Explaining time series downsampling through visualisation

A Preprint

Morgan Frodsham
School of Computing
Newcastle University
Newcastle upon Tyne, UK
`M.C.M.Frodsham2@newcastle.ac.uk`

Matthew Forshaw
School of Computing
Newcastle University
Newcastle upon Tyne, UK
`matthew.forshaw@newcastle.ac.uk`

July 29, 2023

Abstract

Enter the text of your abstract here.

Keywords blah · blee · bloo · these are optional and can be removed

1 INTRODUCTION

The UK Government is committed to making data-driven decisions that engender public trust [1]–[4]. Data-driven decisions are considered to be "more well-informed" [1], effective [4], consistent [3], and better "at scale" [2]. Despite this, there is a lack of trust in government use of data [5]. This suggests that public trust in data-driven decisions goes beyond how the "data complies with legal, regulatory and ethical obligations" [3]. Transparency is needed for the UK public to have "confidence and trust in how data, including personal data, is used" [2], [5].

To make data-driven decisions, government decision-makers also need to trust how the data used (cite user research here). This 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. At every stage of the data processing pipeline, data practitioners have the opportunity to communicate the impact of the assumptions and choices they are making to support decision-makers in trusting the data informing their decisions.

Time series data is used across the UK Government [6] to inform decision-makers across various domains [7]. It is also widely generated and used by industry and research [8]. The volume of time series data is continuously increasingly [9], posing significant challenges for handling and visualising this popular data type [8]. Data practitioners must utilise methods that reduce data volumes to align with limitations like processing time, computing costs, storage capabilities, and sustainability ambitions [8], [10], [11].

Downsampling is an established technique [12], [13] that involves selecting a representative subset of the time series data to preserve its shape while reducing the number of data points [9], [14]. This is a vital part of making voluminous time series understandable for human observation [10] and an essential step in many time series database solutions [9]. However, little attention has been devoted to how downsampling impacts decision-makers trust in the data.

Despite widespread use, how to communicate the impact of downsampling algorithms on time series data remains understudied [9], [10]. Downsampling expands the boundaries of risk for decision-makers as data practitioners may not realise the significance of the data being discarded. Such choices throughout the data pipeline may have disproportionately larger consequences later as their ramifications for future decisions are not fully understood by all. It is important, therefore, that data practitioners are able to communicate the impact of choices made throughout the data pipeline.

To address these challenges, this work proposes a visualisation methodology for understanding and communicating the impact of downsampling algorithms on time series data. Section *II* contextualises the impact of this work for data practitioners and decision-makers by sharing insights from user research. Section *III* provides an overview of previous related work to help assess the contributions of this work. Section *IV* presents how R packages `imputeTS` [15] and `Rcatch22` **Rcatch22?** are combined to identify the time series features that are most sensitive to downsampling. Section *V* outlines how this approach allows data practitioners to communicate which downsampling algorithms and parameters are most appropriate for particular use cases. Section *VI* shares the potential impact of this work and the opportunities for further work to improve decision-makers' trust in data.

2 MOTIVATION

3 RELATED WORK

A. Downsampling applications and processes

Data-driven decision-making necessitates that time series data is kept for future analysis. Technological innovation has generated unprecedented amount of time series data, which continues to grow [2], [16], [17], **TVstore?**. For example, climate simulations that inform recommendations for decision-makers generate tens of terabytes per second. Downsampling plays an important role in addressing how this voluminous data is processed, stored **TVstore?** and visualised [10], [18].

Data practitioners have made recent advances in the performance of value preserving downsampling algorithms [12], [14], [18]–[20], **samping?**. Examples of these advances are set out in the table below:

insert table [9] - EveryNth, also known as sampling or decimation, selects n^{th} datapoint [19] - percentage change - MinMax preserves the minimum and maximum of every data bucket [14] - OM^3 maintains minimum and maximum values at every time interval that is used to rasterize a pixel column in the display window [20] - M4 combines EveryNth and MinMax, selecting the first and last values of each data bucket as well as its minimum and maximum [18], [21] - Largest-TriangleOne-Bucket (LTOB) - Largest Triangle Three Buckets LTTB selects the data point that forms the largest triangular surface between the previously selected data point and the next data bucket's average value [14] - MinMaxLTTB preselects data using MinMax before applying LTTB on the selected datapoints **MinLaxLTTB?**

B. Time series visualisation

C. Trust in Data

You can use directly LaTeX command or Markdown text.

LaTeX command can be used to reference other section. See Section 8. However, you can also use **bookdown** extensions mechanism for this.

3.1 Headings: second level

You can use equation in blocks

$$\xi_{ij}(t) = P(x_t = i, x_{t+1} = j | y, v, w; \theta) = \frac{\alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}$$

But also inline i.e $z = x + y$

3.1.1 Headings: third level

Another paragraph.

4 METHODOLOGY

4.1 ImputeTS

4.2 Rcatch22

4.3 Downsamplng Impat

4.4 User Research

5 RESULTS AND EVALUATION

6 FUTURE WORK

7 CONCLUSION

8 REFERENCES

9 Examples of citations, figures, tables, references

You can insert references. Here is some text **kour2014real?**, **kour2014fast?** and see **hadash2018estimate?**.

The documentation for `natbib` may be found at

You can use custom blocks with LaTeX support from **rmarkdown** to create environment.

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf%7D`

Of note is the command `\citet`, which produces citations appropriate for use in inline text.

You can insert LaTeX environment directly too.

```
\citet{hasselmo} investigated\dots
```

produces

> Hasselmo, et al. (1995) investigated…

`https://www.ctan.org/pkg/booktabs`

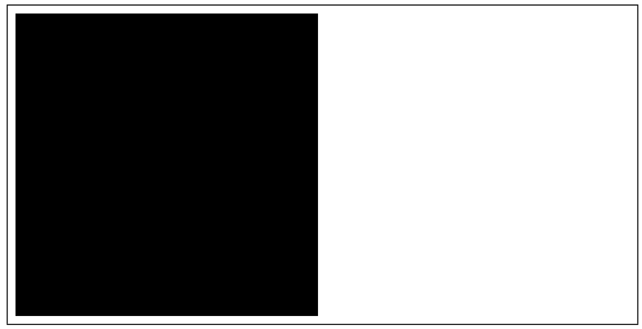

Figure 1: Sample figure caption.

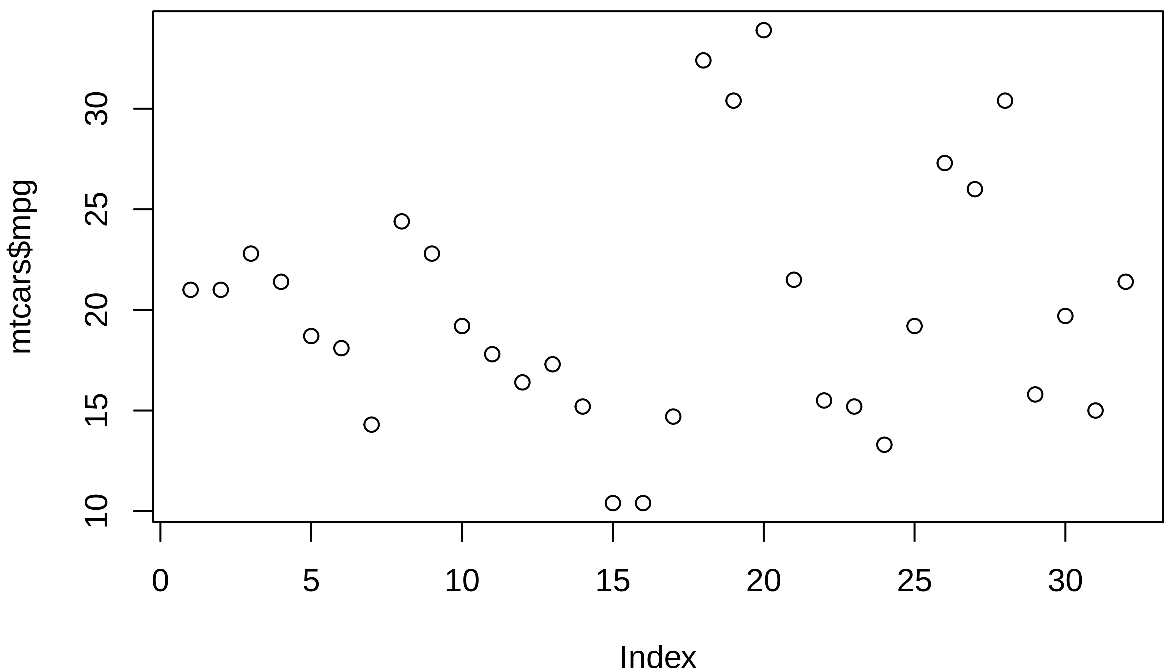

Figure 2: Another sample figure

9.1 Figures

You can insert figure using LaTeX directly.

See Figure 1. Here is how you add footnotes. [^Sample of the first footnote.]

But you can also do that using R.

```
plot(mtcars$mpg)
```

You can use **bookdown** to allow references for Tables and Figures.

Table 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	~100
Axon	Output terminal	~10
Soma	Cell body	up to 10^6

9.2 Tables

Below we can see how to use tables.

See awesome Table~1 which is written directly in LaTeX in source Rmd file.

You can also use R code for that.

```
knitr::kable(head(mtcars), caption = "Head of mtcars table")
```

Table 2: Head of mtcars table

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

9.3 Lists

- Item 1
- Item 2
- Item 3

[1] Cabinet Office and Government Digital Service, "Government transformation strategy: Better use of data." HM Government; https://www.gov.uk/government/publications/government-transformation-strategy-2017-to-2020/government-transformation-strategy-better-use-of-data, 2017.

[2] Department for Digital, Culture, Media & Sport and Department for Science, Innovation & Technology, "National data strategy." HM Government; https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy, 2020.

[3] M. of Defence, "Data strategy for defence," *GOV.UK*. HM Government; https://www.gov.uk/government/publications/data-strategy-for-defence/data-strategy-for-defence, 2021.

[4] Central Digital & Data Office, "Transforming for a digital future: 2022 to 2025 roadmap for digital and data." HM Government; https://www.gov.uk/government/publications/roadmap-for-digital-and-data-2022-to-2025/transforming-for-a-digital-future-2022-to-2025-roadmap-for-digital-and-data, 2022.

[5] Centre for Data Ethics & Innovation, "Addressing trust in public sector data use." https://www.gov.uk/government/publications/cdei-publishes-its-first-report-on-public-sector-data-sharing/addressing-trust-in-public-sector-data-use#introduction--context.

[6] Government Analysis Function, "Types of data in government learning pathway." https://analysisfunction.civilservice.gov.uk/learning-development/learning-pathways/types-of-data-in-government-learning-pathway/, 2022.

[7] Office for National Statistics, "Time series explorer." `https://www.ons.gov.uk/timeseriestool?query=&topic=&updated=&fromDateDay=&fromDateMonth=&fromDateYear=&toDateDay=&toDateMonth=&toDateYear=&size=50`, Unknown.

[8] Y. An, Y. Su, Y. Zhu, and J. Wang, "TVStore: Automatically bounding time series storage via time-varying compression," in *Proceedings of the 20th USENIX conference on file and storage technologies*, in USENIX conference on file and STorage technologies. Santa Clara, CA, USA: USENIX Association, 2022, pp. 83–99.

[9] J. Donckt, J. Donckt, M. Rademaker, and S. Hoecke, "Data point selection for line chart visualization: Methodological assessment and evidence-based guidelines." 2023. doi: 10.48550/arXiv.2304.00900.

[10] S. Steinarsson, "Downsampling time series for visual representation." University of Iceland, Faculty of Industrial Engineering, Mechanical Engineering; Computer Science, School of Engineering; Natural Sciences, University of Iceland, Reykjavik, Iceland, 2013.

[11] The Shift Project, "Implementing digital sufficiency," 2020.

[12] W. Aigner, S. Miksch, W. Muller, H. Schumann, and C. Tominski, "Visual methods for analyzing time-oriented data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 1, pp. 47–60, 2008, doi: 10.1109/TVCG.2007.70415.

[13] B. C. Kwon, J. Verma, P. J. Haas, and C. Demiralp, "Sampling for scalable visual analytics," *IEEE Computer Graphics and Applications*, vol. 37, no. 1, pp. 100–108, 2017, doi: 10.1109/MCG.2017.6.

[14] J. Donckt, J. Donckt, M. Rademaker, and S. Hoecke, "MinMaxLTTB: Leveraging MinMax-preselection to scale LTTB." 2023. Available: `https://arxiv.org/abs/2305.00332`

[15] S. Moritz and T. Bartiz-Beielstein, "imputeTS: Time series missing value imputation in r," vol. 9.1. R Journal, 2017. doi: 10.32614/RJ-2017-009.

[16] A. Visheratin *et al.*, "Peregreen – modular database for efficient storage of historical time series in cloud environments," in *2020 USENIX annual technical conference (USENIX ATC 20)*, USENIX Association, 2020, pp. 589–601. Available: `https://www.usenix.org/conference/atc20/presentation/visheratin`

[17] T. Schlossnagle, J. Sheehy, and C. McCubbin, "Always-on time-series database: Keeping up where there's no way to catch up," *Commun. ACM*, vol. 64, no. 7, pp. 50–56, 2021, Available: `https://doi.org/10.1145/3442518`

[18] A. Kohn, D. Moritz, and T. Neumann, "DashQL – complete analysis workflows with SQL." 2023. doi: 10.48550/arXiv.2306.03714.

[19] U. Jugel, Z. Jerzak, G. Hackenbroic, and V. Markl, "VDDA: Automatic visualization-driven data aggregation in relational databases," *The VLDB Journal*, vol. 25, 2016, doi: 10.1007/s00778-015-0396-z.

[20] W. Yunhai *et al.*, "OM3: An ordered multi-level min-max representation for interactive progressive visualization of time series," in *Proc. ACM manag. data*, ACM, 2023. Available: `https://doi.org/10.1145/3589290`

[21] U. Jugel, Z. Jerzak, G. Hackenbroich, and V. Markl, "M4: A visualization-oriented time series data aggregation. Proceedings of the VLDB endowment," vol. 7, 2014, Available: `https://www.vldb.org/2014/program/http://www.vldb.org/pvldb/vol7/p797-jugel.pdf`