CSC8639 Interim Report: Explaining Time Series Downsampling

Author: 210431461 | Supervisor: Matthew Forshaw

Introduction

Decision-makers must trust that the data being considered sufficiently represents the situation they are deciding on. Trusting the data means trusting which data points are selected, how this data collected and stored, and the capability of data practitioners to understand the quality, insights and limitations of it. This data pipeline can obscure or loose important information making it difficult to determine and explain that the data reliably and truthfully reflects the situation in question. Today's increasing volume of data makes this even more difficult (Yanzhe An and Wang (2022)).

This problem is particularly pertinent to collections of observations obtained through repeated measurements over time (Statistics (2023)), known as time series data. "[W]idely generated by industry and research at an increasing speed" (Yanzhe An and Wang (2022)), voluminous time series data is putting unprecedented demand on resources (Schlossnagle, Sheehy, and McCubbin (2021), Atlam, Walters, and Wills (2018)). This is forcing data practitioners to utilise methods, such as aggregation, windowing, and downsampling, that reduce data volumes to align with cost or time limitations, storage capabilities, and sustainability ambitions (Steinarsson (2013), Yanzhe An and Wang (2022), Tank (2020)). These reduction methods involve discarding data, which could result in the further loss of important information for decision-makers, and reduce the representativeness of the data.

However, discarding this data is a vital part of making voluminous time series understandable for human observation (Steinarsson (2013)). Downsampling reduces "…the number of data points while preserving the overall shape of the time series" (Donckt et al. (2023)), allowing the human eye to observe only the most valuable data points. Line graphs are an effective and popular method for visualising this data (Yunhai Wang and Yu (2023)). Despite effectively conveying the overall shape of the time series data (Aigner et al. (2008)), they offer little insight into which downsampling approach and parameters best represent the original data. Better visualising the impact of downsampling time series data, is likely help data practitioners confidently select their downsampling approach and better explain the insights and limitations of downsampled data. In doing so, data practitioners can better support decision-makers to trust the data they are considering.

Aim and Objectives

The research outlined by this interim report aims to improve how data practitioners better understand and explain the impact of downsampling time series data. It is hoped that this research will support data practitioners to determine and communicate whether data being considered by decision-makers reliably and truthfully reflects the situation in question, and help increase decision-makers trust in data-led decision-making.

To better understand and explain the impact of downsampling voluminous time series data, the research addresses the following five objectives:

- Develop a baseline understanding current downsampling algorithms' impact on original data sets by using the R package `TS Impute` (Moritz and Bartiz-Beielstein (2017)() to compare visualisations of the original and compressed data.

- Conduct exploratory analysis to determine common properties of time series data, attempting to refine the 22 time series features identified by `catch22` (Lubba et al. (2022)) to identify the most useful features for comparing the impacts of downsampling algorithms on original time series data.
- Design comparative visualisations of the most useful features of time series data across different downsampling algorithms to help communicate their impacts on the original data.
- Survey existing metrics used to compare downsampled data representativeness to inform an evaluation method for this research.
- Conduct user research with data practitioners and decision-makers to understand how they engage with downsampled time series data and its trustworthiness.

The aim and objectives set out here are ambitious; it is likely that each objective could be an individual project and the author is a part-time student. Given this, this project will delivered iteratively; the research objectives will be continuously reviewed to successfully deliver the most impact in the available time.

Project Plan

This project is divided in to five activity themes (Milestones, Reading, Exploratory Data Analysis, Visualisation, User Research) and six phases (1-6) to deliver the research aim and objectives by 15 August 2023. This plan is visualised on the next page.

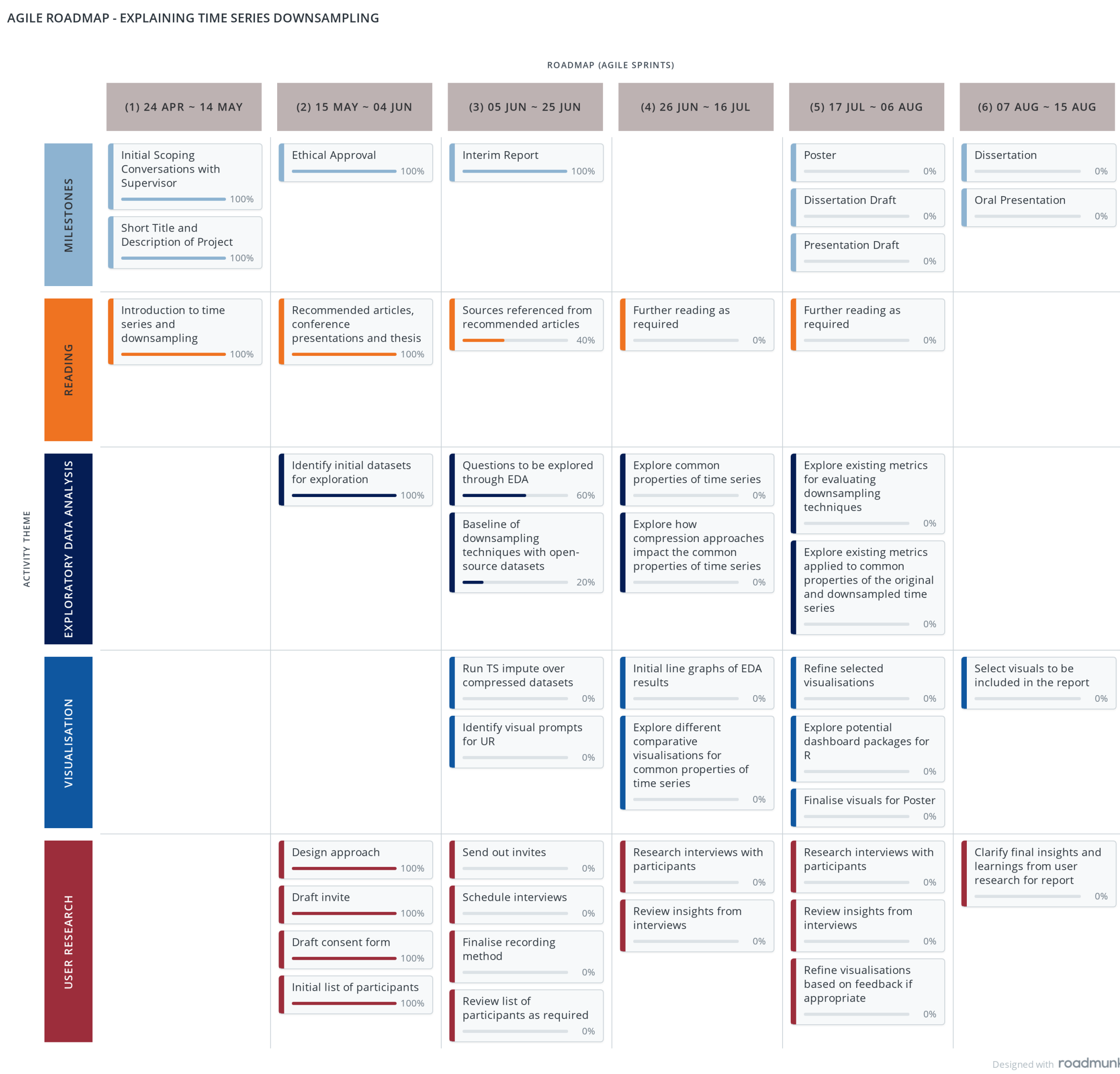
Overview of Progress

Progress on the project so far includes activity specified in phases (1) and (2) as well as some phase (3) activities, where progress to date is visualised as a percentage. Further details on this progress are set out below by activity theme:

- *Milestones:* Four meetings between the project supervisor and author have taken place, where the project scope, approach, aims and objectives have been clarified. A short title, description, and ethical approval have been submitted as required. The Data Management Plan is completed and shared in Annex A.
- *Reading:* Initial exploratory reading around time series data and downsampling was conducted before the supervisor recommended nine sources. These sources were read, and further reading of referenced sources is underway.
- *Exploratory Data Analysis:* The data sets for exploration were identified with support from the supervisor; the author has drafted questions to guide exploration and initial visual exploration is being conducted on selected data sets.
- *Visualisation:* Potential visuals for user research are being collated, but this activity theme is not a focus of phase (1) and (2).
- *User Research:* The approach to user research has been designed and discussed with the supervisor; invites, consent form,question list, and an initial list of participants are drafted. The invite, consent form, and question list are attached to this interim report (Annex B, C, and D).

Overview of Project

The visualisation of the project plan highlights the key activities within each theme across the project phases. Because of this iterative approach, it was agreed with the project supervisor that an agile approach was appropriate; the project plan is visualised as an agile roadmap. This visualisation, created on a platform provided by `roadmunk`, is interactive and will be updated to reflect the iterative nature of the project.



Project Risks and Mitigation

To be delivered successfully, there are several risks that this project may need to mitigate. These are set out in the table below with a risk rating of low, medium or high.

Rating	Risk	Mitigation
Low	Fewer people agreed to participate in User Research than expected.	Invites will be sent in a phased approach to enable further invites to be sent if response numbers are lower than expected.
Low	Number of decision-makers and data practitioners who agreed to participate in User Research is imbalanced.	The numbers for User Research are unlikely to be statistically significant for this project anyway, so any further limitations on the findings of User Research and their impact will be set out in the final report.

Rating	Risk	Mitigation
Medium	There are several elements of this project that are, to some extent, dependent on other project elements. For example, the exploratory data analysis and visualisation.	These dependencies are being mitigated by the agile approach to the project, allowing the author to adapt as needed, and clear communication with the supervisor.
High	The author is a part-time student working towards the same deadlines and criteria as full-time students.	The agile roadmap clearly sets out how the project is likely to progress, and the research objectives have been selected to help maximise impact if there are delays with some components. Progress will be clearly communicated to the supervisor, and an extension may be applied for if required.
Low	Data availability and cleaning take a significant amount of time, and may detract from original research.	A subs-selection of open source data previous time series visualisation research has been chosen in mitigation.
Low	The outputs from User Research, such as interview content or personal details, are not stored and treated securely.	Consent forms will be collected prior to scheduling interviews and participants will be annonymised in a locked spreadsheet saved separated from the project. There will be no personal details associated with the interview content, which will also be saved in a locked folder separate from the project.

This list of risks and mitigations will be reviewed and updated throughout the project as more may arise when the research is iterated.

Conclusion

The research of this project aims to improve how data practitioners better understand and explain the impact of downsampling time series data. This report has introduced the research topic by outlining why this matters, set out the research aim and objectives as well as progress so far, visualised the project plan and explained the risks that may need mitigated. The information provided in the report will be continuously reviewed in consultation with the project supervisor so that the research is delivers the most impact in the time available.

Annex A: Data Management Plan

A data management plan (DMP) is a written document outlining how you are planning to manage your research data both during and after your research project. The plan should address what types of data will be collected and how the data will be documented, stored, shared and preserved.

Annex B: Draft Invite for User Research

Annex C: Draft Question List for User Research

Annex D: Draft Consent Form for User Research

Annex E: References

Aigner, Wolfgang, Silvia Miksch, Wolfgang Müller, Heidrun Schumann, and Christian Tominski. 2008. "Visual Methods for Analyzing Time-Oriented Data." *IEEE Transactions on Visualization and Computer Graphics* 14 (1): 47–60. https://doi.org/10.1109/TVCG.2007.70415.

Atlam, Hany Fathy, Robert Walters, and Gary Wills. 2018. "Internet of Things: State-of-the-Art, Challenges, Applications, and Open Issues" 9 (3): 928–38. http://dx.doi.org/10.20533/ijicr.2042.4655.2018.0112.

Donckt, Jeroen Van Der, Jonas Van Der Donckt, Michael Rademaker, and Sofie Van Hoecke. 2023. "MinMaxLTTB: Leveraging MinMax-Preselection to Scale LTTB." https://arxiv.org/abs/2305.00332.

Lubba, Carl, Ben Fulcher, Trent Henderspn, Brendan Harris, Oliver TL, and Oliver Cliff. 2022. "Catch22: CAnonical Time-Series CHaracteristics." R Journal. https://doi.org/10.5281/zenodo.6673597.

Moritz, Steffen, and Thomas Bartiz-Beielstein. 2017. "imputeTS: Time Series Missing Value Imputation in r." R Journal. https://doi.org/10.32614/RJ-2017-009.

Schlossnagle, Theo, Justin Sheehy, and Chris McCubbin. 2021. "Always-on Time-Series Database: Keeping up Where There's No Way to Catch Up." *Commun. ACM* 64 (7): 50–56. https://doi.org/10.1145/3442518.

Statistics, Australian Bureau of. 2023. "TIme Series Analysis: The Basics." https://www.abs.gov.au/websitedbs/d3310114.nsf/home/time+series+analysis:+the+basics.

Steinarsson, Sveinn. 2013. "Downsampling Time Series for Visual Representation." Faculty of Industrial Engineering, Mechanical Engineering; Computer Science, School of Engineering; Natural Sciences, University of Iceland, Reykjavik, Iceland: University of Iceland.

Tank, The Shift Project: The Carbon Transition Think. 2020. "Implementing Digital Sufficiency." https://theshiftproject.org/wp-content/uploads/2021/07/TSP_DigitalSufficiency2020_Summary_corrige.pdf.

Yanzhe An, Yuqing Zhu, Yue Su, and Jianmin Wang. 2022. "TVStore: Automatically Bounding Time Series Storage via Time-Varying Compression." In *Proceedings of the 20th USENIX Conference on File and Storage Technologies*, 83–99. USENIX Conference on File and STorage Technologies. Santa Clara, CA, USA: USENIX Association.

Yunhai Wang, Xin Chen, Yuchun Wang, and Xiaohui Yu. 2023. "OM3: An Ordered Multi-Level Min-Max Representation for Interactive Progressive Visualization of Time Series." In *Proc. ACM Manag. Data*, 1:145:1–24. 2. ACM. https://doi.org/10.1145/3589290.