From Lockdowns to Code Commits: COVID-19's Footprint on GitHub

Daniel Samira - 316164417

Ben-Gurion University of the Negev, Israel
Department of Software and Information Systems Engineering
samirada@post.bgu.ac.il

Abstract

The COVID-19 pandemic significantly impacted software development practices on platforms like GitHub. This study analyzes GitHub activity trends from 2017 to 2022, employing statistical and time-series techniques to uncover changes in event volumes and work patterns. Key findings include significant increases in activities such as commit comments, releases, and pushes, as well as a shift toward increased weekend activity, reflecting the flexibility of remote work. These insights highlight how developers adapted to the pandemic's challenges. Our analyses and code are available at: https://github.com/DanielSamira96/FromLockdownstoCodeCommits.

1 Introduction

The COVID-19 pandemic brought unprecedented changes to how individuals and organizations operate, including software development workflows. The shift to remote work drastically altered collaboration, productivity, and engagement patterns in online platforms like GitHub, a central hub for software development. Understanding these changes provides insights into how developers adapted to the challenges posed by the pandemic and its impact on their work habits.

This paper aims to analyze the effects of COVID-19 on GitHub activities using two complementary approaches. First, we conducted an **event volume analysis** to investigate overall trends in GitHub events before and during the pandemic, employing statistical and time-series modeling techniques to identify significant changes. Second, we performed a **day-specific adjusted trend analysis** to capture more granular shifts in work patterns, focusing on traffic changes across different days of the week.

Our findings reveal notable shifts in developer behavior, including increased activity on weekends, reflecting the flexibility introduced by remote work. These results provide a deeper understanding of the pandemic's impact on software development practices and offer valuable insights into adapting to future disruptions in collaborative environments.

2 Background

This section provides an overview of the COVID-19 pandemic's outbreak to establish context. Then, it examines how users interact with GitHub. Finally, it reviews related work in the field.

2.1 COVID-19 Timeline

The COVID-19 outbreak significantly impacted the world in early 2020, prompting widespread measures to curb transmission, including social distancing, mask mandates, travel restrictions, and lockdowns [1]. Many companies transitioned to remote work, either to comply with regulations or as a precaution [2]. For this project, January 1, 2020, is designated as the critical starting point of the COVID-19 pandemic.

2.2 Software Development on GitHub

GitHub (https://github.com/) is a leading platform for collaborative version control and software development [3]. It supports branching and forking [4] to enable code modifications.

Forked repositories allow users to make changes and submit pull requests (PRs) to propose updates to the main repository. PRs are reviewed by the core development team, who provide feedback. Users address comments, iterating until the changes meet the required quality, at which point the PR is approved and merged [5].

Core teams can create tags to reference commits and periodically stabilize the code base through testing and bug fixes, resulting in new releases.

2.3 Related Works

Several studies have analyzed GitHub telemetry during the COVID-19 pandemic. Klotzman *et al.* [6] examined GitHub and StackOverflow events from 2017 to 2020, identifying temporary spikes in event volumes during March and April 2020 but found no significant long-term changes. Lu *et al.* [7] analyzed GitHub metrics from 2015 to 2020, noting that remote work led to increased active repositories but fewer pull requests, reflecting shifts in team productivity.

Pejić *et al.* [8] studied GitHub event trends during the pandemic and observed changes in developer activity patterns. Casanueva *et al.* [9] explored the pandemic's effects on women's contributions to public code, highlighting disparities in participation. Wang *et al.* [10] conducted a large-scale empirical study of COVID-19-themed GitHub repositories, identifying collaboration trends and emerging themes. In another study, they *et al.* [11] presented a preliminary analysis of COVID-19's impact on GitHub developers, focusing on productivity and collaboration challenges during the crisis.

Oliveira *et al.* [12] examined both GitHub and StackOverflow activity during the pandemic, observing shifts in behavior and productivity.

GitHub itself reported [13, 14] that productivity remained consistent or improved during the pandemic, with longer workdays, faster pull request merges, and increased collaboration, especially in open-source projects.

In contrast to the aforementioned studies, our analysis adopts a granular approach by examining changes across individual event types and their temporal patterns. This includes day-specific activity adjustments, allowing us to capture shifts in working habits that were not explicitly addressed in previous studies.

Furthermore, this study employs several advanced statistical techniques togheter to normalize pre- and post-COVID data and assess significant changes. This methodology reveals nuanced patterns in developer behavior, providing deep insights into how remote work has reshaped collaboration and engagement within the GitHub ecosystem.

3 Analysis

This section first gives information regarding the dataset that was used in this paper, from details about how it was collected to what the different types of events it contains are. After that, an analysis of the trends of the events is presented, with commentary regarding how the different changes in the event trends could be explained.

3.1 Dataset

For this project, I used GitHub Archive [15] via BigQuery [16]. GitHub Archive is a collection of GitHub events from February 12, 2011, to the present, aggregated hourly. The data is also available as a public dataset on Google BigQuery, which supports processing with GoogleSQL. I queried the dataset to retrieve daily event counts for 2017-2022.

The query results are shown in Figure 1, revealing 16 distinct event types in the dataset. While there are differences in event volumes, some event types exhibit a clear upward trend over time. Notably, increased variance is observed in the latter half of the plot, beginning in early 2020, coinciding with the onset of the COVID-19 pandemic.

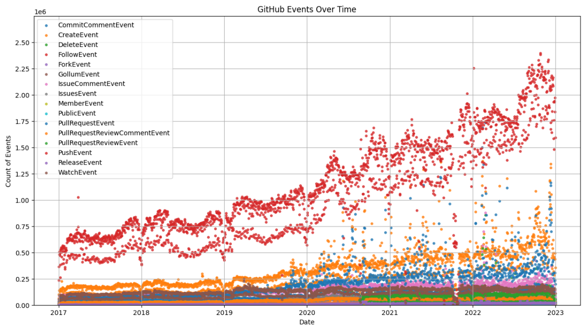

Figure 1: Overview of all GitHub event types over time.

This paper aims to analyze the impact of COVID-19 on GitHub event volumes and assess whether the open-source community maintained its productivity during the pandemic. To achieve this, we defined two time windows:

- **COVID-19 period:** Spanning from 2020 to 2022, covering three full years for consistency in comparison.

- **Pre-COVID-19 period:** Spanning from 2017 to 2019, matching the time window of the COVID-19 period.

Not all event types were present in both periods. To ensure accurate analysis, we examined the first and last occurrences of each event type and their total counts. The summarized data are shown in Table 1. Events like 'FollowEvent' and 'PullRequestReviewEvent', which were not consistently present, were excluded from further analysis.

Table 2 contains short descriptions of each event.

3.2 Event Volume Trends

This subsection examines trends and statistical tests performed on GitHub event volumes before and

Type	First Occurrence	Last Occurrence	Total Count
CommitCommentEvent	2017-01-01	2022-12-31	17,236,618
CreateEvent	2017-01-01	2022-12-31	696,245,413
DeleteEvent	2017-01-01	2022-12-31	149,799,169
FollowEvent	2017-12-01	2017-12-01	95
ForkEvent	2017-01-01	2022-12-31	96,786,951
GollumEvent	2017-01-01	2022-12-31	15,627,774
IssueCommentEvent	2017-01-01	2022-12-31	277,628,365
IssuesEvent	2017-01-01	2022-12-31	124,404,544
MemberEvent	2017-01-01	2022-12-31	16,506,749
PublicEvent	2017-01-01	2022-12-31	12,073,604
PullRequestEvent	2017-01-01	2022-12-31	402,820,967
PullRequestReviewCommentEvent	2017-01-01	2022-12-31	79,095,735
PullRequestReviewEvent	2020-08-18	2022-12-31	63,255,358
PushEvent	2017-01-01	2022-12-31	2,376,223,284
ReleaseEvent	2017-01-01	2022-12-31	22,223,879
WatchEvent	2017-01-01	2022-12-31	270,669,303

Table 1: Summary of GitHub event types: their first and last occurrences and total counts. Rows for `FollowEvent` and `PullRequestReviewEvent` are removed from analysis due to limited data.

Event Type	Description
CreateEvent	Triggered when a git branch or tag is created.
DeleteEvent	Triggered when a git branch or tag is deleted.
PushEvent	Triggered when commits are pushed to a repository.
ReleaseEvent	Triggered when a new release is published.
GollumEvent	Triggered when a wiki page is created or updated.
MemberEvent	Triggered by activities related to repository collaborators.
PublicEvent	Triggered when a private repository is made public.
PullRequestEvent	Triggered by activities related to opening or closing.
PullRequestReviewCommentEvent	Triggered by activities related to leaving a comment.
CommitCommentEvent	Triggered by activities related to comments on specific commits.
IssuesEvent	Triggered by activities related to creating or closing an issue.
IssueCommentEvent	Triggered by activities related to creating or editing a comment.
ForkEvent	Triggered when a user forks a repository.
WatchEvent	Triggered when a user stars a repository.

Table 2: Descriptions of GitHub event types.

during the COVID-19 pandemic. The analysis leverages various time-series modeling and hypothesis testing techniques to assess changes in activity.

3.2.1 Setup

The following statistical models and thresholds were utilized in the analysis:

- **Prophet model:** Prophet is a time-series forecasting model developed by Facebook [17]. For this analysis, Prophet was trained on pre-COVID event data to model expected trends and was then used to compare these forecasts with actual post-COVID event volumes.

- **Paired t-test:** The paired t-test is a statistical method used to compare two related datasets, such as forecasted and actual event volumes [18]. This test evaluates whether the mean difference between the two datasets is statistically significant, with $p < 0.01$ set as the significance threshold in this study.

- **Relative change analysis:** This method calculates the percentage change in event volumes between pre- and post-COVID periods.

Changes exceeding $\pm 10\%$ were considered significant, as smaller variations are likely due to random fluctuations [19].

- **Difference-in-Differences (DID):** DID is a statistical technique commonly used in econometrics to estimate causal effects by comparing changes in outcomes over time between treated and untreated groups [20]. The interaction term's significance ($p < 0.01$) indicates whether the pandemic had a statistically measurable impact on event trends.

- **Interrupted Time Series (ITS):** ITS analysis evaluates the impact of an intervention by analyzing slope and level changes before and after the intervention [21]. This method was applied to detect significant slope changes (ITS effect size > 0.1) in event volumes.

- **ARIMA model:** ARIMA (AutoRegressive Integrated Moving Average) is a statistical model for analyzing and forecasting time-series data [22]. It captures temporal dependencies, trends, and seasonality in the data. In this study, ARIMA was applied to pre-COVID data to assess autoregressive effects, with a threshold of $ARIMA$ effect size > 0.1 used to determine significance.

3.2.2 Results

Table 3 presents the results of the statistical analysis on GitHub event volumes before and during the COVID-19 pandemic. The analysis involved t-tests, relative change assessments, DID, ITS, and ARIMA modeling to evaluate changes in event trends. For example, Figure 2 shows smoothed plots for the `CommitCommentEvent`, highlighting trends before and after the pandemic. Data smoothing was performed using a rolling window size of 10 days to reduce noise and better illustrate patterns. we can observe how the trend of this even was a significant change after the starting point of the corona virus.

Analysis of GitHub event volumes during the COVID-19 pandemic revealed notable trends. Significant increases were observed in certain events, such as `CommitCommentEvent` (130.36%), `ReleaseEvent` (58.39%), and `PushEvent` (22.28%), highlighting heightened collaboration and activity in commit discussions, software releases, and code integration. These trends

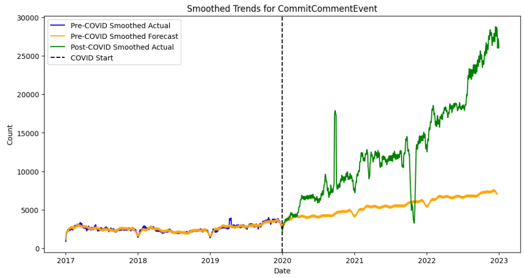

Figure 2: Smoothed trends for `CommitCommentEvent` GitHub event before and during COVID-19.

suggest that developers adapted to remote work by focusing on key collaborative tasks.

Other events, such as `PullRequestReviewCommentEvent` (-3.66%) and `CreateEvent` (-4.13%), showed minor decreases or stability, indicating minimal disruption to activities like branch creation and pull request reviews. Similarly, events like `GollumEvent` (2.71%) exhibited negligible change, reflecting stable engagement with wiki-related activities.

The events highlighted with gray in Table 3—`CommitCommentEvent`, `ReleaseEvent`, `ForkEvent`, `IssuesEvent`, `PushEvent`, `WatchEvent`, `PullRequestEvent`, and `DeleteEvent`—demonstrated significance across all tests (T-Test, Relative Change, DID, ITS, and ARIMA). This highlights their consistent and robust changes during the analyzed period, underscoring their critical role in the observed trends.

3.3 Day-Specific Adjusted Trend Analysis

This analysis focuses on evaluating day-specific changes in GitHub traffic patterns before and after the start of the COVID-19 pandemic.

3.3.1 Setup

To isolate meaningful changes, we introduced a trend ratio, calculated by comparing average traffic across all other days of the week, pre- and post-COVID. This ratio multiply the pre-COVID data for a given day to account for broader trends, providing a normalized baseline for comparison.

Using this adjusted pre-COVID data, we performed paired t-tests against post-COVID data for each event and day of the week. This identified significant changes ($p < 0.01$) in traffic patterns. We also computed the relative change between adjusted pre-COVID and post-COVID means to quantify the magnitude of these shifts.

Event	T-Test Stat.	T-Test P-Val.	T-Test Conc.	Rel. Change (%)	Rel. Change Conc.	DID Eff.	DID P-Val.	DID Conc.	ITS Eff.	ITS Conc.	ARIMA Eff.	ARIMA Conc.
CommitCommentEvent	42.25	$p < 0.0001$	Sig.	130.36	Increase	18.59	$p < 0.0001$	Sig.	18.59	Sig.	5.07	Sig.
PullRequestReviewCommentEvent	-5.23	$p < 0.0001$	Sig.	-3.66	NS	-9.14	$p < 0.0001$	Sig.	-9.14	Sig.	134.90	Sig.
CreateEvent	-5.16	$p < 0.0001$	Sig.	-4.13	NS	132.19	$p < 0.0001$	Sig.	132.19	Sig.	1758.34	Sig.
GollumEvent	0.09	0.9286	NS	2.71	NS	-1.14	0.0003	Sig.	-1.14	Sig.	2.32	Sig.
ReleaseEvent	35.74	$p < 0.0001$	Sig.	58.39	Increase	7.08	$p < 0.0001$	Sig.	7.08	Sig.	1.98	Sig.
ForkEvent	10.51	$p < 0.0001$	Sig.	15.84	Increase	3.56	$p < 0.0001$	Sig.	3.56	Sig.	0.95	Sig.
IssuesEvent	18.94	$p < 0.0001$	Sig.	9.84	Increase	5.47	$p < 0.0001$	Sig.	5.47	Sig.	2.76	Sig.
PushEvent	23.12	$p < 0.0001$	Sig.	22.28	Increase	12.03	$p < 0.0001$	Sig.	12.03	Sig.	4.02	Sig.
IssueCommentEvent	5.23	$p < 0.0001$	Sig.	8.07	Increase	-0.95	0.3542	NS	-0.95	NS	0.82	Sig.
MemberEvent	-2.94	0.0032	Sig.	-14.73	Decrease	-3.12	0.1234	NS	-3.12	NS	1.12	Sig.
WatchEvent	15.45	$p < 0.0001$	Sig.	42.95	Increase	10.09	$p < 0.0001$	Sig.	10.09	Sig.	2.12	Sig.
PullRequestEvent	-8.32	$p < 0.0001$	Sig.	-9.09	Decrease	-5.76	$p < 0.0001$	Sig.	-5.76	Sig.	1.45	Sig.
PublicEvent	12.47	$p < 0.0001$	Sig.	27.99	Increase	6.03	0.3489	NS	6.03	NS	0.67	Sig.
DeleteEvent	6.87	$p < 0.0001$	Sig.	8.67	Increase	4.56	$p < 0.0001$	Sig.	4.56	Sig.	0.89	Sig.

Table 3: Analysis Results of GitHub Event Volume Trends. Events highlighted in gray indicate significance across all tests (T-Test, Relative Change, DID, ITS, and ARIMA). Criteria: T-Test ($p < 0.01$), Relative Change ($\pm 10\%$ threshold), DID ($p < 0.01$), ITS ($effect > 0.1$), and ARIMA ($effect > 0.1$). Sig. = Significant, NS = Not Significant, $p < 0.0001$ denotes very small p-values.

3.3.2 Results

The analysis of day-specific adjusted trends revealed patterns in traffic activity across different days of the week and event types during the COVID-19 pandemic.

Aggregate Trends Across All Events: Figure 3 compares the adjusted pre-COVID and post-COVID average traffic across all events for each day of the week. Notably, there is a significant increase in activity during the weekend (Saturday and Sunday) compared to weekdays, where activity shows a relative decrease. This suggests a shift in work habits, likely influenced by the pandemic and the transition to remote work, where individuals may have adapted their schedules to work more flexibly during weekends.

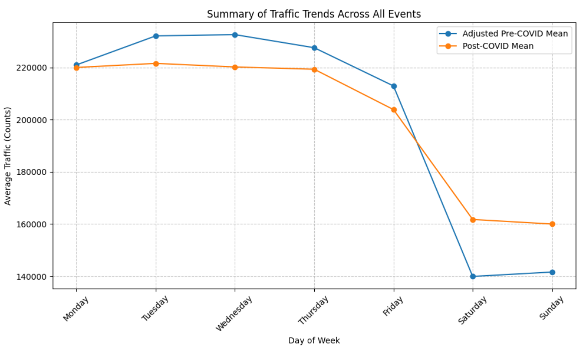

Figure 3: Summary of Traffic Trends Across All Events.

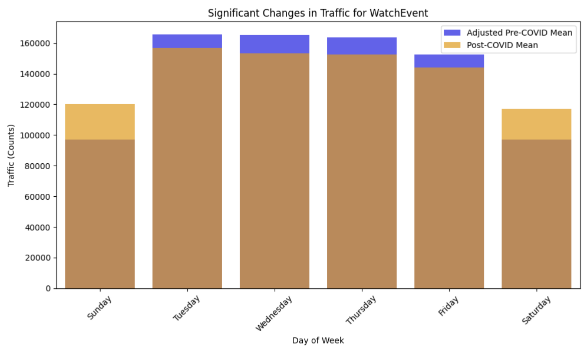

Figure 4: Significant Changes in Traffic for WatchEvent.

Specific Event Analysis: Figure 4 focuses exclusively on the days with significant changes for the `WatchEvent`. The figure illustrates the expected activity for each day compared to the actual post-COVID trends.

Relative Change Heatmap: Figure 5 shows a heatmap of significant relative changes across events and days of the week. The heatmap further emphasizes the higher activity levels during weekends, especially for events such as `PushEvent`, `ReleaseEvent`, and `WatchEvent`.

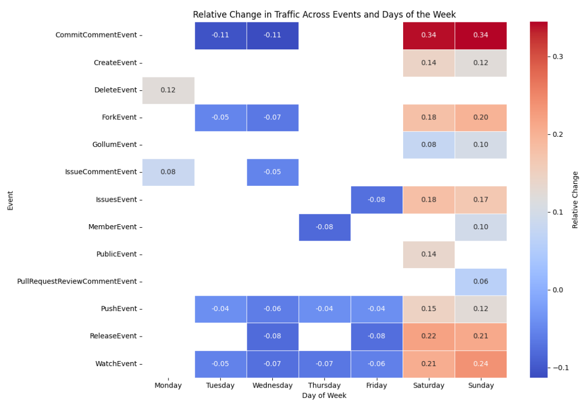

Figure 5: Relative Change in Traffic Across Events and Days of the Week.

These observations highlight the broader impact of the pandemic on work habits, with developers increasingly working on weekends, possibly due to the flexibility afforded by remote work environments.

4 Acknowledgment

I would like to thank Michael Hiestand for his help and support with this project. In this project, I used the ChatGPT-4 model.

References

[1] World Health Organization. Timeline of who's response to covid-19, 2020. URL https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline. Accessed: 2024-12-26.

[2] Statista. Remote work frequency before and after covid-19 in the united states, 2020. URL https://www.statista.com/statistics/1122987/change-in-remote-work-trends-after-covid-in-usa/. Accessed: 2024-12-26.

[3] Yasset Perez-Riverol, Laurent Gatto, Rui Wang, Timo Sachsenberg, Julian Uszkoreit, Felipe da Veiga Leprevost, Christian Fufezan, Tobias Ternent, Stephen J Eglen, Daniel S Katz, et al. Ten simple rules for taking advantage of git and github, 2016.

[4] Jing Jiang, David Lo, Jiahuan He, Xin Xia, Pavneet Singh Kochhar, and Li Zhang. Why and how developers fork what from whom in github. *Empirical Software Engineering*, 22: 547–578, 2017.

[5] Nikola Pejić, Zaharije Radivojević, and Miloš Cvetanović. Helping pull request reviewer recommendation systems to focus. *IEEE Access*, 2023.

[6] Vanessa Klotzman, Farima Farmahinifarahani, and Cristina Lopes. Public software development activity during the pandemic. In *Proceedings of the 15th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–12, 2021.

[7] Xuan Lu, Wei Ai, Yixin Wang, and Qiaozhu Mei. Team resilience under shock: An empirical analysis of github repositories during early covid-19 pandemic. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 578–589, 2023.

[8] Nikola Pejić, Zaharije Radivojević, and Miloš Cvetavnović. Analyzing the impact of covid-19 on github event trends. *Sustainability*, 15 (19):14622, 2023.

[9] Annalí Casanueva, Davide Rossi, Stefano Zacchiroli, and Théo Zimmermann. The impact of the covid-19 pandemic on women's contribution to public code. *Empirical Software Engineering*, 30(1):1–35, 2025.

[10] Liu Wang, Ruiqing Li, Jiaxin Zhu, Guangdong Bai, and Haoyu Wang. A large-scale empirical study of covid-19 themed github repositories. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 914–923. IEEE, 2021.

[11] Liu Wang, Ruiqing Li, Jiaxin Zhu, Guangdong Bai, Weihang Su, and Haoyu Wang. Understanding the impact of covid-19 on github developers: A preliminary study. In *SEKE*, pages 249–254, 2021.

[12] Pedro Almir M Oliveira, Pedro A Santos Neto, Gleison Silva, Irvayne Ibiapina, Werney L Lira, and Rossana MC Andrade. Software development during covid-19 pandemic: an analysis of stack overflow and github. In *2021 IEEE/ACM 3rd International Workshop on Software Engineering for Healthcare (SEH)*, pages 5–12. IEEE, 2021.

[13] GitHub. Octoverse spotlight: An analysis of developer productivity, work cadence, and collaboration in the early days of covid-19, 2020. URL https://github.blog/2020-05-06-octoverse-spotlight-an-analysis-of-de Accessed: 2023-07-26.

[14] Microsoft. Octoverse—finding balance between work and play, 2021. URL https://arxiv.org/ftp/arxiv/papers/2110/2110.10248.pdf. Accessed: 2023-07-25.

[15] GH Archive. Gh archive, 2023. URL https://www.gharchive.org/. Accessed: 2023-05-21.

[16] Google Cloud. Github on bigquery: Analyze all the open source code, 2023. URL https://cloud.google.com/blog/topics/publicdatasets/github-on-bigquery-analyze-all-the-open-source-c Accessed: 2023-12-26.

[17] Sean J. Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018. doi: 10.1080/00031305.2017.1380080.

[18] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908. doi: 10.2307/2331554.

[19] Joe W. Kotrlik and Heather A. Williams. The incorporation of effect size in information technology, learning, and performance research. *Information Technology, Learning, and Performance Journal*, 19(1):43–50, 2001.

[20] Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, NJ, 2008. ISBN 9780691120355.

[21] Alicia K. Wagner, Sharon B. Soumerai, Fang Zhang, and Dennis Ross-Degnan. Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics*, 27(4): 299–309, 2002. doi: 10.1046/j.1365-2710.2002.00430.x.

[22] Gregory C. Reinsel George E. P. Box, Gwilym M. Jenkins and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley Sons, Hoboken, NJ, 5th edition, 2015. ISBN 978-1118675021.