A Hands-on Tutorial on Time Series Imputation with ImputeGAP

Quentin Nater
University of Fribourg
Fribourg, Switzerland
quentin.nater@unifr.ch

Mourad Khayati*
University of Fribourg
Fribourg, Switzerland
mourad.khayati@unifr.ch

Philippe Cudré-Mauroux
University of Fribourg
Fribourg, Switzerland
pcm@unifr.ch

Abstract

Although missing gaps are common in time series data, most existing imputation libraries have a narrow focus. They typically rely on a limited set of techniques and make overly simplistic assumptions about the nature of missing data. Consequently, they fail to model the true intricate complexity of real-world time series. To overcome these challenges, we developed ImputeGAP, a versatile and comprehensive library for time series imputation. ImputeGAP supports a wide range of imputation algorithms and modular missing data simulation, catering to datasets with varying characteristics. It also streamlines imputation analysis with features such as automated hyperparameter tuning, benchmarking, explainability, and downstream evaluation.

In this tutorial, we will provide an engaging hands-on tutorial where you will learn time series imputation using the powerful Python library, ImputeGAP. The session is divided into two parts. In the first part, we will dive into building an end-to-end imputation workflow with the library, where you will explore real-world missingness patterns simulation, leverage automated tuning for optimal imputation, and benchmark imputation techniques—all with extensive customization options. In the second part, we will unlock advanced functionalities, including assessing the impact of imputation on downstream analytics and understanding how time series features influence imputation outcomes. Whether you are a researcher or practitioner, this tutorial will provide you with the expertise to handle missing data in time series. The ImputeGAP library is accessible at: https://imputegap.readthedocs.io.

Keywords

Time series, missing data, data cleaning, imputation library, explainability, downstream impact.

ACM Reference Format:
Quentin Nater, Mourad Khayati, and Philippe Cudré-Mauroux. 2025. A Hands-on Tutorial on Time Series Imputation with ImputeGAP. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3711896.3737601

*Corresponding author.

KDD '25, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3737601

1 Introduction

As the Internet of Things (IoT) rapidly expands, sensor failures are becoming an unavoidable challenge. Even brief disruptions can lead to consecutive gaps in time series data, severely compromising its quality and undermining its reliability for critical applications such as predictive analytics, similarity search, and real-time monitoring. Incomplete or faulty time series often produce inaccurate or suboptimal results, leading to costly errors and inefficiencies [2, 7, 14]. Despite the growing need for robust imputation frameworks, existing libraries [4, 5, 13] fall short in providing a comprehensive, end-to-end imputation pipeline—one that seamlessly integrates dataset collection, realistic missingness pattern simulation, algorithm fine-tuning, and clear interpretability of results.

ImputeGAP is a powerful library that aims to bridge this gap and to ensure reliable time series imputation. Our library offers a diverse range of advanced imputation algorithms, along with a configurable contamination module that simulates real-world missingness patterns. Additionally, ImputeGAP includes tools to analyze the behavior of these algorithms and assess their impact on key downstream tasks in time series analysis, such as forecasting.

Learning Outcomes. The Interactive part of the tutorial will provide participants with experience in applying time series imputation techniques to real-world datasets and missingness scenarios. They will also know how to (1) deploy ImputeGAP to build a full imputation pipeline for time series with various customization options, (2) create a common test-bed for comparing imputation algorithms, (3) assess the effects of data imputation on downstream applications, and (4) provide insights into the imputation behavior.

We expect that the attendees will gain a deep understanding of time series imputation techniques and their underlying mechanisms, including their theoretical foundations, practical implementation, and real-world implications.

Target Audience and Prerequisites. Our tutorial is intended for both beginners and intermediate-level time series practitioners, such as data scientists and software engineers, who frequently engage in data cleaning tasks. It also serves time series researchers, especially those who focus on improving data quality for machine learning applications.

Engaging Experience. Attendees will benefit from in-depth demonstrations and carefully designed step-by-step hands-on materials. Given that most attendees are expected to have access to their personal computers, they will have the opportunity to explore and experiment with ImputeGAP firsthand during the tutorial.

2 Tutorial Outline (3 Hours)

This tutorial is intended to be hands-on using Jupyter Notebooks. It is organized into two main parts: the first focuses on creating

foundational components for time series imputation, while the second delves into advanced analysis to evaluate the imputation results and assess their impact.

2.1 Part I: Building Imputation Pipelines (2h)

The first part begins by presenting the background and motivations underlying ImputeGAP, followed by a demonstration of its key features through quickstart examples to engage the audience. Next, we will guide participants through the steps of designing and implementing an imputation pipeline. The first building block of the pipeline is to stimulate a malfunctioning sensor in a real-world dataset by creating missingness patterns in a systematic way.

Then, we will delve into different techniques to impute the missing data. As imputation techniques, we will cover both classical and more nascent deep learning algorithms, discussing their imputation mechanisms and key properties. We will explore various ways to parametrize those algorithms, including sophisticated techniques from the Ray Tune framework [10]. Additionally, this part will give guidance on benchmarking multiple categories of imputation algorithms and explaining the trade-offs they strike according to various performance metrics.

2.2 Part II: Impact and Explainability (1h)

In the second part, we will showcase the power of ImputeGAP in debugging imputation results, offering invaluable insights into the algorithms' behavior. Using SHapley Additive exPlanations (SHAP) [11], we will unveil how different time series features influence imputation outcomes. Furthermore, we will demonstrate the critical impact of imputation on downstream analysis, with a particular focus on forecasting. By leveraging various time series forecasters, we will highlight the performance gains achieved through advanced imputation compared to leaving the data missing. Lastly, we will cover how our framework can be extended to include new imputation algorithms and downstream models.

3 Related Materials

To support this tutorial, we have published a comprehensive paper that introduces the architecture of ImputeGAP and compares it against other imputation libraries [12]. Additionally, we have developed multiple cutting-edge imputation tools and open-sourced their code. Those tools have been presented in top-tier venues, including ADARTS [7] at ICDE 2025, ImputeVIS [9] at PVLDB 2024, ORBITS [6] at PVLDB 2021, ImputeBench [8] at PVLDB 2020, and RecovDB [1] at ICDE 2019.

4 Societal Impact

This tutorial is positioned to significantly advance the field of time series data cleaning by providing practitioners and researchers with the necessary methodologies and tools to implement cutting-edge techniques across various domains. We anticipate that this tutorial will drive innovation within both academic and industrial sectors. Leveraging ImputeGAP's robust capabilities, researchers will be equipped to design novel imputation algorithms, systematically evaluate them across diverse datasets, analyze their performance characteristics, and assess their influence on downstream analytical and predictive tasks.

5 Tutors' Biography

Quentin Nater is a PhD student jointly supervised by Mourad Khayati and Philippe Cudré-Mauroux at the Department of Computer Science of the University of Fribourg in Switzerland. His main research interests revolve around time series analytics, with a focus on data imputation and multimodal learning.

Mourad Khayati is a Senior Researcher and Lecturer at the Department of Computer Science of the University of Fribourg in Switzerland. His research interests include time series analytics and data quality, with a special focus on temporal data cleaning. He is the recipient of the VLDB 2020 Best Experiments and Analysis Paper Award for this time series imputation benchmark [8].

Philippe Cudré-Mauroux is a full professor at the Department of Computer Science of the University of Fribourg and has several significant contributions to the database and data engineering field, including the first paper that implements a storage system for trajectory data [3].

References

[1] Ines Arous, Mourad Khayati, Philippe Cudré-Mauroux, Ying Zhang, Martin L. Kersten, and Svetlin Stalinlov. 2019. RecovDB: Accurate and Efficient Missing Blocks Recovery for Large Time Series. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*. IEEE, 1976–1979. doi:10.1109/ICDE.2019.00218

[2] José Cambronero, John K. Feser, Micah J. Smith, and Samuel Madden. 2017. Query Optimization for Dynamic Imputation. *Proc. VLDB Endow.* 10, 11 (aug 2017), 1310–1321. doi:10.14778/3137628.3137641

[3] Philippe Cudré-Mauroux, Eugene Wu, and Samuel Madden. 2010. TrajStore: An adaptive storage system for very large trajectory data sets. In *Proceedings of the 26th International Conference on Data Engineering, ICDE 2010, March 1-6, 2010, Long Beach, California, USA*. IEEE Computer Society, 109–120. doi:10.1109/ICDE.2010.5447829

[4] Wenjie Du. 2023. PyPOTS: A Python Toolbox for Data Mining on Partially-Observed Time Series. *CoRR* abs/2305.18811 (2023). doi:10.48550/ARXIV.2305.18811 arXiv:2305.18811

[5] James Honaker, Gary King, and Matthew Blackwell. 2011. Amelia II: A Program for Missing Data. *Journal of Statistical Software* 45, 7 (2011), 1–47. doi:10.18637/jss.v045.i07

[6] Mourad Khayati, Ines Arous, Zakhar Tymchenko, and Philippe Cudré-Mauroux. 2020. ORBITS: Online Recovery of Missing Values in Multiple Time Series Streams. *Proc. VLDB Endow.* 14, 3 (2020), 294–306. doi:10.5555/3430915.3442429

[7] Mourad Khayati, Guillaume Chacun, Zakhar Tymchenko, and Philippe Cudré-Mauroux. 2025. A-DARTS: Stable Model Selection for Data Repair in Time Series. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*. IEEE Computer Society, Los Alamitos, CA, USA, 2009–2023. doi:10.1109/ICDE65448.2025.00153

[8] Mourad Khayati, Alberto Lerner, Zakhar Tymchenko, and Philippe Cudré-Mauroux. 2020. Mind the Gap: An Experimental Evaluation of Imputation of Missing Values Techniques in Time Series. *Proc. VLDB Endow.* 13, 5 (2020), 768–782. doi:10.14778/3377369.3377383

[9] Mourad Khayati, Quentin Nater, and Jacques Pasquier. 2024. ImputeVIS: An Interactive Evaluator to Benchmark Imputation Techniques for Time Series Data. *Proc. VLDB Endow.* 17, 12 (2024), 4329–4332. doi:10.14778/3685800.3685867

[10] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A Research Platform for Distributed Model Selection and Training. *arXiv preprint arXiv:1807.05118* (2018).

[11] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[12] Quentin Nater, Mourad Khayati, and Jacques Pasquier. 2025. ImputeGAP: A Comprehensive Library for Time Series Imputation. (2025). arXiv:2503.15250 [cs.LG] https://arxiv.org/abs/2503.15250

[13] Mayur Kishor Shende, Andrés E. Feijóo-Lorenzo, and Neeraj Dhanraj Bokde. 2022. cleanTS: Automated (AutoML) tool to clean univariate time series at microscales. *Neurocomputing* 500 (2022), 155–176. doi:10.1016/J.NEUCOM.2022.05.057

[14] Shaoxu Song and Aoqian Zhang. 2020. IoT Data Quality. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. ACM, 3517–3518. doi:10.1145/3340531.3412173