THE EXPLAINABILITY OF TIME SERIES DOWNSAMPLING

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

August 14, 2023

Abstract

Trusting data-driven decision-making goes beyond demonstrating compliance with legal, regulatory and ethical obligations; decision-makers need to trust how the data is used. Handling, storing and visualising the volume of data being generated today requires data practitioners to make assumptions and processing choices that remain opaque to decision-makers. The research outlined by this paper explores decision-makers' trust in data, data pracitioners' experience of communicating data insights to decision-makers and a new visualisation methodology for explaining the impact of downsampling on high-volume time series data. It uniquely combines user research with R packages `imputeTS` [1] and `Rcatch22` [2] to identify and visualise time series features that are most sensitive to downsampling. It is hoped this will improve decision-makers' trust in data by helping data practitioners to create transparency in the data processing pipeline, communicate the impact of downsampling, and support conversations about which algorithms or parameters are most appropriate for particular decision-maker use cases.

***K*eywords** time series · downsampling · decision-making · visualisation

1 INTRODUCTION

HM Government is committed to making data-driven decisions that engender public trust [3]–[6]. Data-driven decisions are considered to be "more well-informed" [3], effective [6], consistent [5], and better "at scale" [4]. Despite this, there is a lack of trust in government use of data [7]. This suggests that public trust in data-driven decisions goes beyond how the "data complies with legal, regulatory and ethical obligations" [5]. The UK public need to have "confidence and trust in how data, including personal data, is used" [4], and this requires transparency [7].

To make data-driven decisions, it is likely that government decision-makers also need transparency of how the data used to trust it. 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 data impact of the assumptions and choices they are making. This could not only support decision-makers in trusting the data informing their decisions, but also help decision-makers understand the limitations of the data and thresholds at which the data insights can be relied upon for each decision.

The research outlined in this paper explores this dynamic with a two pronged approach. User research with sixteen UK Civil Servants investigates what decision-makers need to trust the data insights they are presented with, and how data practitioners assess and communicate the trustworthiness of the data. These insights are combined with analysis of how to explain the impact of downsampling algorithms on time series data without detailing extensive statistical evidence.

Downsampling involves selecting a representative subset of data to preserve its original shape while reducing the number of data points [8], [9]. Time series data is used across HM Government [10] to inform decision-makers across various domains [11]; it is also widely generated and used by industry and research [12]. The volume of time series data is continuously increasingly [8], posing significant challenges for handling and visualising this popular data type [12].

Data practitioners must utilise methods that reduce data volumes to align with limitations like processing time, computing costs, storage capabilities, and sustainability ambitions [12]–[14]. Downsampling is an established technique [15], [16] for this, and is a vital part of making voluminous time series understandable for human observation [13] and an essential step in many time series database solutions [8]. However, little attention has been devoted to how downsampling impacts decision-makers trust in the data and how to communicate the impact of downsampling algorithms on time series data remains also understudied [8], [13].

Downsampling is an important part of how time series data is used and a technical issue that, this research posits, whose impact is important for decision-makers to understand. Downsampling expands the boundaries of risk for decision-makers as data practitioners may not realise the significance of the data being discarded. For example, a decision-maker may be looking at analysis of CO_2 levels on an open-plan office floor to decide on how many officials can work in the office each day post-Coivd. Different downsampling algorithms preserve different data points, which could change the threshold of risk for the decision-maker. In complex decisions, such Choices throughout the data pipeline may have disproportionately larger consequences later as their future ramifications are unlikely to be fully understood. It is important, therefore, that data practitioners are able to communicate the impact of choices made throughout the data pipeline.

This paper shares initial insights from user research on decision-makers' trust in data and suggests a visualisation methodology for communicating the impact of downsampling algorithms on time series. This methodology combines user research insights with R packages `imputeTS` [1] and `Rcatch22` [2] to identify time series features that are most sensitive to downsampling and visualise the impact of different downsampling algorithms on these features. It is hoped this will improve how data practitioners communicate the impact of downsampling, support conversations about which algorithms or parameters are most appropriate for particular decision-maker use cases, and increase decision-makers' trust in data by creating transparency in the data processing pipeline,.

2   RELATED WORK

This section provides an overview of previous related work to create a clear understanding of the most relevant fields of research and identify the gaps being addressed by the paper.

2.1 Data transparency

Technology transparency, "including institutional data practices," is sociopolitical in nature [17]. There is a growing number of researchers reflecting on "societal needs in terms of what is made transparent, for whom, how, when and in what ways, and, crucially, who decides" [18].

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" [19]. However, without sufficient agency to explore the information being shared, seeing a phenomenon often results in "information overload" [20] that obfuscates or diverts [21]. Without agency, transparency is increasingly considered to be a fallacy [22].

Meaningful transparency is only realised when the information is provided with the tools to turn "access to agency" [19], [22]. This suggests that data practitioners communicating the assumptions and choices made throughout the data processing pipeline with decision-makers is not likely to create trust in how the data is used. Instead, data practitioners should be encouraged to find tools, such as interactive visualisations [8], that put agency into the hands of decision-makers.

2.2 Time series visualisation

Time series data is commonly visualised as a line graph [13], [23]. Line graphs help the human eye to observe only the most important data points [13] by convey the overall shape and complexity of the time series data [8], [15]. The most effective time series visualisations are, however, interactive [23], [24], turning access into agency [22] by allowing the user to access details on demand. Evaluation of time series visualisation is, therefore, a growing field of research [23].

However, this growing field of research does not extend to visualisations of choices and assumptions made during data processing pipline. Indeed, such visualisations are a side effect of the research. This dynamic is exemplified by the R package `imputeTS` [1] where the impact of imputation choices made by the user is only visualised to support the user through the complete process of replacing missing values in time series [25]. The research set out in this paper harnesses the capabilities of `imputeTS` and its 'process' visualisations to help data practitioners communicate the impact of downsampling choices made in the data processing pipeline.

2.3 Value preserving data aggregation

Technological innovation has generated unprecedented amount of time series data and this data continues to grow [4], [12], [26], [27]. For example, tackling climate change is the UK Government's "number one international priority" [28], yet climate simulations that help inform decision-makers generate tens of terabytes per second [12], [29]. Value preserving data aggregation (a subset of downsampling) algorithms play an important role in addressing how this voluminous data is processed, stored [12] and visualised [13], [30] by minimising computing resources needed [12], reducing network latency, and improving rendering time [8], [9].

Value preserving data aggregation algorithms preserve the original data points that aim to be most representative of the original data. Examples of these algorithms are provided in the table below [8], [9], [13], [30]–[33]:

Table 1: Downsampling Algorithms

Algorithm	Description
EveryNth	Selects every nth data point.
Percentage Change	Selects every data point beyond a specified percentage of the previous data poin
Mode-Median Bucket	Selects mode or median within equally sized data buckets.
Min-Std-Error-Bucket	Selects data points with standard error of the estimate (SEE) in linear regressio
MinMax	Preserves minimum and maximum data points in each data bucket.
OM3	Maintains the minimum and maximum data points at every time interval for ras
M4	Selects the first and last as well as the minimum and maximum data points
Longest-Line Bucket	Calculates line length (euclidean distance) between data buckets.
Largest Triangle Three Bucket (LTTB)	Calculates the largest triangular surface between the previously selected data po
MinMaxLTTB	Preselects data points using MinMax before applying Largest Triangle Three Bu
EveryNth	Selects every nth data point.

Data practitioners have made recent advances in the performance and evaluation of downsampling approaches [8], [9], [15], [16], [24], [30]–[32]. These advances focus on the effectiveness of the algorithm in delivering downsampled data that represents the original data as accurately as possible. This is vital part of enabling and improving data-driven decision-making, but is focused on supporting data practitioners in their analysis of the data. Instead, the research set out in this paper aims to support data practitioners to communicate the impact of their downsampling choices for decision-makers.

2.4 Time series feature analysis

The increasing size of modern time series data sets has also generated new research into the dynamical characteristics of time series [34]. These characteristics are often used to identify features that enable efficient clustering and classification of time series data, especially for machine learning. A comprehensive library of such features is the *hctsa* (highly comparative time series analysis) toolbox. This shares the 4791 best performing features after computationally comparing thousands of features from across scientific time series analysis literature [35].

Utilising such a library, however, is computationally expensive [34]. C. H. Lubba et. al have attempted to address this by identified a subset of 22 features that are tailored to time series data mining tasks [34], [36]. Although further research is needed to evaluate the relative performance of different feature sets on different types of problems, `catch22` performs well against other feature libraries across 800 diverse real-world and model-generated time series [37].

Features used to classify time series data could provide a common framework by which to consistently compare different downsampling algorithms and parameters. The research set out in this paper utilises the `Rcatch22` subset of features to explore impact of downsampling and create a visual tool for explaining this impact.

3 USER RESEARCH

One-to-one virtual interviews with 16 UK Civil Servants (nine decision-makers and seven data practitioners) highlight the importance of transparency in enabling data-driven decision-making across government. This user research utilises the same interview script for all decision-makers and data practitioners, respectively. Follow-up questions vary to support the flow of each interview and unpack the participants' responses. The interviews are recorded and transcribed on MS teams for later analysis.

Qualitative thematic analysis combined with natural language processing keyword extraction is used to identify the the key points of each interview, which are similar for both decision-makers and data practitioners. Transparency of data provenance, collection, and analysis pipeline is desired by decision-makers to understand the reliability, repeatability, and quality of the data they're presented with. All the decision-makers and data-practitioners shared that this is best communicated through engaging conversations and interactive visualisations; they both emphasised the importance of communicating the story behind data, its assumptions, and how the data changes through time. Everyone interviewed acknowledged that this is difficult to achieve and how to do this better is a topic of live debate.
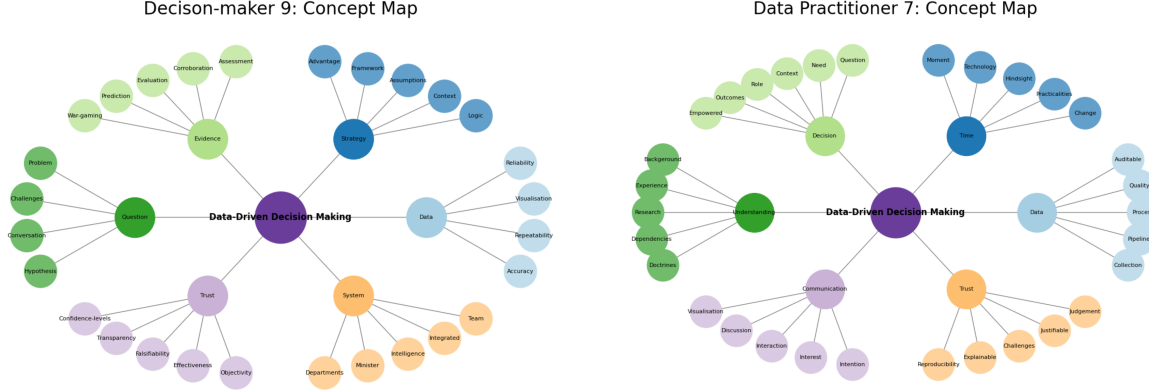
Insights from this user research indicate that there is a spectrum of transparency that decision-makers desire for decisions. This spectrum depends on the question being addressed by the data; it's complexity, importance and potential impact. To trust the data informing data-driven decisions, decision-makers and data practitioners repeatedly requested transparency of the data context, source and limitations as well as an assessment of overall confidence in the data.

Decision-makers shared difficulties in developing correct views of "how much they can trust the data" they're seeing. They addressed their desire for more emphasis on the "metadata and how to appropriately describe the level of uncertainty and provenance of the data" to help ensure the data is being applied to the context it was collected for. They shared the importance of understanding if "there is anything that is baked into what is being provided" in order to trust the data and, how the source of this trust changes as data practitioners become decision-makers because they rarely see the raw data and so "have to have an understanding of what preprocessing has been done, because it almost always has been [done]." Decision-makers shared that visualisations are important for conveying "specific insights that simply looking at numbers can't get over" and emphasised that if analysts "clearly explain what the limitations are and and and rigorously presented that, then I'll trust [the analysis] more."

3.1 [IS THERE SPACE FOR A DATA PRACTITIONER VERSION?]

-
-

Below concept maps outlining the key themes from anonymised decision-maker and data-practitioner interviews are shared as examples. All the interview concept maps are shared in Annex A. Only summative content from these interviews is presented in compliance with the consent forms shared with the user research participants.

The findings from this user research indicate that decision-makers need transparency of how the data is used to trust the insights informing data-driven decisions. This suggests that endeavors to create greater transparency in data pipelines, such as the research outlined in the subsequent sections of this paper, are worthy of further study. It also suggests that evaluating the objectivity and quality of data sources as well as mechanisms to develop common descriptions of data insight confidence-levels are potential avenues for data science to have greater policy impact.

It is important to note that, like many interview studies [38], the findings of this user research is limited; the sample of participants is not drawn from the overall population. The participants have been selected by reaching out to relevant individuals in the researcher's professional network, and this introduces bias. However, given the consistency of insights from the user research participants, further interviews with a wider, random subset of the population would be beneficial.

4 ANALYSIS METHODOLOGY

This section outlines the analysis design undertaken to explain the impact of downsampling algorithms on time series data without detailing extensive statistical evidence; the decisions, justifications and limitations of this methodology are shared.

4.1 Data

The eight Alan Turing Institute `Annotated Change` [39] synthetic time series are selected for initial analysis. There is an inherent limitation in using synthetic time series for analysis because the data has been generated for a different purpose. However, these time series have been choices because they are previously cleaned time series from a reputable institution with known change points that are designed to provide examples of different types of time series. These synthetic time series (demo_100, demo_200, demo_300, demo_400, demo_500, demo_600, demo_700, demo_800) are used for demonstration to support change point annotators annotate real-world data sets for the `Turing Change Point Dataset` [39]. The eighth `Annotated Change` synthetic time series (demo_800) is multivariate and split in two (800A and 800B) to create nine synthetic time series for this research and enable better comparison.

4.2 Downsampling

These nine synthetic time series are imported into `RStudio` from JSON scripts and two downsampling algorithms are applied to all with the `Jettison MVP Code` [40]. This enables initial insights into the impacts of downsampling with minimally complex code. The two algorithms used by the `Jettison MVP Code` [40]

are *EveryNth*, which selects every other data point, and *Percentage Change*, which selects every data point that has greater than one per cent difference between the last transmitted and newest values.

[SHOULD I ADD AN EQUATION HERE FOR THE ALGORITHMS?]

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

The `Jettison MVP Code` [40] applies these algorithms iteratively across each of the nine synthetic time series. This creates parameters 1 to 50 for each downsampling algorithm and `Annotated Change` time series. There are now 900 time series for investigation; 450 for each downsampling algorithm.

The R package `imputeTS` [1] is applied to all 900 time series to obtain the remaining data volume after downsampling, number of missing values (NAs), and number of gaps in the time series. This is important to understand the effect of each downsampling algorithm. Initial analysis highlights that the effect of each downsampling algorithm is unevenly distributed across the 50 parameters; for example, *EveryNth* discards data more quickly than *Percentage Change* so comparing the time series for each parameter is insufficient. The remaining data volume, therefore, provides a common variable that enables like-for-like comparison between the downsampling algorithms and their impact.

The missing values for each of the 900 time series are replaced by the `imputeTS` linear interpolation function. This is a simple imputation algorithm that returns each time series to its original length by using curve fitting to create new data points between the remaining data points. [DO I NEED TO EXPLAIN MORE THAN THIS?] The selection of this simple imputation function may limit the findings of this analysis, but returning the time series to their original length enables a like-for-like comparison of downsampling impact. For each `Annotated Change` synthetic time series there is now an imputed time series for parameters 1 to 50 and each downsampling algorithm.

4.3 Features

The `Rcatch22` package [2] enables the computation of a value for each of the catch22 time series features [34]. Values for the 22 time series features are calculated for each of the nine original synthetic time series and 900 imputed time series. The difference between the original and imputed time series feature values is calculated and that difference scaled to accurately compare how the downsampling algorithms impact these catch22 features. The standard deviation across the 900 imputed time series for each catch22 feature is also calculated to indicate the sensitivity of each time series feature to downsampling. The catch22 features are considered to be more sensitive if the standard deviation across the 900 imputed time series is higher.

The application of catch22 features in this way is potentially limited; the time series features reflect characteristics that help cluster and classify different time series. Further analysis is needed to confirm whether features sensitive to downsampling are applicable to all time series types.

4.4 Visualisation

Combining `imputeTS` and `Rcatch22` in this way unstudied; initial visual analysis to observe the impact of downsampling on catch22 features is conducted utilising bar and line graphs alongside heatmaps. All the decision-makers who were interviewed as part of the aforementioned user research emphasised the importance of visualisation in data transparency and data-driven decision-making. In line with this, different approaches to visualising the impact of downsampling are tried to explore how best the impact of downsampling could be visually communicated by data practitioners for decision-makers. However, the visualisations created by this research are under-developed; useability research needs to be conducted with decision-makers and data practitoners to ensure the visualisations meet the aim of this research.

5 RESULTS AND EVALUATION

The results of the chosen analysis methodology are shared in this section with a discussion of their the limitations and potential avenues for further investigation.

5.1 Downsampling sensitivity

Seven catch22 features appear to a wider dispersion across the 900 imputed time series than the other 15 features. These seven features are considered to be more sensitive to the impacts of downsampling. To demonstrate the sensitivity of the seven features, the ten features with the highest standard deviation across

all the imputed time series in the bar graph below. The standard deviation of all 22 features is shared in Annex B.

Standard Deviation of Sensitive Catch22 Features
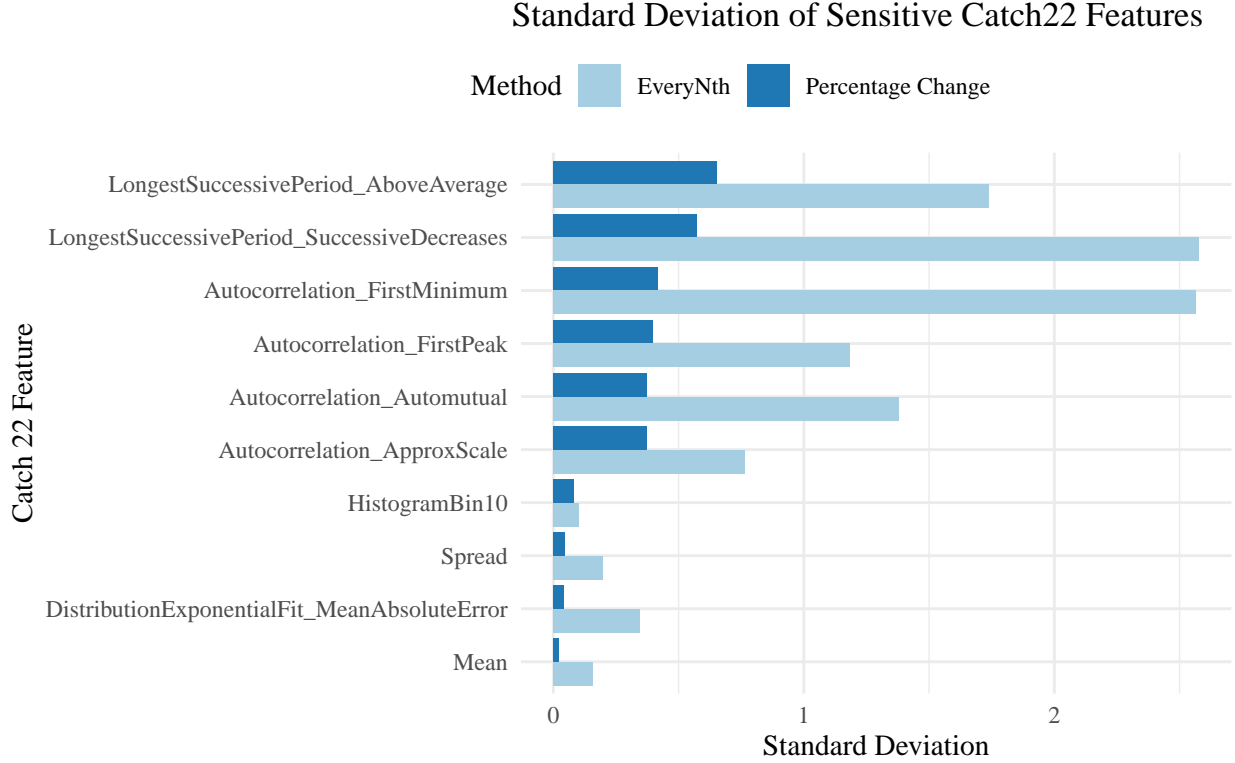
This visualisation suggests that the *EveryNth* algorithm impacts the catch22 time series features more than the *Percentage Change* algorithm. Initial analysis indicates that this is because the *EveryNth* algorithm discards more data, more quickly in the 50 parameters of downsampling. The sensitivity of these features to the *Percentage Change* is limited by the number of parameted used for this research. It would be beneficial to investigate the impact of *Percentag Change* further by creating downsampling parameters beyond 50 until the same volume of data is discarded by both algorithms.

The standard deviation of the seven most sensitive catch22 features is presented in the table below. This table includes the standard deviation of these features for the *EveryNth* algorithm ('everyNthSensitivity'), the *Percentage Change* algorithm ('PercentageChangeSensitivity'), and the combined standard deviation ('CombinedSensitivity').

Table 2: Standard Deviation of the Seven Most Sensitive Catch22 Features

| Feature	everyNthSensitivity	PercentageChangeSensitivity	CombinedSensitivity
Autocorr_FirstMinimum	2.5639294	0.4181318	2.5253400
LongestPeriod_Decreases	2.5778227	0.5711553	2.0901832
LongestPeriod_AboveAverage	1.7389139	0.6524714	1.7545755
Autocorr_Automutual	1.3808674	0.3736266	1.3890246
Autocorr_FirstPeak	1.1844657	0.3954080	0.9515096
Autocorr_ApproxScale	0.7661026	0.3713025	0.7172823
DistributionExponentialFit_MAE	0.3440829	0.0437543	0.3605538

The combined standard deviation is used to identify the seven catch22 features that appear to be most sensitive. The table suggests that the first minimum of the autocorrelation function ('Autocorr_FirstMinimum' or 'CO_FirstMin_ac') is most sensitive across both of the algorithms; that the catch22 feature measuring the longest sequence of successive steps that decrease ('LongestPeriod_Decrease' or 'SB_BinaryStats_diff_longstretch0') is most sensitive to the *EveryNth* algorithm;

and, the longest sequence of successive values greater than the mean ('LongestPeriod_AboveAverage' or 'SB_BinaryStats_mean_longstretch1') is most sensitive to the *Percentage Change* algorithm.

These results suggest that it may be possible to communicate the impact of downsampling via the sensitivity of the catch22 features. However, standard deviation is a limited measure that suggests, rather than confirms, sensitivity to the impacts of downsampling algorithms. Further statistical analysis is needed to confirm that the changes in the features is directly caused by the impact of downsampling.
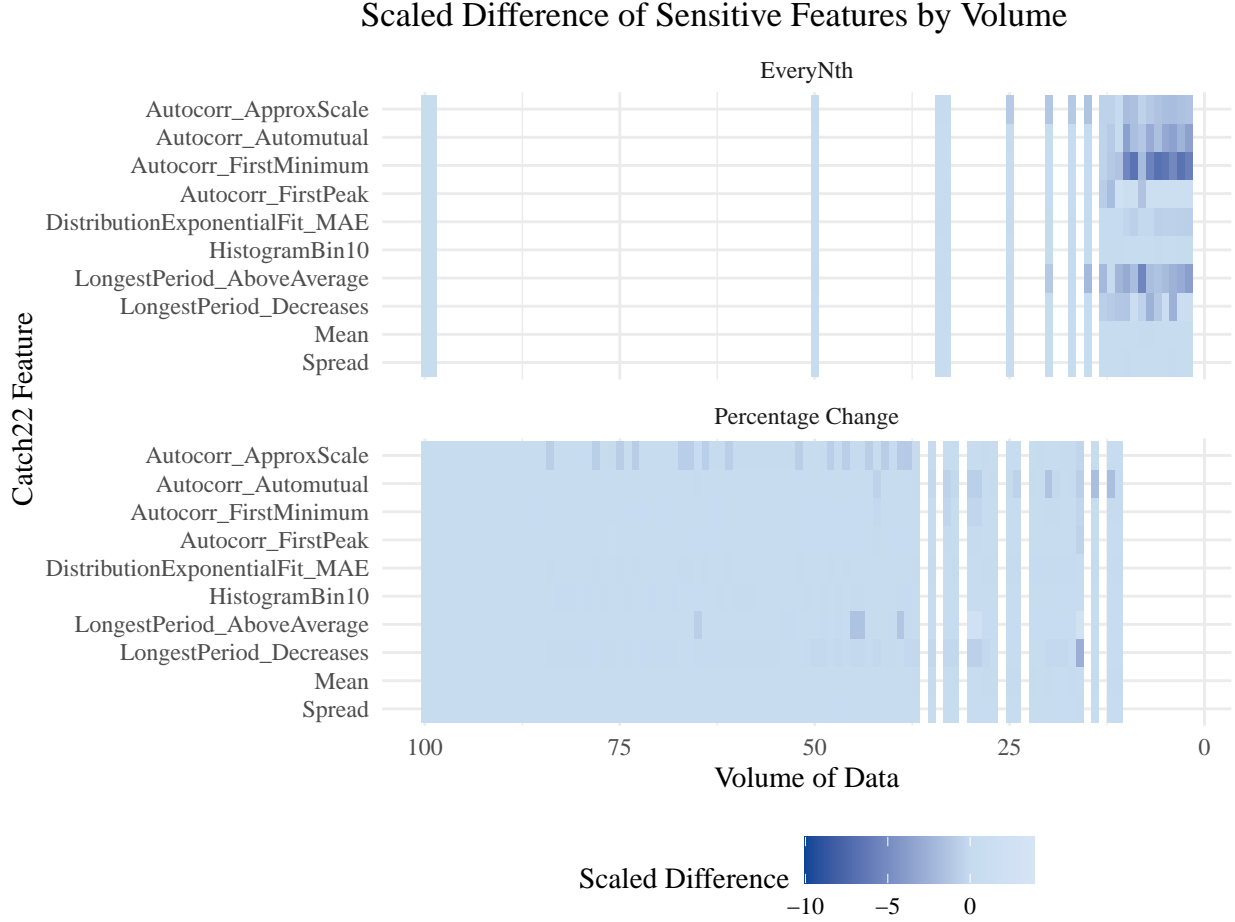
5.2 Downsampling Impact

Identifying that seven catch22 features appear to be more sensitive to the impacts of downsampling enables data pratitioners to evaluate and communicate the impact of different downsampling algorithms without deep statistical analysis. The difference between the values of catch22 features for the nine original time series and 900 imputes time series is used to measure the impact of downsampling. The heapmap below visualises this difference, scaled, for the seven most sensitive catch22 features across the 900 imputed time series and each of the 50 parameters.

Scaled Difference of Sensitive Features by Parameter
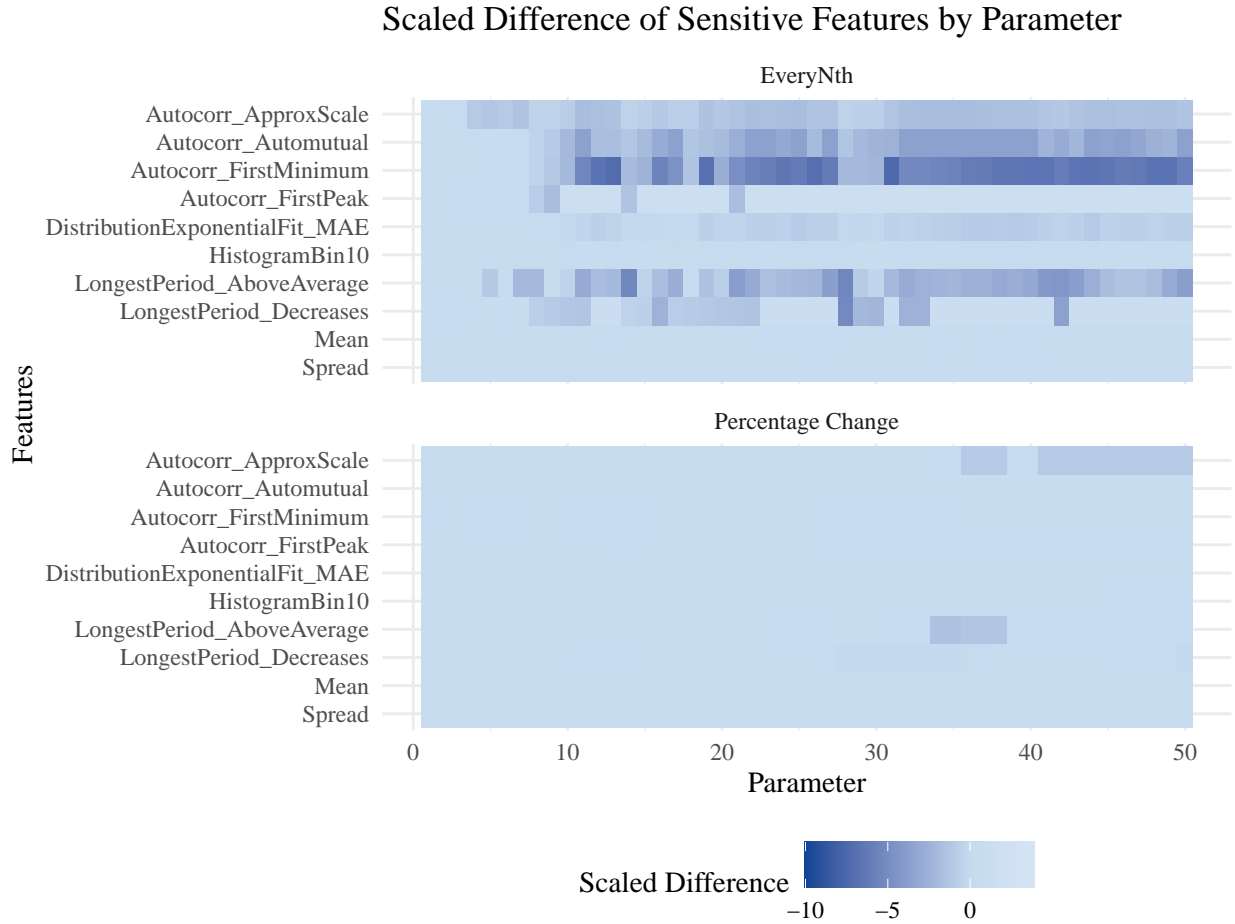
This visualisation suggests that the approximate scale of autocorrelation ('Autocorr_ApproxScale' or 'CO_f1ecac') and the longest sequence of successive values greater than the mean ('LongestPeriod_AboveAverage' or 'SB_BinaryStats_mean_longstretch1') are the first catch22 features to be impacted by the downsampling algorithms. Interestingly, this visualisation also demonstrates that the catch22 features that are impacted by both downsampling algorithms tend to increase in comparison the original feature values. This dynamic warrants further investigation.

However, the parameters prevent a like-for-like comparison of the two downsampling algorithms as *EveryNth* discards data more quickly than *Percentage Change*. The heatmap below also visualises the scaled difference between the values of the original catch22 features and the imputed time series, but across the volume of data retained by the downsampling algorithms before `imputeTS` imputed the missing values.

## Scaled Difference of Sensitive Features by Volume

This visualisation demonstrates that data volume retained by the downsampling algorithms creates acute differences, especially when less than 20 data points remained after *EveryNth* is applied. Interestingly, the impact of *Percentage Change* appears to be inconsistent; there are some retained data volumes where `Autocorr_ApproxScale` and `LongestPeriod_AboveAverage` do not appear to be impacted by *Percentage Change* even though larger retained data volumes appear to be impacted. This dynamic warrants further investigation.

It is not possible, however, to understand the impact of both downsampling algorithms from this visualisation. The *Percentage Change* algorithm needs to be applied for more than 50 parameters so that it discards volumes of data comparable to the *EveryNth* algorithm. Despite this, the heatmap visualisation of downsampling impact holds potential; it suggests that, with the same volume of data, the downsampling algorithms impact the catch22 features differently.

5.3 Feature Variation

The impact of both downsampling algorithms on the most sensitive features varies across different time series types. This is exemplified by the catch22 feature 'Autocorr_FirstMinimum' ('CO_FirstMin_ac'): the first minimum of the autocorrelation function ('Autocorr_FirstMinimum' or 'CO_FirstMin_ac.' This feature has the highest combined standard deviation across the 900 imputed time series represents "the number of steps into the future at which a value of the time series at the current point and that future point remain substantially ($>1/e$) correlated" **feature_book?**. The line graphs below highlight how this feature changes over the time series interpolated from different volumes of data after both downsampling algorithms are applied. The difference between the catch22 feature value of the original and imputed time series has been scaled for better comparisons.

## Variation of the First Minimum of the Autocorrelation Function
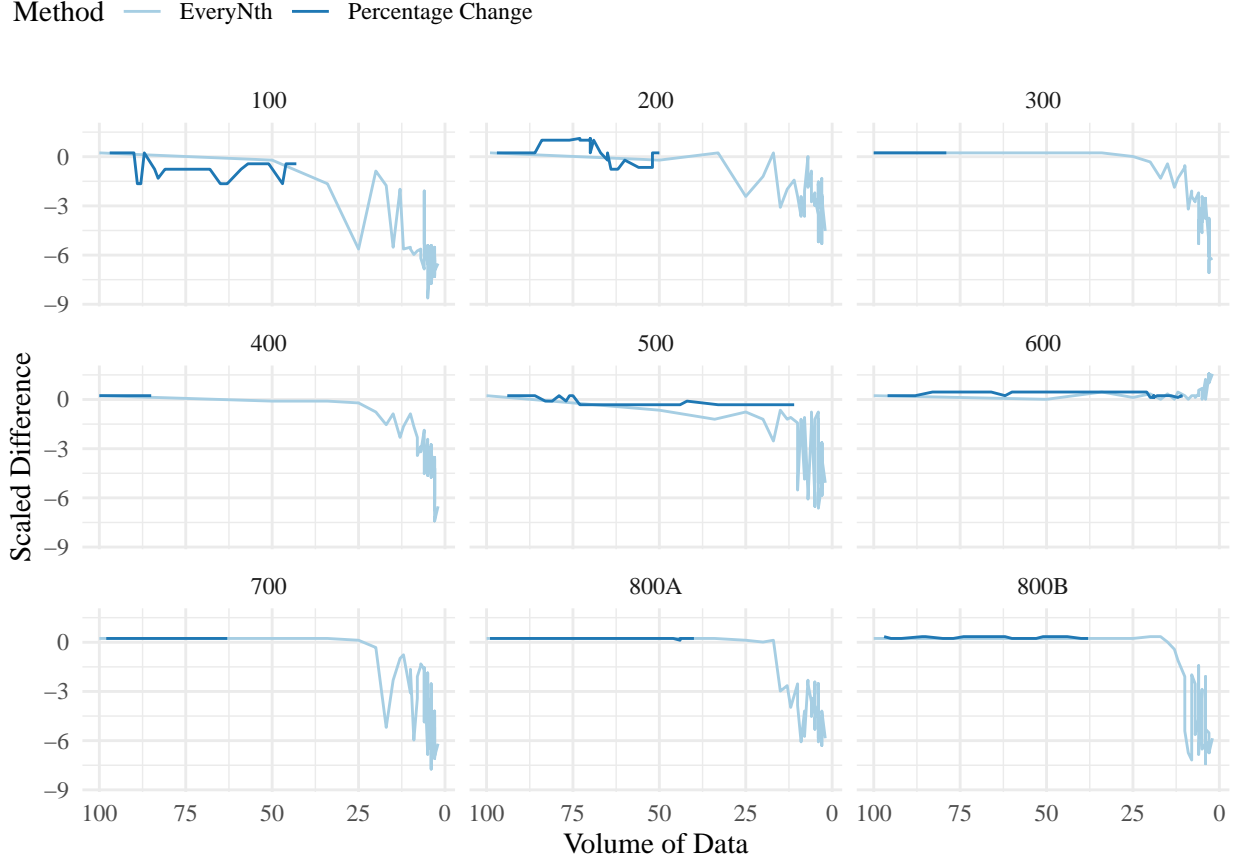
The visualisations present the impact of discarding data for both downsampling algorithms; the volume of data retained by *Percentage Change* algorithm varies by the type of time series whereas the volume of data retained by the *EveryNth* algorithm trends towards zero.

This variation across time series types also suggests that different catch22 features may be more or less sensitive to downsampling for different time series types. This dynamic is invesitgated by the line graphs below, which plot the scaled difference of the six most sensitive catch22 features for time series '500.' Time series '500' was selected as more data is discarded by the *Percentage Change* algorithm, offering a better comparison.

The visualisation below highlights that, although 'Autocorr_FirstMinimum' is identified as the most sensitive across both downsampling algorithms, 'Autocorr_Automutual' might better indicate the impact for time series '500.' 'Autocorr_Automutual' is the minimum of the automutual information function ('Autocorr_Automutual' or 'IN_AutoMutualInfoStats_40_gaussian_fmmi'); this catch22 feature outputs a measure of "autocorrelation in the time series, as the minimum of the automutual information function" **feature_book?**.

Variation of the most sensitive catch22 time series features is shared in Annex []. It would be beneficial to investigate this dynamic further as it may offer data practitioners method for identifying features that best indicate the preservation of a time series after downsampling is applied.

Variation of Catch22 Features: Time Series 500
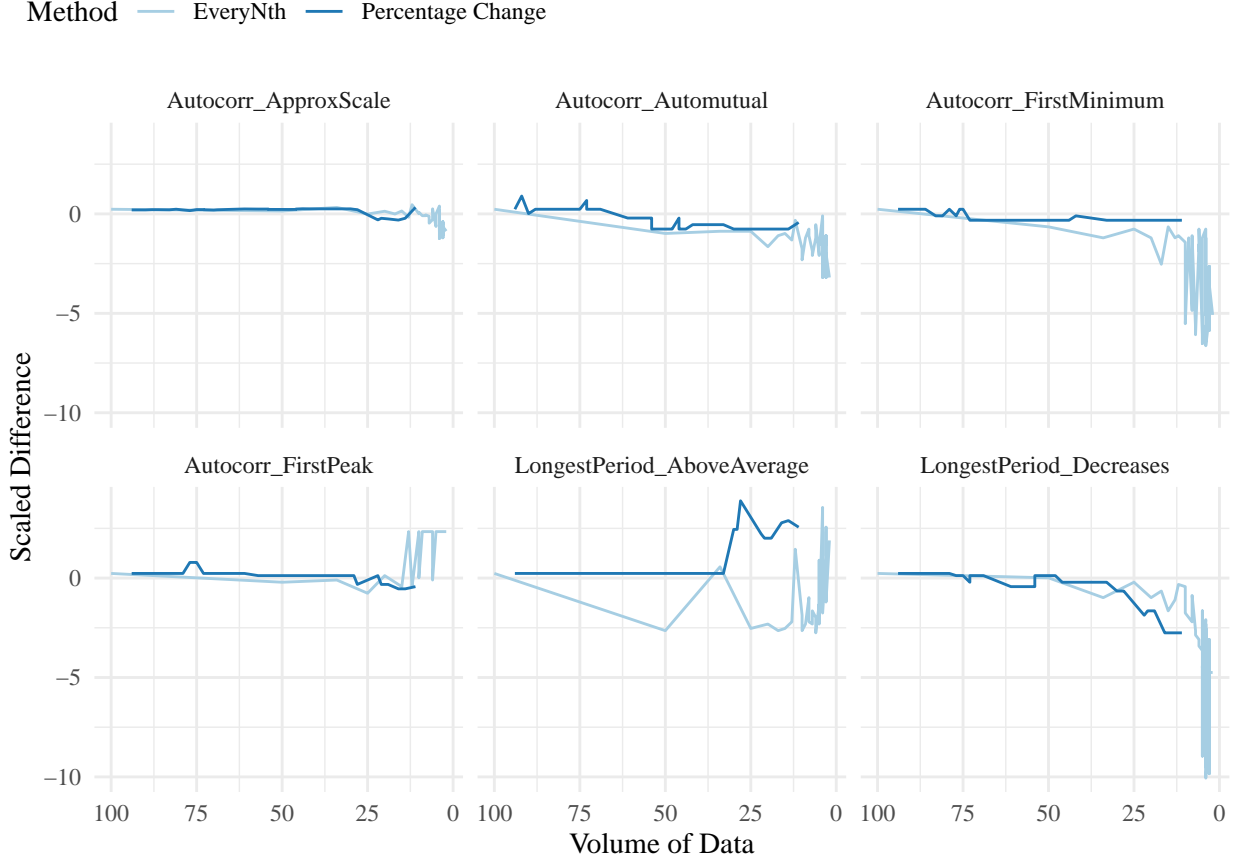
6 FUTURE WORK

As outlined throughout the previous section, there are many avenues for future work to better understand the findings and address the limitations of this research.

6.1 User research There is additional analysis that could be conducted on the transcripts and recordings of the user research. For example, using topic modelling in natural language processing to better identify the connections between the key themes identified by this research. It would also be helpful to conduct user research with a wider group of decision-makers and data practitioner to develop statsitically significant findings for the UK Government.

6.2 Downsampling Sensitivity The data pipeline developed by C. H. Lubba et. al [34] could be used to generate other subsets of time series features for distinct tasks in different domains. Examining the sensitivity of the features in these new subsets to downsampling is likley to generate the information needed to understand why certain features are more sensitive to downsampling across different types of time series. Further

6.3 Downsampling Impact Further iterations of dowsampling on the synthetic time series used in this research would be beneficial. This will enable a better comparison of the *EveryNth* and *Percentage Change* algorithms. Repeating this for another downsampling algorithm, such as LTTB, and applying this research on real-world time series shared by the Alan Turing Change Point Data set would allow deeper investigation into the impact of downsampling. It would also generate the information needed to realise the full ambitions of this research by developing better visualisations that could then be evaluated with users through usability research.

6.4 Feature Variation Combined, the further work to understand downsampling impact and sensitivity will also generate a better understanding of how the variation across time series features and types could enable the development of thresholds and benchmarks. This could enable the most relevant and sensitive time series features to be embedded in smart sensors, enabling downsampling at the point of collection and reducing demand on computing resources. It could also facilitate more nuanced discussions on the thresholds of downsampling and the impact of these thresholds for decisions and data storage.

7 CONCLUSION

- reflect on the original goal ad addressed the aims
- summarise the significance and meaning of the results (clearly evidenced by work in project)
- summarise the main benefits drawbacks of the work and how the project will impact the field. # REFERENCES

[1]   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.

[2]   C. H. Lubba, B. Fulcher, T. Henderspn, B. Harris, O. r. TL, and O. Cliff, "catch22: CAnonical time-series CHaracteristics." R Journal, 2022. doi: 10.5281/zenodo.6673597.

[3]   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.

[4]   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.

[5]   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.

[6]   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.

[7]   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`.

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

[10]   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.

[11]   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.

[12]   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.

[13]   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.

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

[15]   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.

[16]   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.

[17]   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.

[18]   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.

[19]   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.

[20]   R. Matheus, M. Janssen, and T. Janowski, "Design principles for creating digital transparency in government," *Government Information Quarterly*, vol. 38, no. 1, 2021, doi: `https://doi.org/10.1016/j.giq.2020.101550`.

[21]   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.

[22]   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.

[23]   J. Walker, R. Borgo, and MW. Jones, "TimeNotes: A study on effective chart visualization and interaction techniques for time-series data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, 2016, doi: 10.1109/TVCG.2015.2467751.

[24]   J. Donckt, J. Donckt, E. Deprost, and S. Hoecke, "Plotly-resampler: Effective visual analytics for large time series," *IEEE Visualization and Visual Analytics*, 2022, doi: 10.1109/VIS54862.2022.00013.

[25]   S. Moritz and T. Bartiz-Beielstein, "imputeTS: Time series missing value imputation in r." R Journal, 2017. Available: `https://cran.r-project.org/web/packages/imputeTS/vignettes/imputeTS-Time-Series-Missing-Value-Imputation-in-R.pdf`

[26]   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`

[27]   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`

[28]   HM Government, "Global britain in a competitive age: The integrated review of security, defence, development and foreign policy." GOV.UK, 2021. Available: `https://www.gov.uk/government/publications/global-britain-in-a-competitive-age-the-integrated-review-of-security-defence-development-and-fo global-britain-in-a-competitive-age-the-integrated-review-of-security-defence-development-and-fo`

[29]   I. Foster *et al.*, "Computing just what you need: Online data analysis and reduction at extreme scales," in *Euro-par 2017*, F. F. Rivera, T. F. Pena, and J. C. Cabaleiro, Eds., in Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). Germany: Springer Verlag, 2017, pp. 3–19. doi: 10.1007/978-3-319-64203-1_1.

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

[31]   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.

[32]   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`

[33] 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`

[34] C. H. Lubba, S. S. Sarab, P. Knaute, S. R. Schultz, B. D. Fulcher, and N. S. Jones, "catch22: CAnonical time-series CHaracteristics," *Data Mining and Knowledge Discovery*, vol. 33, 2019, doi: 10.1007/s10618-019-00647-x.

[35] B. D. Fulcher and N. S. Jones, "Highly comparative feature-based time-series classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 3026–3037, 2014, doi: 10.1109/TKDE.2014.2316504.

[36] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, "The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances," *Data Mining and Knowledge Discovery*, vol. 31, 2017, doi: 10.1007/s10618-016-0483-9.

[37] T. Henderson and B. D. Fulcher, "An empirical evaluation of time-series feature sets," in *2021 international conference on data mining workshops (ICDMW)*, 2021, pp. 1032–1038. doi: 10.1109/ICDMW53433.2021.00134.

[38] S. Alspaugh, N. Zokaei, A. Liu, C. Jin, and M. A. Hearst, "Futzing and moseying: Interviews with professional data analysts on exploration practices," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 22–31, 2019, doi: 10.1109/TVCG.2018.2865040.

[39] G. J. J. van den Burg and C. K. I. Williams, "An evaluation of change point detection algorithms." 2022. doi: 10.48550/arXiv.2003.06222.

[40] M. F. et. al, "Jettison MVP code." 2023. Available: `https://github.com/MattForshaw/Jettison/tree/main/MVPCode`