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 What data is selected, how the chosen data is collected and stored, and the capability of data practitioners to understand the quality, insights and limitations of this data. 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 [reference].

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 (look up 6, 79 in Yanzhe An and Wang (2022)). This is forcing data practitioners to select 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 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 (no 2 for Donckt et al. (2023)), 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 voluminous 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 should address the following objectives:

- Develop a baseline understanding current downsampling algorithms' impact on original data sets by using the R package `TS Impute` [Reference] 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* [reference] 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.

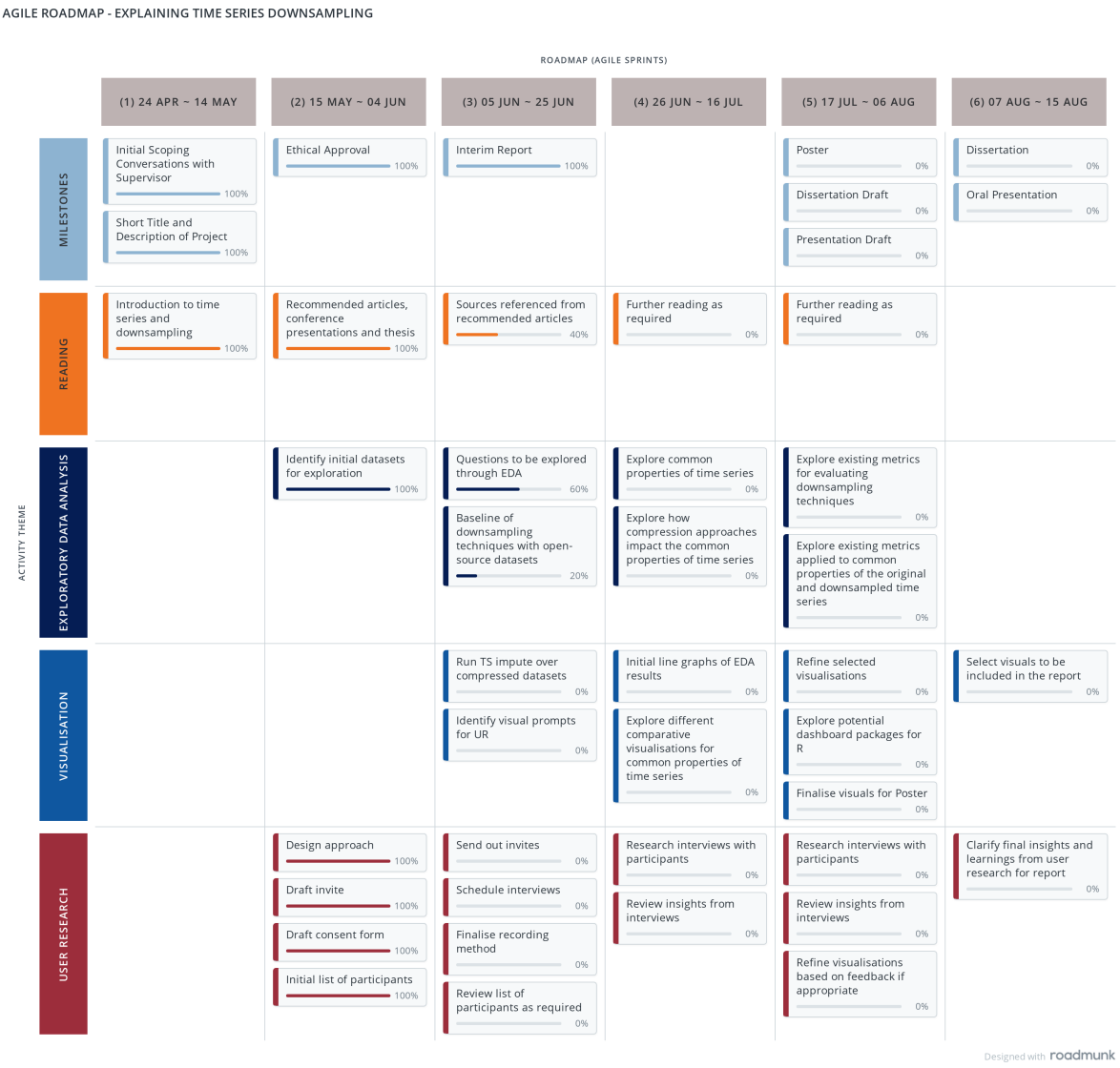
Overview of Progress

Progress of the project 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.
- *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 a 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 forms and an initial list of participants are drafted and are being reviewed by the supervisor as required.

Overview of Project

The visualisation of the project plan highlights the key activities within each theme across each phase. Because of this iterative approach, it was agreed with the project supervisor that an agile approach was appropriate, which 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.



Risks:

- Not enough UR (balance practitioners decision-makers / send invite to as many as possible)
- UR interest, but not scheduled
- Too many properties to create clear common principles for comparison. (iterate score and approach as needed)
- data availability (use open source, cleaned data from turing).
- time (part time - scope)
- finding common properties that are sufficiently clear for visuals- take too long / disrupts other project elements (use pre-exiting visuals as prompts in UR?)

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.

References

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.

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.