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 23, 2023

Abstract

Enter the text of your abstract here.

***K*eywords** 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 the data and how it is used. 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 for decision-makers across various domains [7]. The volume of time series data has been increasingly continuously [8], posing significant challenges for handing and visualising this popular data type [9]. Data practitioners must utilise methods that reduce data volumes to align with limitations like processing time, computing costs, storage capabilities, and sustainability ambitions [9]–[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 [8], [14]. This is an essential step in many time series database solutions [8] and a vital part of making voluminous time series understandable for human observation [10].

Despite widespread use, the how to communicate the impact of downsampling algorithms on time series data remains understudied [8], [10]. Downsampling expands the boundaries 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.

.

2 RELATED WORK

This section provides a comprehensive overview of related work in the field. (Summarise section) By doing so, we aim to offer a clear understanding of the current state-of-the-art and identify the gaps that our work seeks to address.

Trust

Although easy to grasp intuitively, transparency is hard to define and even harder to realize.@digital_transparency - Merely opening data does not result in digital transparency and might only result in information overload for those wanting to examine such data.

transparency" initiatives become part of an obfuscation process that often uses the rhetoric of placation and diversion [15]

Ananny and Crawford argue transparency alone can not create accountable systems as simply looking is insufficient. [16] Mike Ananny and Kate Crawford (2018) reinforce this point, noting that "the implicit assumption behind calls for transparency is that seeing a phenomenon creates opportunities and obligations to make it accountable and thus to change it" (Ananny and Crawford, 2018: 974, emphasis in original). They describe the promise of transparency as being rooted in the connection between seeing, knowing, and controlling.

Socially meaningful transparency moves away from meaninfulness in relation to individuals' specific needs to o focus attention on societal needs in terms of what is made transparent, for whom, how, when and in what ways, and, crucially, who decides. [17]

told that access to information is essential, but without the tools for turning that access to agency - transparency fallacy - archieving meaningful transparency is difficult [18]

The political valence of data transparency is a critical reminder of the inherently sociopolitical nature of all technologies, including institutional data practices. [19]

- time series volume

[9] - "...mounting demands have emerged for keeping time series data for future analysis [94]. But time series data are generated at a growing speed that is outpacing the increase of computing capabilities [17, 79]. Many application scenarios cannot afford enough computing resources such as storage and network bandwidth to accommodate the processing needs for time series data." pg 84

- time series visualisation
- characteristics of time series
- downsampling
- trust
- masters thesis
- imputeTS

In statistics this process of replacing missing values is called imputation.

At the moment imputeTS (Moritz, 2016a) is the only package on CRAN that is solely dedicated to univariate time series imputation and includes multiple algorithms. Nevertheless, there are some other packages that include imputation functions as addition to their core package functionality. Most noteworthy being zoo (Zeileis and Grothendieck, 2005) and forecast (Hyndman, 2016). Both packages offer also some advanced time series imputation functions. The packages spacetime (Pebesma, 2012), timeSeries (Rmetrics Core Team et al., 2015) and xts (Ryan and Ulrich, 2014) should also be mentioned, since they contain some very simple but quick time series imputation methods.

Univariate means there is just one attribute that is observed over time. Which leads to a sequence of single observations o1 , o2, o3, ... on at successive points t1 , t2, t3, ... tn in time

- Rcatch22

Selecting an appropriate feature-based representation of time series for a given application can be achieved through systematic comparison across a comprehensive time-series feature library, such as those in the hctsa toolbox. However, this approach is computationally expensive and involves evaluating many similar features, limiting the widespread adoption of feature-based representations of time series for real-world applications. In this work, we introduce a method to infer small sets of time-series features that (i) exhibit strong classification performance across a given collection of time-series problems, and (ii) are minimally redundant. Applying our method to a set of 93 time-series classification datasets (containing over 147,000 time series) and using a filtered version of the hctsa feature library (4791 features), we introduce a set of 22 CAnonical Time-series CHaracteristics, catch22, tailored to the dynamics typically encountered in time-series data-mining tasks.

This dimensionality reduction, from 4791 to 22, is associated with an approximately 1000- fold reduction in computation time and near linear scaling with time-series length, despite an average reduction in classification accuracy of just 7%

An ideal starting point for such an exercise is the comprehensive library of over 7500 features provided in the hctsa toolbox (Fulcher et al. 2013; Fulcher and Jones 2017).

- visualisation of time series
- turing change point / annotated change

You can use directly LaTeX command or Markdown text.

LaTeX command can be used to reference other section. See Section 7. However, you can also use **bookdown** extensions mechanism for this.

2.1 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 $z = x + y$

2.1.1 Headings: third level

Another paragraph.

3 METHODOLOGY

3.1 ImputeTS

3.2 Rcatch22

3.3 Downsamplng Impat

3.4 User Research

4 RESULTS AND EVALUATION

5 FUTURE WORK

6 CONCLUSION

7 REFERENCES

8 Examples of citations, figures, tables, references

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

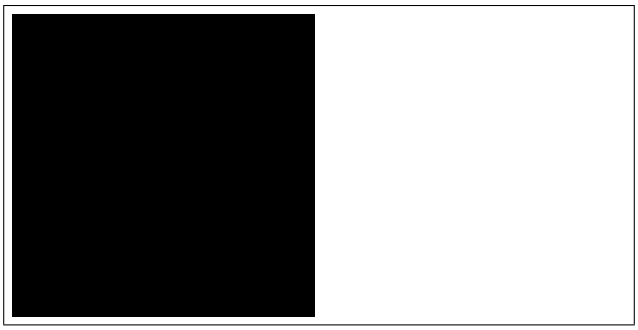

Figure 1: Sample figure caption.

Table 1: Sample table title

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

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`

8.1 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.

8.2 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.

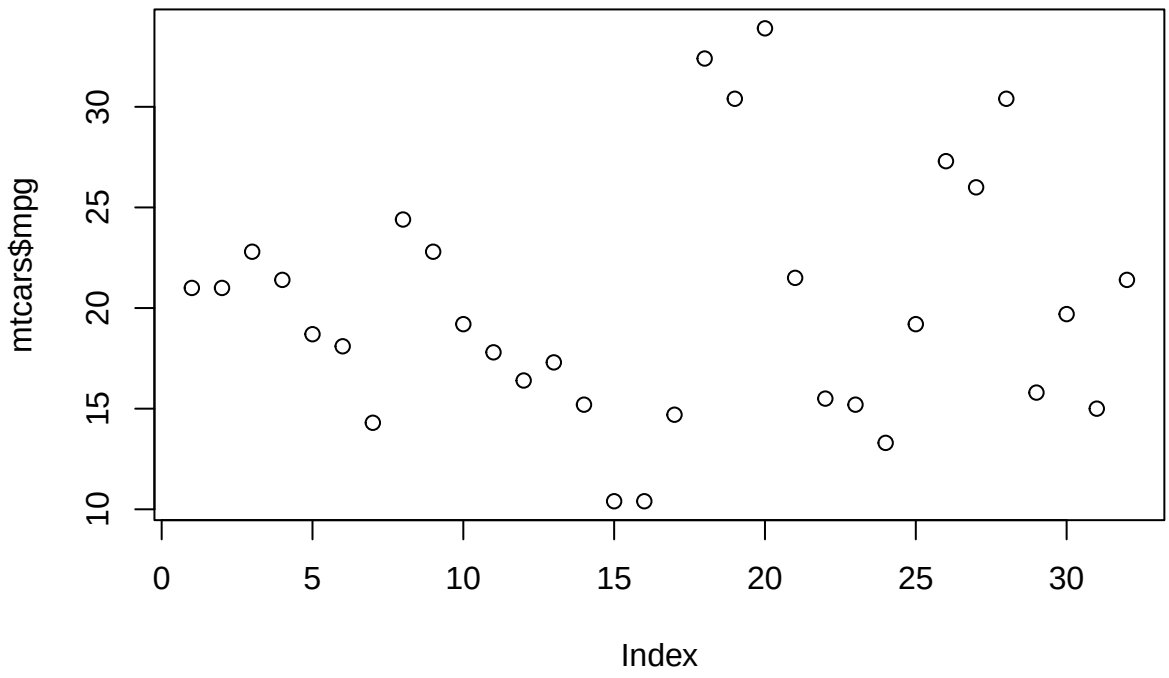

Figure 2: Another sample figure

```
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

8.3 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] 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.

[9] 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.

[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] N. A. Draper and J. Turow, "The corporate cultivation of digital resignation," *New Media & Society*, vol. 21, no. 8, pp. 1824–1839, 2019, doi: 10.1177/1461444819833331.

[16] M. Ananny and K. Crawford, "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability," *New Media & Society*, vol. 20, no. 3, pp. 973–989, 2018, doi: 10.1177/1461444816676645.

[17] J. Bates, H. Kennedy, I. Medina Perea, S. Oman, and L. Pinney, "Socially meaningful transparency in data-based systems: Reflections and proposals from practice," *Journal of Documentation*, vol. ahead–of–print, 2023, doi: 10.1108/JD-01-2023-0006.

[18] J. A. Obar, "Sunlight alone is not a disinfectant: Consent and the futility of opening big data black boxes (without assistance)," *Big Data & Society*, vol. 7, no. 1, 2020, doi: 10.1177/2053951720935615.

[19] K. E. Levy and D. M. Johns, "When open data is a trojan horse: The weaponization of transparency in science and governance," *Big Data & Society*, vol. 3, no. 1, 2016, doi: 10.1177/2053951715621568.