

Are We Speeding Up or Slowing Down? On Temporal Aspects of Code Velocity

Gunnar Kudrjavets
University of Groningen
9712 CP Groningen, Netherlands
g.kudrjavets@rug.nl

Nachiappan Nagappan
Meta Platforms, Inc.
Menlo Park, CA 94025, USA
nnachi@meta.com

Ayushi Rastogi
University of Groningen
9712 CP Groningen, Netherlands
a.rastogi@rug.nl

Abstract—This paper investigates how the duration of various code review periods changes over a projects’ lifetime. We study four open-source software (OSS) projects: Blender, FreeBSD, LLVM, and Mozilla. We mine and analyze the characteristics of 283,235 code reviews that cover, on average, seven years’ worth of development. Our main conclusion is that neither the passage of time or the project’s size impact code velocity. We find that (a) the duration of various code review periods (time-to-first-response, time-to-accept, and time-to-merge) for FreeBSD, LLVM, and Mozilla either becomes shorter or stays the same; no directional trend is present for Blender, (b) an increase in the size of the code bases (annually 3–17%) does not accompany a decrease in code velocity, and (c) for FreeBSD, LLVM, and Mozilla, the 30-day moving median stays in a fixed range for time-to-merge. These findings do not change with variabilities in code churn metrics, such as the number of commits or distinct authors of code changes.

Index Terms—Code review, code velocity, developer productivity

I. INTRODUCTION

One critical goal in the software industry is to develop, review, integrate, and deploy code changes *fast*. The software industry focuses on *increasing the code velocity* [1], [2], [3] using different tools or process enhancements. Various adaptations of Continuous Integration (CI) and Continuous Deployment (CD) [4] have become default practices for most of the current projects in both industry and OSS communities. Similarly, the lightweight Modern Code Review [5], [6] is now a de facto standard process to conduct code reviews. For existing non-agile projects, the initial switch from methodologies such as the waterfall model to CI/CD or mailing lists to contemporary code collaboration tools can have an immediate and noticeable impact on increasing *code velocity* [7], [8]. The longevity of these improvements to code velocity has yet to be thoroughly investigated.

We focus on studying the direction of a change in the duration of various code review periods as the software projects evolve. We intend to determine how the speed of code reviews changes over time. We mine code review data from four different OSS projects: Blender [9], FreeBSD [10], LLVM [11], and Mozilla [12]. We analyze 283,235 code reviews and evaluate how different code review periods (time-to-first-response, time-to-accept, and time-to-merge) trend over time. Our study suggests that the speed of code reviews remains the same as the projects evolve.

II. BACKGROUND AND MOTIVATION

Different interpretations and scopes for code velocity exist. A general definition is “the time between making a code change and shipping the change to customers” [13]. In this paper, we focus on a more quantifiable metric related to the duration of code reviews. We use *time-to-merge* as a proxy metric to quantify how fast code changes propagate. The time-to-merge covers a period from when an engineer publishes a set of code changes that are ready for code review till these changes are merged to the target branch [14].

There are two prevailing and contradictory theories within industry about the direction of code velocity over time. The first hypothesis states that due to an increase in the size and complexity of a system (Lehman’s second law “Increasing Complexity” [15]), e.g., increase in the complexity of communication due to a bigger team, code velocity decreases. A second hypothesis argues that code velocity increases over time. The increase in code velocity is because engineers become more familiar with the code base, interpersonal communication becomes more efficient, and the tooling infrastructure improves.

Another of Lehman’s laws of software evolution is “Continuing Growth” [15]. Lehman’s sixth law states that the “[f]unctional content of a program must be continually increased to maintain user satisfaction over its lifetime” [16]. As a result of additional functionality, it is reasonable to assume that the size of the code base increases. The size of the code base is typically measured in source lines of code (SLOC). Based on our industry experience, we also observe that in conjunction with the new demands on a project, the size of the development teams tends to increase rather than decrease. According to Brooks’ law, “[a]dding manpower to a late software project makes it late” [17]. While the software project does not necessarily have to be late while developing new features, it is unknown how increased code base and team sizes impact code velocity. To investigate this subject further, we formulate the following research question:

RQ: How does a project’s code velocity trend over time?
Does the code velocity increase, decrease, or stay neutral?

TABLE I: The p values from the modified Mann-Kendall (MK) test and magnitude of change for Sen’s slope (Theil-Sen estimator). We present the 30-day moving median and all-time median for different code review periods. Statistically significant Mann-Kendall p value indicates the presence of either an upward or downward monotonic trend. Sen’s slope (presented at a 95% confidence level) identifies the magnitude of the trend per unit time step. The time step is 30 days. The unit for code review periods is hours.

Name	Time-to-first-response				Time-to-accept				Time-to-merge			
	30-day median		All-time median		30-day median		All-time median		30-day median		All-time median	
	MK	Sen’s	MK	Sen’s	MK	Sen’s	MK	Sen’s	MK	Sen’s	MK	Sen’s
Blender	.231		.229		.715		.846		.609		.666	
FreeBSD	< .001	−0.003	< .001	−0.002	< .001	−0.011	< .01	−0.003	.193		< .05	0.006
LLVM	< .001	0.000	< .001	0.000	< .001	−0.005	< .001	−0.002	< .001	−0.009	< .001	−0.002
Mozilla	< .001	−0.001	< .001	−0.001	< .01	0.000	< .001	0.000	< .01	−0.001	.060	

III. METHODOLOGY

A. Choice of data

Section II states that our primary code velocity metric of interest is time-to-merge [14]. We also know that research from both Google [18] and Microsoft [5] finds that time-to-first-response and time-to-accept are additional code review metrics important for developers.

Existing code review datasets focus either on GitHub [19] or Gerrit [20], [21], [22]. By default, Gerrit immediately merges changes once they are accepted. That behavior means that treating time-to-accept and time-to-merge as separate events is neither valuable nor valid. GitHub added the ability to approve changes only in 2016 [23]. We randomly chose 100 GitHub projects and inspected pull requests in those projects. The adoption rate and consistent use of that feature still need to be higher to yield valuable data.

One code collaboration tool that exposes a data model that formally tracks various code review periods is Phabricator [24]. Based on public information about existing Phabricator projects [25], we mine data for four major OSS projects with a multi-year development history. Those projects are Blender, FreeBSD, LLVM, and Mozilla. Our initial dataset contains 283,235 code reviews. We removed the code reviews where the Modern Code Review did not happen. Examples of nonconforming code reviews are the ones that do not contain any lines of code, where the author has accepted their changes, or code reviews that were committed without any acceptance. After filtering and applying consistency checks, the final dataset contains 280,456 code reviews.

B. Statistical analysis

As a first step, we investigate if there is a trend in code velocity. We use the Mann-Kendall test [26], [27] to determine if there is a monotonic upward or downward trend. The monotonic trend [28] means that “the variable consistently increases (decreases) through time, but the trend may or may not be linear” [29]. The null hypothesis (H_0) for the Mann-Kendall test is that there is no monotonic trend. The alternative hypothesis (H_1) is the presence of a monotonic trend. Secondly, if a statistically significant Mann-Kendall

correlation is present, then we calculate the Sen’s slope (Theil-Sen estimator) [30], [31] to evaluate the magnitude of the trend. Sen’s slope indicates the rate of change per unit time step.

How to handle autocorrelation is a challenge for time series analysis. One popular approach is to aggregate the time series to use coarser time granularities such as monthly or yearly samples [32]. Another mitigation is to use a modified Mann-Kendall test to adjust for autocorrelation [33], [34]. We use both techniques to reduce the chance of falsely concluding that a trend is present when it is not.

We conduct these calculations for three code review periods: time-to-first-response, time-to-accept, and time-to-merge. We look at the 30-day rolling (moving) and all-time median. The all-time median is the median of the specific metrics up to a given time. Both metrics help to evaluate the trend of a specific variable [35]. We chose the median values as opposed to the mean values because the median is more resistant to outliers. We use an α level of .05 for our statistical tests.

IV. RESULTS

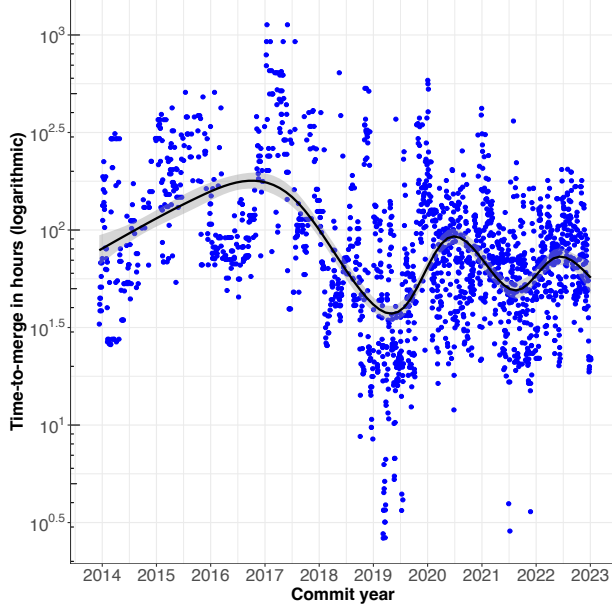
The results from the statistical analysis are displayed in Table I. The main observations from that analysis are the following: (a) except for Blender, there is a statistically significant monotonic trend for all the projects, (b) while a statistically significant trend is present, based on its numeric values and visual representation (see Figure 1), it is minimal, and (c) all the 30-day median slope values are negative or zero, which suggests a *minor increase in code velocity*.

Using a 30-day rolling mean or median is standard practice. Depending on the context, 90-day (quarterly) technical indicators are also used to check for trends. We also calculate the Mann-Kendall and Sen’s slope values as an additional data point using the 90-day rolling median. The conclusions do not change as a result.

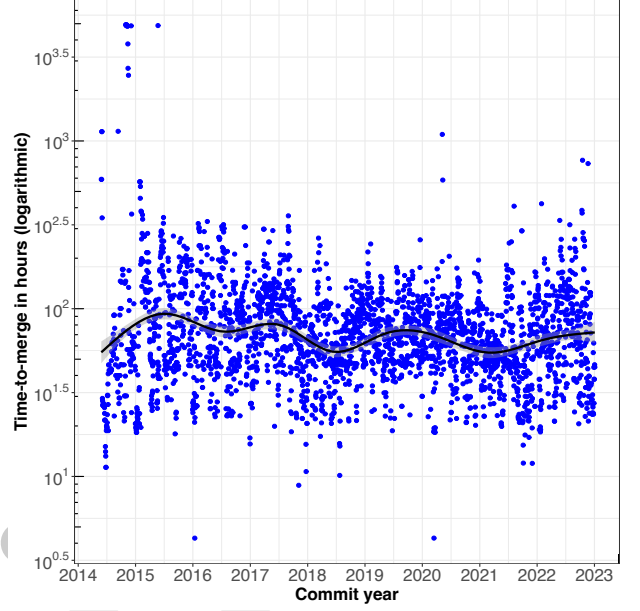
Complimentary to the analysis above, the visualization of the trend for 30-day rolling median time-to-merge is displayed in Figure 1. Based on visual observation, we note that for FreeBSD, LLVM, and Mozilla, most median values stay in a relatively fixed range.

For Blender, there is a noticeable cluster of results that indicate increased code velocity between 2017 and 2019.

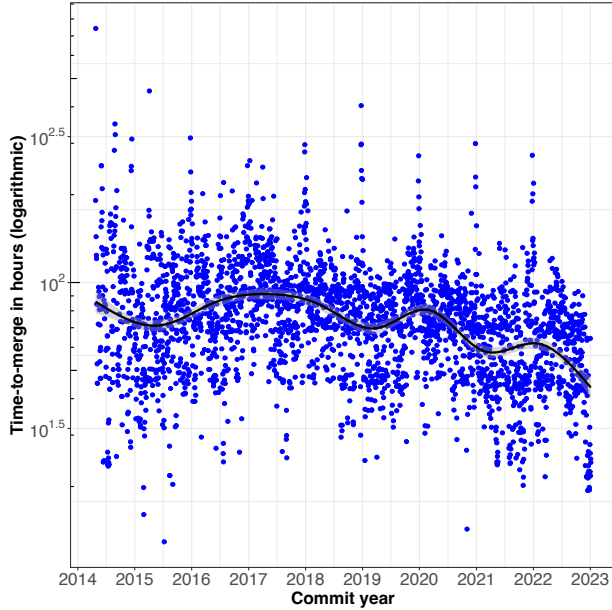
Fig. 1: The trend of 30-day rolling medians of time-to-merge for different projects.



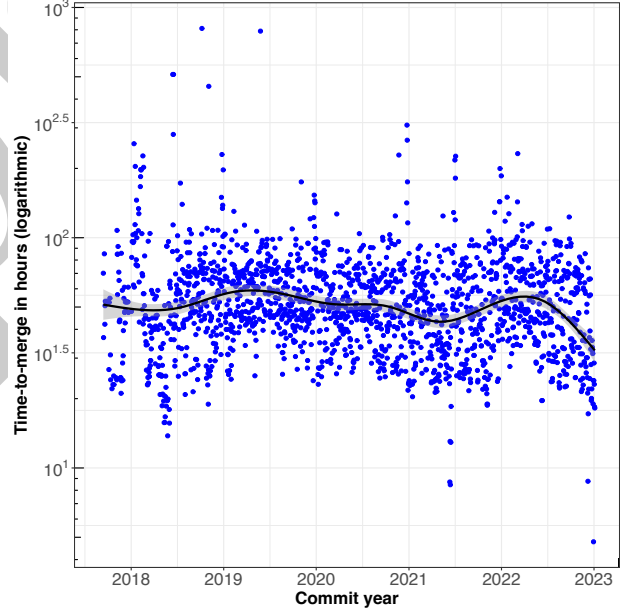
(a) Blender.



(b) FreeBSD.



(c) LLVM.



(d) Mozilla.

To further investigate this result, we inspected 50 random Blender code reviews between 2017 and 2019. We cannot find conclusive evidence that explains the drop in 30-day rolling median values. Based on the public information (“Blender now has a much larger team of people working on core development”), we speculate that a sudden increase in the number of engineers may have caused a temporary increase in code velocity [36]. The trend may have normalized in 2020 because of “unprecedented number of 108 new contributors” [37].

Various data points and events can influence a project’s code

velocity. Only a few of these potential variables are formally tracked. Decisions about feature development, project management, organizational challenges, or changing business priorities are not always documented and available to the public. We have only limited insight into all the confounding variables. To understand the potential impact of metrics associated with project development, we mine the ones we can access. Our findings are displayed in Table II. The only common indicator between the projects is the continuous increase in the source lines of code. We focus on *objective characteristics* that we

can mine from the source control management system and defect database. We look at the annual change in total source lines of code (calculated using scc [38]), number of commits, number of distinct authors, and number of distinct committers. We use each project’s default branch from the original Git repository or an available GitHub read-only mirror.

TABLE II: Median annual increase or decrease percentage in various code churn metrics per project. The scope is a code review period covered in Figure 1. We separate the roles of an author and a committer because they can differ.

Project	Period	SLOC	Commit count	Distinct author	Distinct committer
Blender	2014–2022	6.2%	−2.7%	13.5%	2.3%
FreeBSD	2014–2022	3.5%	−1.2%	0.5%	−4.3%
LLVM	2014–2022	17.4%	6.8%	18.5%	15.6%
Mozilla	2017–2022	11.1%	−4.8%	−6.9%	−13.8%

We investigate the contents of defect-tracking databases associated with the projects we analyze. We find the metrics related to defects to be an unreliable indicator of the project’s workload. We find no formal rules related to defect counts and their implication for the development process. In addition, we do not observe that developers use the defects assigned to them as a primary list of work items.

V. DISCUSSION

Our main finding is that *there is no significant change in the trend for various code review periods*. The lack of trend applies to all the projects regardless of the increase or decrease in commit count, number of distinct authors, and number of distinct committers. The median annual change in code base size for all the projects is an increase of 3–17%. At the same time, the 30-day median and all-time median for different code review periods change only in a second or third decimal place (see Table I). If anything, there is a minor decrease in the duration for time-to-first-response, time-to-accept, and time-to-merge.

The main finding is surprising. The conventional wisdom in software engineering is that communication and processes slow down with an increase in the project’s size and age. There could be multiple explanations for the results that we see. One possibility is that while the code complexity and team size continue to increase, engineers get more familiar and efficient while working in the project’s code base. With time the developers also build better interpersonal relationships that improve communication efficiency. It is reasonable to assume that development infrastructure also improves. Those factors can counteract the time spent on tasks such as complex debugging issues or the comprehension speed for the code sent for review.

Another possibility is that *regardless of the project size, the code velocity stays in a specific range due to factors such as human nature*. Engineers will get to reviewing the code when they get to it. Developers will spend a fixed amount of time on code reviews regardless of their familiarity with

the code or other parallel priorities. The “flatness” or lack of a trend is visible in Figure 1 for FreeBSD, LLVM, and Mozilla. The Blender project has more variability in time-to-merge. The variability can be explained by Blender being the smallest of all the projects we investigate. For example, the median number of commits and distinct authors across the years in Blender are 8,589 and 89, respectively, while in LLVM, it is 29,555 and 592.

One more explanation is related to organizational self-correcting behavior. Each product we study has an informal or formal core team. The core team contains the most active or senior project members. Based on our experience with OSS, the code velocity falling into a specific range can also result from core team members ensuring that code reviews get a timely response and the number of pending issues decreases.

VI. THREATS TO VALIDITY

Our study is subjected to a specific category of threats. One threat relates to application of our findings in other contexts or *external validity* [39]. The projects that we investigate are all OSS. The incentive structure in OSS development is different from industrial projects. In addition, because of the code review granularity that we target, the project selection is limited to the ones that use Phabricator for code collaboration. We do not recommend generalizing these results without further replication in the target environment.

Another threat relates to *internal validity*. This threat type relates to the interpretation of the results and if correct conclusions are drawn from the data. We use the standard recommended nonparametric statistical apparatus to draw our conclusions. The metrics such as rolling 30-day means and median are widely used in quantitative finance as trend indicators [35]. We corroborate our findings by calculating 90-day and all-time medians that indicate a similar trend.

VII. CONCLUSIONS AND FUTURE WORK

We investigate the trend for code velocity in four major OSS projects: Blender, FreeBSD, LLVM, and Mozilla. Our analysis is based on 283,235 code reviews that span, on average, seven years of development activity per project. We find that *code velocity does not decrease over time*. While the size of the code base in these projects increases on median between 3–17% annually, the code velocity either stays the same or slightly improves.

We intend to replicate our findings for other code collaboration tools such as Gerrit and GitHub. While they do not provide the same granularity level as Phabricator, the findings will help invalidate or strengthen our claims. In addition, our observation about the code velocity falling into a specific range is worth additional research.

VIII. DATA AVAILABILITY

The Phabricator data, various R scripts that are used to perform statistical analysis, and relevant SQL queries are available on Figshare.¹

¹<https://figshare.com/s/4558d92adc8d5d262bd6>

REFERENCES

- [1] C. Maddila, S. S. Upadrasta, C. Bansal, N. Nagappan, G. Gousios, and A. v. Deursen, “Nudge: Accelerating Overdue Pull Requests Towards Completion,” *ACM Trans. Softw. Eng. Methodol.*, May 2022. [Online]. Available: <https://doi.org/10.1145/3544791>
- [2] P. Riggs. (2022, Nov.) Move faster, wait less: Improving code review time at Meta. [Online]. Available: <https://engineering.fb.com/2022/11/16/culture/meta-code-review-time-improving/>
- [3] L. Chen, P. C. Rigby, and N. Nagappan, “Understanding Why We Cannot Model How Long a Code Review Will Take: An Industrial Case Study,” in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2022. New York, NY, USA: Association for Computing Machinery, 2022, pp. 1314–1319. [Online]. Available: <https://doi.org/10.1145/3540250.3558945>
- [4] M. Fowler. (2006, May) Continuous Integration. [Online]. Available: <https://martinfowler.com/articles/continuousIntegration.html>
- [5] C. Bird, T. Carnahan, and M. Greiler, “Lessons Learned from Building and Deploying a Code Review Analytics Platform,” in *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories (MSR)*. Los Alamitos, CA, USA: IEEE Computer Society, May 2015, pp. 191–201. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/MSR.2015.25>
- [6] C. Sadowski, E. Söderberg, L. Church, M. Sipko, and A. Bacchelli, “Modern Code Review: A Case Study at Google,” in *Proceedings of the 40th International Conference on Software Engineering: Software Engineering in Practice*, ser. ICSE-SEIP ’18. Gothenburg, Sweden: Association for Computing Machinery, May 2018, pp. 181–190. [Online]. Available: <https://doi.org/10.1145/3183519.3183525>
- [7] M. M. Jha, R. M. F. Vilardell, and J. Narayan, “Scaling Agile Scrum Software Development: Providing Agility and Quality to Platform Development by Reducing Time to Market,” in *2016 IEEE 11th International Conference on Global Software Engineering (ICGSE)*, 2016, pp. 84–88. [Online]. Available: <https://doi.org/10.1109/ICGSE.2016.24>
- [8] K. Kaur, M. Khurana, and Manisha, “Impact of Agile Scrum Methodology on Time to Market and Code Quality—A Case Study,” in *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 2021, pp. 1673–1678. [Online]. Available: <https://doi.org/10.1109/ICAC3N53548.2021.9725375>
- [9] Blender Foundation. (2023) Blender’s Phabricator. [Online]. Available: <https://developer.blender.org/>
- [10] FreeBSD Foundation. (2023) FreeBSD’s Phabricator. [Online]. Available: <https://reviews.freebsd.org/>
- [11] LLVM Foundation. (2023) LLVM’s Phabricator. [Online]. Available: <https://reviews.llvm.org/>
- [12] Mozilla Foundation. (2023) Mozilla’s Phabricator. [Online]. Available: <https://phabricator.services.mozilla.com/>
- [13] Microsoft Research. (2019, May) 14th IEEE/ACM International Workshop on Automation of Software Test. Montreal, Canada. [Online]. Available: <https://www.microsoft.com/en-us/research/event/14th-ieee-acm-international-workshop-on-automation-of-software-test/>
- [14] D. Izquierdo-Cortazar, N. Sekitoleko, J. M. Gonzalez-Barahona, and L. Kurth, “Using Metrics to Track Code Review Performance,” in *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*, ser. EASE’17. Karlskrona, Sweden: Association for Computing Machinery, Jun. 2017, pp. 214–223. [Online]. Available: <https://doi.org/10.1145/3084226.3084247>
- [15] M. Lehman, “Programs, life cycles, and laws of software evolution,” *Proceedings of the IEEE*, vol. 68, no. 9, pp. 1060–1076, 1980. [Online]. Available: <https://doi.org/10.1109/PROC.1980.11805>
- [16] M. M. Lehman, “Software Engineering, the Software Process and Their Support,” *Softw. Eng. J.*, vol. 6, no. 5, pp. 243–258, Sep. 1991. [Online]. Available: <https://doi.org/10.1049/sej.1991.0028>
- [17] F. P. Brooks, Jr., *The Mythical Man-Month: Essays on Software Engineering*, 2nd ed. Boston, MA, USA: Addison Wesley, Aug. 1995.
- [18] Google. (2023) Speed of Code Reviews. [Online]. Available: <https://google.github.io/eng-practices/review/reviewer/speed.html>
- [19] G. Gousios, “The GHTorrent dataset and tool suite,” in *Proceedings of the 10th Working Conference on Mining Software Repositories*, ser. MSR ’13. San Francisco, CA, USA: IEEE Press, May 2013, pp. 233–236. [Online]. Available: <https://dl.acm.org/doi/10.5555/2487085.2487132>
- [20] M. Mukadam, C. Bird, and P. C. Rigby, “Gerrit Software Code Review Data from Android,” in *Proceedings of the 10th Working Conference on Mining Software Repositories*, ser. MSR ’13. IEEE Press, 2013, pp. 45–48. [Online]. Available: <https://doi.org/10.1109/MSR.2013.6624002>
- [21] M. Paixao, J. Krinke, D. Han, and M. Harman, “CROP: Linking code reviews to source code changes,” in *2018 IEEE/ACM 15th International Conference on Mining Software Repositories (MSR)*, 2018, pp. 46–49. [Online]. Available: <https://doi.org/10.1145/3196398.3196466>
- [22] X. Yang, R. G. Kula, N. Yoshida, and H. Iida, “Mining the Modern Code Review Repositories: A Dataset of People, Process and Product,” in *Proceedings of the 13th International Conference on Mining Software Repositories*, 2016, pp. 460–463. [Online]. Available: <https://doi.org/10.1145/2901739.2903504>
- [23] C. Wanstrath. (2016, Sep.) A whole new GitHub Universe: Announcing new tools, forums, and features. [Online]. Available: <https://github.blog/2016-09-14-a-whole-new-github-universe-announcing-new-tools-forums-and-features/>
- [24] Phacility. (2021, Jun.) Phacility—Home. [Online]. Available: <https://www.phacility.com/>
- [25] Phabricator. (2021, Jan.) Organizations using Phabricator. [Online]. Available: <https://secure.phabricator.com/w/usage/companies/>
- [26] H. B. Mann, “Nonparametric tests against trend,” *Econometrica*, vol. 13, no. 3, pp. 245–259, 1945. [Online]. Available: <https://doi.org/10.2307/1907187>
- [27] M. Kendall, *Rank correlation methods*, 4th ed. London, England: Hodder Arnold, Jan. 1976.
- [28] R. M. Hirsch, J. R. Slack, and R. A. Smith, “Techniques of trend analysis for monthly water-quality data,” U.S. Geological Survey, Tech. Rep., 1981. [Online]. Available: <https://doi.org/10.3133/ofr81488>
- [29] Pacific Northwest National Laboratory. (2022, Dec.) Mann-Kendall Test For Monotonic Trend. [Online]. Available: https://vsp.pnnl.gov/help/vsample/design_trend_mann_kendall.htm
- [30] P. K. Sen, “Estimates of the regression coefficient based on Kendall’s Tau,” *Journal of the American Statistical Association*, vol. 63, no. 324, pp. 1379–1389, 1968. [Online]. Available: <https://doi.org/10.1080/01621459.1968.10480934>
- [31] R. O. Gilbert, *Statistical methods for environmental pollution monitoring*. Nashville, TN, USA: John Wiley & Sons, Feb. 1987. [Online]. Available: <https://www.osti.gov/biblio/7037501>
- [32] M. C. Coen, E. Andrews, A. Bigi, G. Martucci, G. Romanens, F. P. A. Vogt, and L. Vuilleumier, “Effects of the prewhitening method, the time granularity, and the time segmentation on the Mann–Kendall trend detection and the associated Sen’s slope,” *Atmospheric Measurement Techniques*, vol. 13, no. 12, pp. 6945–6964, Dec. 2020. [Online]. Available: <https://doi.org/10.5194/amt-13-6945-2020>
- [33] K. H. Hamed and A. R. Rao, “A modified Mann-Kendall trend test for autocorrelated data,” *Journal of Hydrology*, vol. 204, no. 1–4, pp. 182–196, Jan. 1998. [Online]. Available: [https://doi.org/10.1016/s0022-1694\(97\)00125-x](https://doi.org/10.1016/s0022-1694(97)00125-x)
- [34] S. Yue, P. Pilon, B. Phinney, and G. Cavadias, “The influence of autocorrelation on the ability to detect trend in hydrological series,” *Hydrological Processes*, vol. 16, no. 9, pp. 1807–1829, 2002. [Online]. Available: <https://doi.org/10.1002/hyp.1095>
- [35] R. J. Hyndman, “Moving averages,” in *International Encyclopedia of Statistical Science*, M. Lovric, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 866–869. [Online]. Available: https://doi.org/10.1007/978-3-642-04898-2_380
- [36] T. Roosendaal. (2019, Dec.) Blender top 16 highlights of 2019. [Online]. Available: <https://www.blender.org/press/blender-top-16-highlights-of-2019/>
- [37] F. Siddi. (2021, Jul.) Blender by the Numbers—2020. [Online]. Available: <https://www.blender.org/news/blender-by-the-numbers-2020/>
- [38] B. E. C. Boyter. (2022, Mar.) Sloc Cloc and Code (scc). [Online]. Available: <https://github.com/boyter/scc/>
- [39] F. Shull, J. Singer, and D. I. K. Sjøberg, *Guide to Advanced Empirical Software Engineering*. London, England: Springer, 2008.