country	geo_country
Same country submitter/integrator			same country (yes=1; no=1)	geo_same_country

既存研究で明らかにされた
PRの受け入れに
影響のある要因

モデルで利用する制御変数

モデル上での変数名

本研究で追加した要因

- PR投稿者の国
- PR投稿者と
レビュワーの国が
同じかどうか

4．方法 | ログスティック回帰分析（2/2）

2つの回帰モデルを作成

目的変数はどちらもPRが受け入れられる確率

1． PR投稿者の国のみを説明変数として追加したモデル

PRの受け入れに国が関係するかどうか分かる

2． モデル1+PR投稿者とレビュワーが同じ国かどうかを説明変数として追加したモデル

PRの受け入れにPR投稿者の国とレビュワーの国が関係しているかどうか分かる

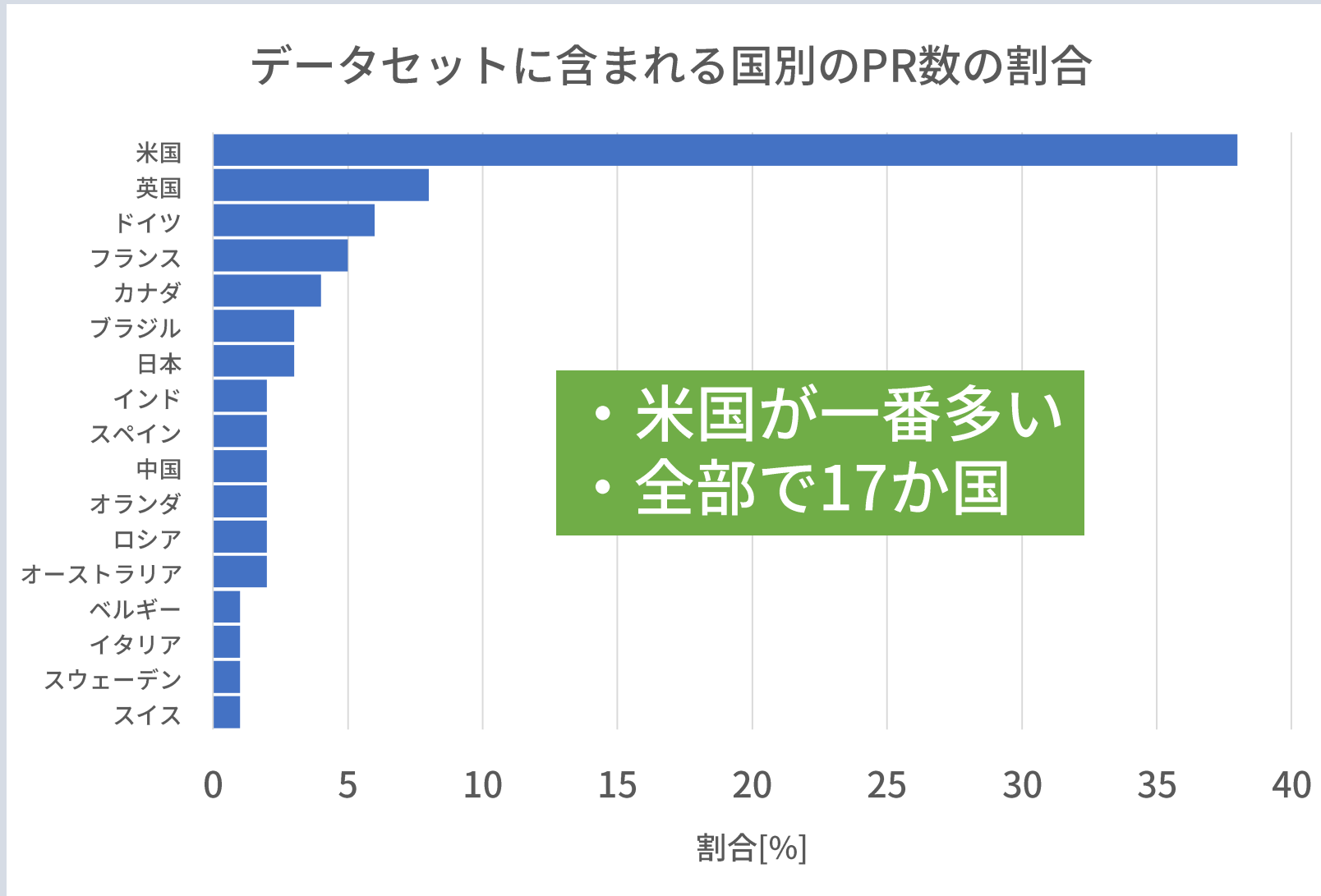
4. 方法 | データ収集 (1/2)

Gousiosら[13]によって公開されたデータセット
計1,069のプロジェクト ⇒ 370,411のPRを抽出

- GHTorrentから個々のPR投稿者の住んでいる国を取得
- 投稿者自身によってマージされたPRを除去

⇒ 最終的に，国名が取得できた計70,740個のPRが分析対象

4. 方法 | データ収集 (2/2)



5. 結果 | モデルの適合度

Table 2: Logistic regression models of factors influencing pull request acceptance [AUC: 0.7]

AUC > 0.5
分析に使える

5. 結果 | 表の見方 (1/2)

geo_country_switzerland	0.38 (0.11)***
geo_country_netherlands	0.26 (0.09)**
geo_country_japan	0.25 (0.08)***
geo_country_united kingdom	0.13 (0.04)**
geo_country_canada	0.12 (0.07)
geo_country_belgium	0.09 (0.12)
geo_country_spain	0.08 (0.10)
geo_country_australia	0.05 (0.07)
geo_country_india	0.02 (0.07)
geo_country_france	0.02 (0.06)
geo_country_russia	-0.06 (0.07)
geo_country_sweden	-0.21 (0.09)*
geo_country_germany	-0.25 (0.04)***
geo_country_brazil	-0.27 (0.06)***
geo_country_italy	-0.31 (0.08)***
geo_country_china	-0.39 (0.09)***

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

- 係数の対数オッズ
PR数が一番多い米国を基準に設定
PRが受け入れられる確率が
 - 係数が正：米国より大きい
 - 係数が負：米国より小さい
- 結果の解釈例
米国からのPR100個が
受け入れられるなら、
日本からのPRは、
 $\exp(0.25) \times 100 \text{個} \simeq 128 \text{個}$
が受け入れられる

5. 結果 | 表の見方 (2/2)

geo_country_switzerland	0.38	(0.11)	***
geo_country_netherlands	0.26	(0.09)	**
geo_country_japan	0.25	(0.08)	***
geo_country_united kingdom	0.13	(0.04)	**
geo_country_canada	0.12	(0.07)	
geo_country_belgium	0.09	(0.12)	
geo_country_spain	0.08	(0.10)	
geo_country_australia	0.05	(0.07)	
geo_country_india	0.02	(0.07)	
geo_country_france	0.02	(0.06)	
geo_country_russia	-0.06	(0.07)	
geo_country_sweden	-0.21	(0.09)	*
geo_country_germany	-0.25	(0.04)	***
geo_country_brazil	-0.27	(0.06)	***
geo_country_italy	-0.31	(0.08)	***
geo_country_china	-0.39	(0.09)	***

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

- 標準誤差
信頼区間を示すのに使う
 $\exp(\text{係数} \pm 1.96 \times \text{標準誤差})$

5. 結果 | PRの受け入れに国が関係するかどうか (1/3)

geo_country_switzerland	0.38 (0.11)***
geo_country_netherlands	0.26 (0.09)**
geo_country_japan	0.25 (0.08)***
geo_country_united kingdom	0.13 (0.04)**
geo_country_canada	0.12 (0.07)
geo_country_belgium	0.09 (0.12)
geo_country_spain	0.08 (0.10)
geo_country_australia	0.05 (0.07)
geo_country_india	0.02 (0.07)
geo_country_france	0.02 (0.06)
geo_country_russia	-0.06 (0.07)
geo_country_sweden	-0.21 (0.09)*
geo_country_germany	-0.25 (0.04)***
geo_country_brazil	-0.27 (0.06)***
geo_country_italy	-0.31 (0.08)***
geo_country_china	-0.39 (0.09)***

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

米国と比べてPRを受け入れる可能性が高い国

- スイス
- オランダ
- 日本
- 英国
- カナダ

5. 結果 | PRの受け入れに国が関係するかどうか (2/3)

geo_country_switzerland	0.38 (0.11)***
geo_country_netherlands	0.26 (0.09)**
geo_country_japan	0.25 (0.08)***
geo_country_united kingdom	0.13 (0.04)**
geo_country_canada	0.12 (0.07)
geo_country_belgium	0.09 (0.12)
geo_country_spain	0.08 (0.10)
geo_country_australia	0.05 (0.07)
geo_country_india	0.02 (0.07)
geo_country_france	0.02 (0.06)
geo_country_russia	-0.06 (0.07)
geo_country_sweden	-0.21 (0.09)*
geo_country_germany	-0.25 (0.04)***
geo_country_brazil	-0.27 (0.06)***
geo_country_italy	-0.31 (0.08)***
geo_country_china	-0.39 (0.09)***

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

米国と比べてPRを受け入れる可能性に差がなかった国

- ベルギー
- スペイン
- オーストラリア
- インド
- フランス
- ロシア

5. 結果 | PRの受け入れに国が関係するかどうか (3/3)

geo_country_switzerland	0.38 (0.11)***
geo_country_netherlands	0.26 (0.09)**
geo_country_japan	0.25 (0.08)***
geo_country_united kingdom	0.13 (0.04)**
geo_country_canada	0.12 (0.07)
geo_country_belgium	0.09 (0.12)
geo_country_spain	0.08 (0.10)
geo_country_australia	0.05 (0.07)
geo_country_india	0.02 (0.07)
geo_country_france	0.02 (0.06)
geo_country_russia	-0.06 (0.07)
geo_country_sweden	-0.21 (0.09)*
geo_country_germany	-0.25 (0.04)***
geo_country_brazil	-0.27 (0.06)***
geo_country_italy	-0.31 (0.08)***
geo_country_china	-0.39 (0.09)***

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

米国と比べてPRを受け入れる可能性が低い国

- スウェーデン
- ドイツ
- ブラジル
- イタリア
- 中国

5. 結果 | PR投稿者とレビュワーが同じ国の場合

geo_same_countryf	0.18 (0.03)***
-------------------	----------------

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

PRを受け入れる可能性が
 $\exp(0.18) \simeq 1.2$ 倍

国ごとにデータを分析
⇒ インド以外のすべての国で、
PRが受け入れられる可能性が2.4%～13.4%上がった
(インドは、受け入れ率が7.4%低かった)

レビュワーが自国からの貢献をより多く受け入れている

6. 考察 (1/2)

米国と比べてPRを受け入れる
可能性が高い国

スイス
オランダ
日本
英国
カナダ

米国と比べてPRを受け入れる
可能性が低い国

スウェーデン
ドイツ
ブラジル
イタリア
中国

← →
違いは？

- 同じ地域，大陸が一貫して高い？ ⇒ No
日本，イギリス，カナダ → 別々の地域
- etc …

結局なんでこうなったかは
この研究では分からない

6. 考察 (2/2)

- ある国では大学生が貢献し，別の国ではほとんどが専門家
⇒ 大きな違いが国ごとに存在するかも？
- なぜ国ごとに差がある（ように見える）のか
- どのように予防できるのか

を知るためには，さらに定性的な研究が必要

補足資料・結果 | 説明変数の逸脱度

変数をモデルに加えた時と
加えなかった時で
どのくらい逸脱度が変わるかを計算
カイ二乗検定で統計的有意差を確認
⇒ 地理的位置は
PRの受け入れられやすさに
影響がある

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			70739	52354.02	
proj_months_existence	1	509.46	70738	51844.56	0.0000
proj_watchers	1	175.49	70737	51669.07	0.0000
proj_external_contribs	1	321.47	70736	51347.59	0.0000
proj_test_loc_per_lloc	1	47.84	70735	51299.75	0.0000
log(proj_ncloc + 1)	1	30.85	70734	51268.90	0.0000
dev_months_participation	1	25.46	70733	51243.45	0.0000
log(dev_prev_pull_requests + 1)	1	1014.64	70732	50228.81	0.0000
dev_success_rate	1	389.62	70731	49839.19	0.0000
pr_test_inclusion	1	22.73	70730	49816.46	0.0000
log(pr_changed_loc + 1)	1	164.02	70729	49652.44	0.0000
log(pr_changed_files + 1)	1	0.11	70728	49652.34	0.7444
pr_comments	1	2.62	70727	49649.72	0.1055
log(dev_followers + 1)	1	57.46	70726	49592.26	0.0000
log(pr_comments + 1)	1	273.45	70725	49318.81	0.0000
dev_watched_project	1	2.82	70724	49315.99	0.0931
dev_followed_integrator	1	6.79	70723	49309.20	0.0092
geo_country	16	143.61	70707	49165.59	0.0000
geo_same_country	1	35.39	70706	49130.20	0.0000