
THE EXPLAINABILITY OF TIME SERIES DOWNSAMPLING

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

Abstract

Trusting data-driven decision-making goes beyond demonstrating compliance with legal, regulatory and ethical obligations; decision-makers also 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. Downsampling is an established technique to select a representative subset of time series data that preserves the original data shape while reducing the number of data points in the time series. The research outlined by this paper explores decision-makers' trust in data, data practitioners' 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 combines user research insights from interviews with 16 UK Civil Servants with analysis of time series features for 900 imputed time series to identify and visualise the features that are most sensitive to downsampling. This research shows the potential for improving 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.

1 INTRODUCTION

HM 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]. The UK public need to have “*confidence and trust in how data, including personal data, is used*” [2], and this requires transparency [5].

To make data-driven decisions, it is likely that government decision-makers' need transparency of how data is used to trust the data insights they are presented with. Trust implies trusting which data points are selected, how this data is 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 assumptions and choices they are making. This could support decision-makers in trusting the data informing their decisions while helping decision-makers understand limitations of the data and thresholds at which data insights can be relied upon for each decision.

The research shared in this paper explores this dynamic with a two pronged approach. User research with 16 UK Civil Servants investigates what helps decision-makers 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 [6], [7]. Time series data is used throughout HM Government [8] to inform decision-makers across various domains [9]; it is also widely generated and used by industry and research [10]. However, the volume of time series data is continuously increasing [6], posing significant challenges for handling and visualising this popular data type [10].

Data practitioners must use methods that reduce the quantity of data (data volume) to align with limitations like processing time, computing costs, storage capabilities, and sustainability ambitions [10]–[12]. Downsampling is an established technique [13], [14] for reducing data volumes. Data reduction is vital for making voluminous time series understandable for human observation [11] and an essential step in many time series database solutions [6]. However, little attention has been devoted to how downsampling impacts decision-makers trust in the data. How to communicate the impact of downsampling algorithms on time series data also remains understudied [6], [11].

Downsampling is an important part of how time series data is used and is a technical issue whose impact, this research posits, is important for decision-makers to understand. Downsampling expands the boundaries of uncertainty and risk for decision-makers as data practitioners may not realise the significance of the data being discarded. For example, it is time and computing resource intensive to transport the terabytes of data generated daily by offshore oil platforms [15] without data discarding techniques like downsampling [10]. For decision-makers of platform safety and productivity, it is vital to have this information as quickly as possible and understand how the discarded data limits the insights they are considering. It is also important a decision-maker looking at analysis of CO_2 levels on an open-plan office floor to decide how many officials can work in the office each day post-Covid. Different downsampling algorithms preserve different data points [6], [7], [11], which could change the risk threshold for the decision-maker. In complex decisions, such Choices throughout the data pipeline may have cumulative or disproportionately larger consequences later as their future ramifications are unlikely to be fully understood [16]. 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 how data is used 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` [17] and `Rcatch22` [18] 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 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 for this research and identify the gaps being addressed by the paper.

2.1 Data transparency

Technology transparency, including institutional data practices, is sociopolitical in nature [19]. There is a growing number of researchers reflecting on society's transparency needs “*in terms of what is made transparent, for whom, how, when and in what ways, and, crucially, who decides*” [20].

The implicit assumption behind calls for transparency is that “*seeing a phenomenon creates opportunities and obligations to make it accountable*” and, therefore, to change it [21]. However, without sufficient agency to explore the information being shared, seeing a phenomenon often results in information overload that obfuscates or diverts [22]. Without agency, transparency is increasingly considered to be a fallacy [23].

Meaningful transparency is only realised when the information is provided with the tools to transform access into agency [21], [23]. This suggests that data practitioners simply communicating the assumptions and choices made throughout the data processing pipeline to 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 [6], that put agency into the hands of decision-makers.

2.2 Time series visualisation

Time series data is commonly visualised as a line graph [11], [24], [25]. Line graphs help the human eye to observe only the most important data points [11] by conveying the overall shape and complexity of the time series data [6], [13]. The most effective time series visualisations are, however, interactive [24]–[26], turning access into agency [23] by allowing the user to access details on demand. Evaluation of time series visualisation is, therefore, a growing field of research [24].

However, this growing field of research does not extend to visualisations of choices and assumptions made during data processing pipeline. Indeed, such visualisations are often a side effect of research. This is

exemplified by the R package `imputeTS` [17] 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 [27], [28]. 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 an unprecedented amount of time series data and this data volume continues to grow [2], [10], [29], [30]. For example, tackling climate change is the UK Government’s “*number one international priority*” [31], yet climate simulations that help inform decision-makers generate tens of terabytes per second [10], [32]. Value preserving data aggregation (a subset of downsampling) algorithms play an important role in addressing how this voluminous data is processed, stored [10], and visualised [11], [33] by minimising the computing resources needed [10], reducing network latency, and improving rendering time [6], [7].

Value preserving data aggregation algorithms preserve the original data points that are considered to best represent the original data. This research uses two of the simplest algorithms: *EveryNth* [34] and *Percentage Change* [35]. Descriptions of commonly used value preserving data aggregation algorithms are provided in the table below [6], [7], [11], [33]–[37]:

Table 1: Example Downsampling Algorithms

Algorithm	Description
EveryNth	Selects every nth data point.
Percentage Change	Selects every data point beyond a specified percentage of the previous data point
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 regression.
MinMax	Preserves minimum and maximum data points in each data bucket.
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 and selects one point per bucket which forms the longest total line length through all the buckets.
Largest Triangle Three Bucket (LTB)	Calculates the largest triangular surface between the previously selected data point and the average value of the next data bucket to select the highest point.
MinMaxLTB	Preselects data points using MinMax before applying Largest Triangle Three Buckets.

Data practitioners have recently made advances in the performance and evaluation of downsampling approaches [6], [7], [13], [14], [26], [33]–[36]. These advances focus on the effectiveness of the algorithm in delivering downsampled data that represents the original data as accurately as possible. This is a vital part of enabling and improving data-driven decision-making, but is focused on supporting data practitioners in their data analysis. 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 sizes of modern time series data sets have also generated new research into the characteristics of time series [38]. 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 highly comparative time series analysis `hctsa` toolbox. This shares the 4791 high-performing features after computationally comparing thousands of characteristics from across scientific time series analysis literature [39].

Utilising such a library, however, is computationally expensive [38]. C. H. Lubba et. al have attempted to address this by identifying a subset of 22 features (as well as mean and standard deviation spread) that are tailored to time series data mining tasks [38], [40], which are shared in the `Rcatch22` R package. Although further research is needed to evaluate the relative performance of different feature sets on different types of problems, the ‘`catch22`’ features performs well against other feature libraries across 800 diverse real-world and model-generated time series [41].

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

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 section outlines how the user research is designed and the key insights it generated.

3.1 Design

User research produced over 16 hours of recorded and transcribed interviews with decision-makers and data practitioners (these outputs are confirmed in Annex A). These interviews are hosted virtually on MS Teams, instead of in-person, to enable recording and transcription; this creates the potential to move beyond qualitative thematic analysis and implement alternative techniques, such as natural language processing. Interviews are selected over surveys because concepts like transparency and trust are complex and nuanced. One-to-one interviews are selected instead of focus groups to minimise authority bias and potential groupthink.

Like many interview studies [42], the participants for this user research are selected by contacting relevant individuals in the researcher’s professional network. This introduces bias as the sample of participants is neither representative nor drawn from the overall population. This could engender confirmation, false consensus and group think biases as these participants are likely to operate in the same professional bubble, which would skew the insights generated by this user research. To mitigate further biases, two interview scripts used; one for all decision-makers and the other for all data practitioners. However, it is important to note that follow-up questions vary to ease the flow of each interview and help participants to share further detail in their responses.

Qualitative analysis is conducted on the content of each interview to understand the themes and identify quotes that reflect the participant’s sentiments. This is combined with natural language processing keyword extraction. The python libraries `docx` and `nltk` are used to import and clean the interviews text data; capitalisation, ‘stopwords,’ and words with three or fewer letters are removed. Interviews are anonymised as the time stamps, interviewer name and participant name are removed [43]; text data from the interviewer is also removed so that natural language processing is only conducted on participant responses. Limited natural language processing is then applied by tokenising the remaining words and counting the number of each unique word. These word counts alongside the qualitative thematic analysis inform concept maps [44] which aim to show the key themes and words.

3.2 Insights

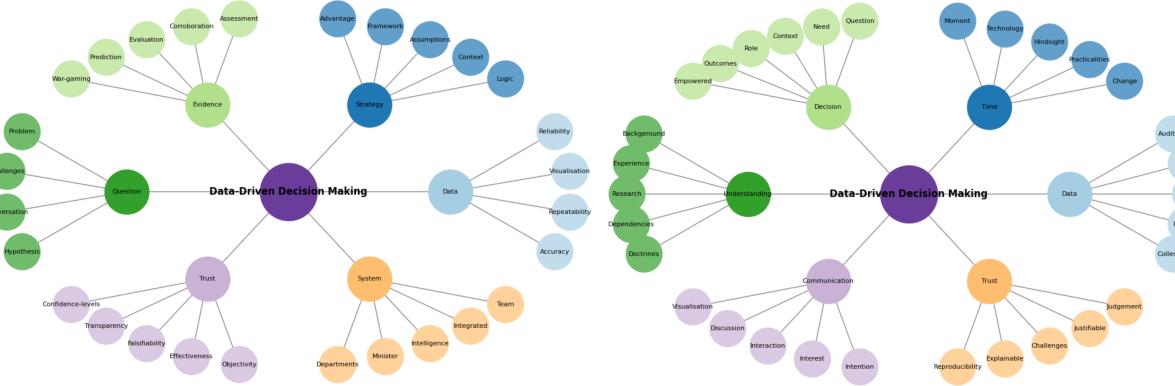
Transparency of data provenance, collection, and analysis pipeline is clearly desired by decision-makers to understand the reliability, repeatability, and quality of the data they’re presented with. All interviewed decision-makers and data-practitioners shared that this is best communicated through engaging conversations and interactive visualisations. Both decision-makers and data practitioners 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 improve this is a topic of live debate.

Insights from this user research indicate that decision-makers desire a spectrum of transparency for decisions. This spectrum depends on the question being addressed by the data: its complexity, importance and potential impact. To trust the data informing data-driven decisions, decision-makers and data practitioners repeatedly request transparency of the data context, source and limitations as well as an assessment of overall confidence in the data.

Quotes from the interviews highlight that decision-makers find it difficult to develop the correct views of “*how much they can trust the data*”. They address 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 share 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 share that visualisations are important for conveying “*specific insights that simply looking at numbers can't get over*” and emphasise that if analysts “*clearly explain what the limitations are and rigorously present that, then I'll trust [the analysis] more*”. Quotes from data practitioner interviews are shared in Annex B.

Figure 1 below shares example concept maps outlining the key themes from two anonymised decision-maker and data-practitioner interviews. All the interview concept maps are shared in Annex C; only summative content from these interviews is presented in compliance with the consent forms shared with the user research participants.

Figure 1: Example Concept Maps



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 endeavours 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 uncertainty or confidence-levels are potential avenues for data science to have greater policy impact.

It is important to reiterate that the findings of this user research are limited and, potentially, biased. Given the consistency of insights from the user research participants, it is clear that further interviews with a wider, random subset of the decision-maker and data practitioner populations in UK Government would be beneficial.

4 ANALYSIS METHODOLOGY

This section outlines the analysis design undertaken to identify and communicate the impact of downsampling algorithms on time series data without extensive statistical evidence.

4.1 Data

The eight Alan Turing Institute Annotated Change [45] synthetic time series are selected for initial analysis. These are selected because they are cleaned time series from a reputable institution with known change points that are designed to provide examples of different types of time series. However, there is an inherent limitation in using synthetic time series for analysis because the data has been generated for a different purpose. These synthetic time series ('100,' '200,' '300,' '400,' '500,' '600,' '700,' '800') are used in demonstrations to support

change point annotators annotate real-world data sets for the **Turing Change Point Data set** [45]. The eighth **Annotated Change** synthetic time series ('800') is multivariate and thus split in two ('800A' and '800B') to create nine synthetic time series for this research and enable better comparison. The nine time series are visualised in Annex D.

4.2 Downsampling

These synthetic time series are imported into **RStudio** from **JSON** scripts and two downsampling algorithms are applied to all with the **Jettison MVP Code** [46]. This enables initial insights into the impacts of downsampling with minimally complex code. The two algorithms used by the **Jettison MVP Code** [46] are *EveryNth*, which selects every other data point, and *Percentage Change*, which selects every data point that has greater than 1% difference between the last transmitted and newest values.

The **Jettison MVP Code** [46] 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** [17] 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 and is a commonly used approach to present line graphs [47]. 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 [18] enables the computation of a value for each of the **catch22** time series features [38]. Values for the 24 time series features are calculated for each of the nine original time series and 900 imputed time series. The difference between the original and imputed feature values is calculated and that difference scaled, dividing by the standard deviation with the **scale()** R function, to accurately compare how the downsampling algorithms impact these **catch22** features. The coefficient of variation across the 900 imputed time series for each **catch22** feature is also calculated to indicate and compare the sensitivity of each feature to downsampling. The **catch22** features are considered to be more sensitive if their coefficient of variation across the 900 imputed time series is higher.

It is important to note that the **catch22** time series features are selected to reflect characteristics that help cluster and classify different time series, not selected for their impact on downsampling. This feature analysis design could be taken further by investigating whether features sensitive to downsampling are equally applicable to all time series types.

4.4 Visualisation

Combining **imputeTS** [17] and **Rcatch22** [18] in this way is unstudied. Initial visual analysis to observe the impact of downsampling on **catch22** features is conducted by using bar and line graphs alongside heatmaps. All the decision-makers interviewed as part of the user research stressed 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 the impact of downsampling could best be visually communicated by data practitioners. The visualisations created by this research are under-developed and would benefit from usability research to ensure they meet the aim of this research.

5 RESULTS AND EVALUATION

This section shares and discusses the results of this research, identifying their limitations and potential avenues for further investigation. These results are presented in three themes: downsampling sensitivity, downsampling impact, and feature variation.

5.1 Downsampling sensitivity

Seven `catch22` features show 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 these seven features, features with the highest absolute coefficient of variation values across all the imputed time series are visualised in the bar graph below. The absolute coefficient of variation values for all 24 features is shared in Annex E.

Figure 2: Coefficient of Variation for Most Sensitive Features

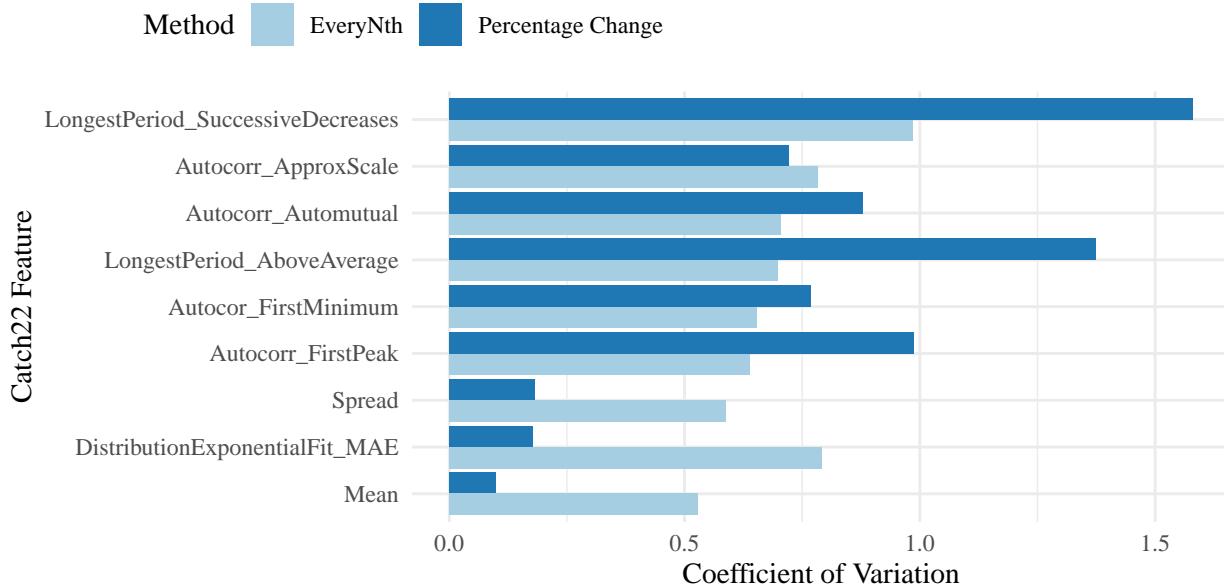


Figure 2 suggests that the *Percentage Change* algorithm impacts the `catch22` time series features more than the algorithm. This is surprising as the *EveryNth* algorithm discards more data, more quickly, in the 50 parameters of downsampling. The sensitivity of these features to the *Percentage Change* is worthy of further statistical analysis to understand why it impacts the feature's coefficient of variation values more. It would also be beneficial to investigate the impact of *Percentage Change* further by creating downsampling parameters beyond 50 until the same volume of data is discarded by both algorithms.

The coefficient of variation values for the seven most sensitive `catch22` features, rounded to two decimal points, are presented in table 2.

Table 2: Coefficient of Variation for the Seven Most Sensitive Catch22 Features

Feature	everyNth	Percentage Change	Combined
LongestPeriod_Decreases	0.99	1.58	1.47
Autocorr_FirstMinimum	0.65	0.77	1.18
Autocorr_Automutual	0.71	0.88	1.14
LongestPeriod_AboveAverage	0.70	1.37	1.12
Autocorr_ApproxScale	0.78	0.72	0.97
Autocorr_FirstPeak	0.64	0.99	0.96
DistributionExponentialFit_MAE	0.79	0.18	0.74

The combined absolute coefficient of variation values are used to identify the seven `catch22` features that appear to be most sensitive. Table 2 shows that the mean absolute error of an exponential fit to a probability distribution [48] (`DistributionExponentialFit_MAE` or `C0_Emb2_Dist_tau_d_expfit_meandiff`) has the highest

combined sensitivity. The longest sequence of successive steps that decrease (`LongestPeriod_Decrease` or `SB_BinaryStats_diff_longstretch0`) is most sensitive to the the *Percentage Change* algorithm. The `catch22` feature measuring the first minimum of the autocorrelation function (`Autocorr_FirstMinimum` or `C0_FirstMin_ac`) is most sensitive to the *EveryNth* algorithm.

These results indicate that it may be possible to communicate the impact of downsampling via the sensitivity of the `catch22` features. This analysis could be taken further by examining whether the coefficient of variation is the best measure of sensitivity and investigating whether the changes in the `catch22` 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 practitioners 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. Figure 3 below visualises this difference scaled (divided by the standard deviation) for the seven most sensitive `catch22` features across the 900 imputed time series and each of the 50 parameters.

Figure 3: Scaled Difference of Sensitive Features by Parameter

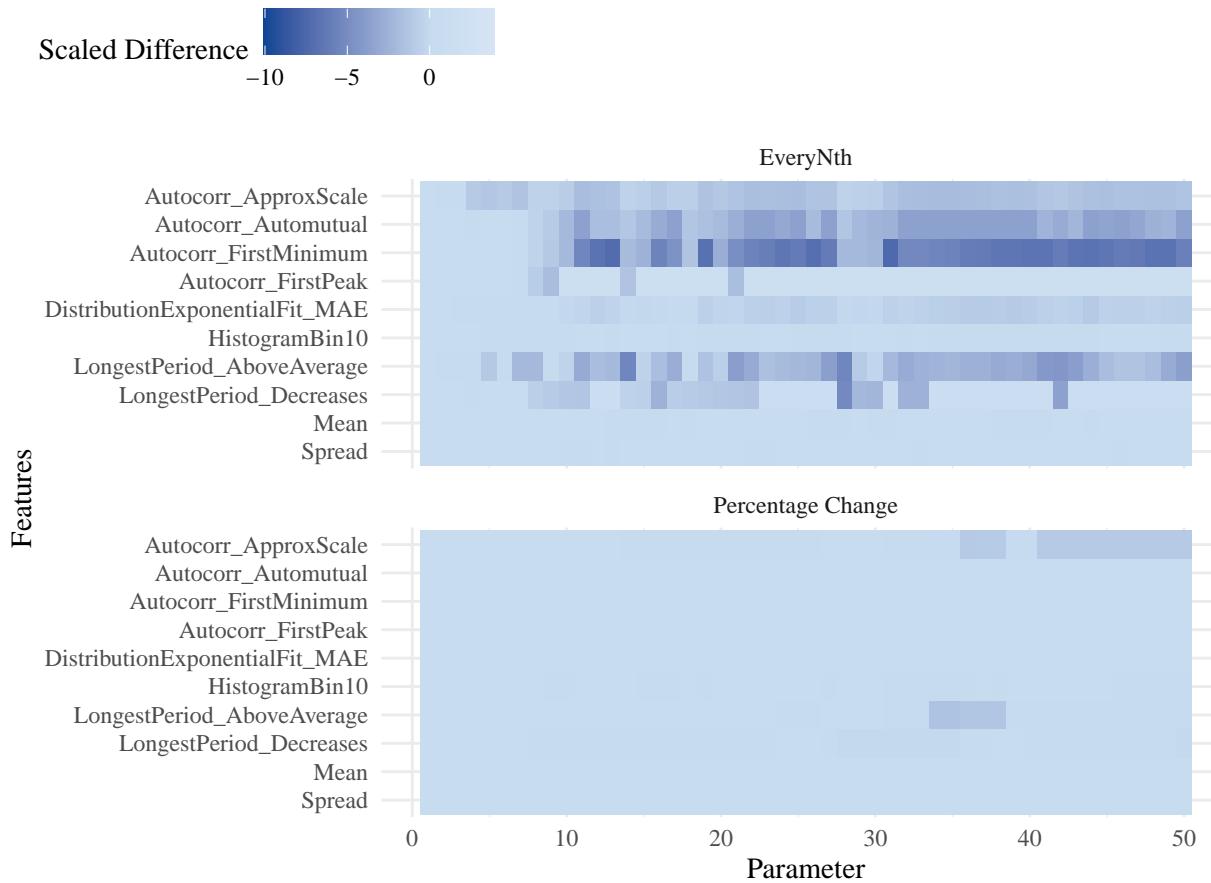


Figure 3 suggests that the approximate scale of autocorrelation (`Autocorr_ApproxScale` or `C0_f1ecac`) and the longest sequence of successive values greater than the mean (`LongestPeriod_AboveAverage` or `SB_BinaryStats_mean_longstretch1`) are the first `catch22` features to be impacted by the downsampling algorithms. Interestingly, Figure 3 also demonstrates that the features values impacted by both downsampling algorithms tend to increase in comparison the original values. This 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*. Figure 4 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` [17] imputes the missing values.

Figure 4: Scaled Difference of Sensitive Features by Volume

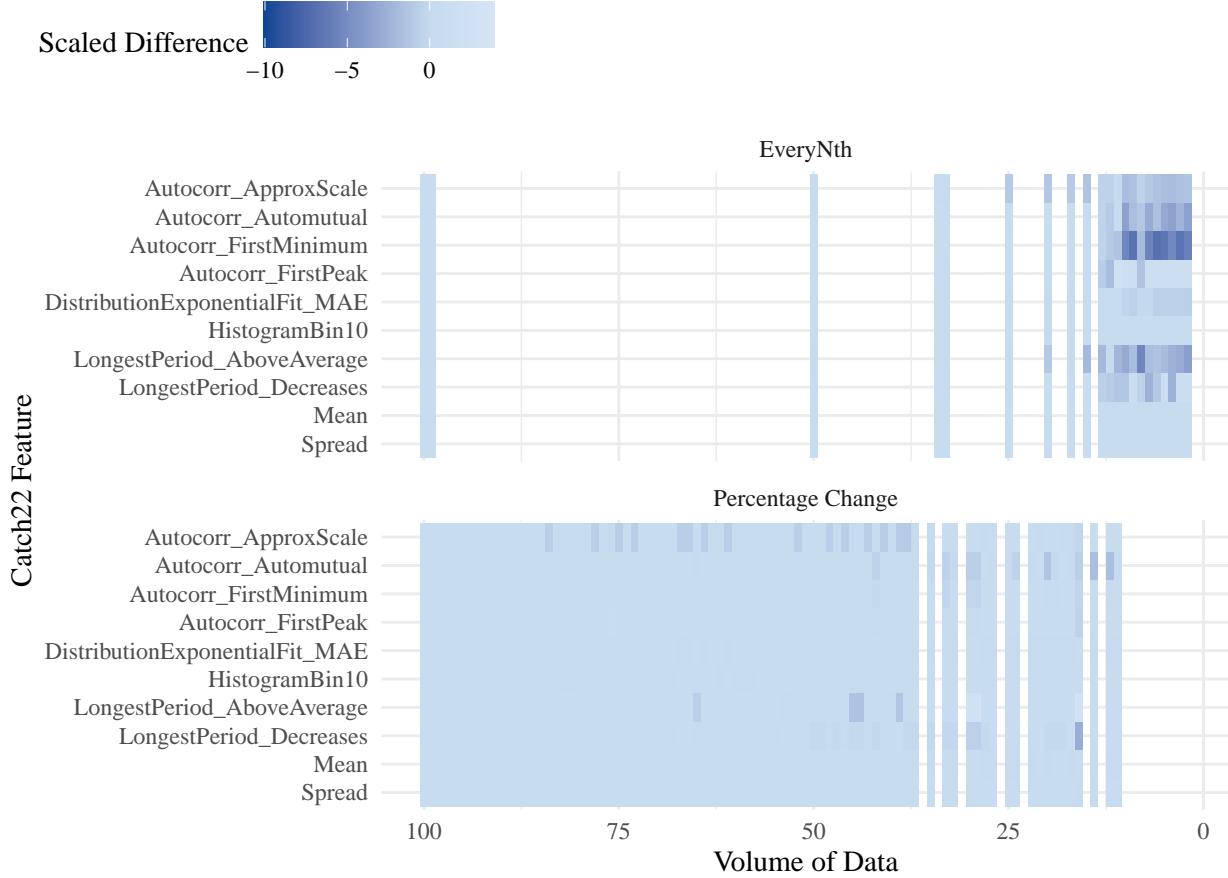


Figure 4 demonstrates that data volume retained by the downsampling algorithms creates acute differences, especially when fewer than 20 data points remained after *EveryNth* is applied. However we cannot conclude that there is a significant difference between the algorithms over and above the volume of data remaining without further investigation. 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 also warrants further investigation.

Downsampling can perturb the statistical properties [45] and visual perception [49] of the data. `Catch22` features are statistical properties of time series. It is likely that some of these statistical properties will be more or less affected by downsampling because downsampling algorithms preserve different data points from the original time series. A full table of the statistical properties represented by each `catch22` feature is shared in Annex F.

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; Figure 4 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 appears to vary across different time series types. This is exemplified by the `catch22` feature `LongestPeriod_Decreases` (`SB_BinaryStats_diff_longstretch0`), which “calculates the longest sequence of successive steps in the

time series that decrease" [48]. The feature has the highest coefficient of variation across the 900 time series; the line graphs in Figure 5 below share the differences between the `catch22` feature value of the original time series and the 50 time series imputed from these after downsampling; these differences are scaled for better comparisons. The title of each line graph refers to the original nine synthetic time series (for example, '100,' '200,' and '300').

Figure 5: Variation of the Longest Sequence of Successive Steps that Decrease

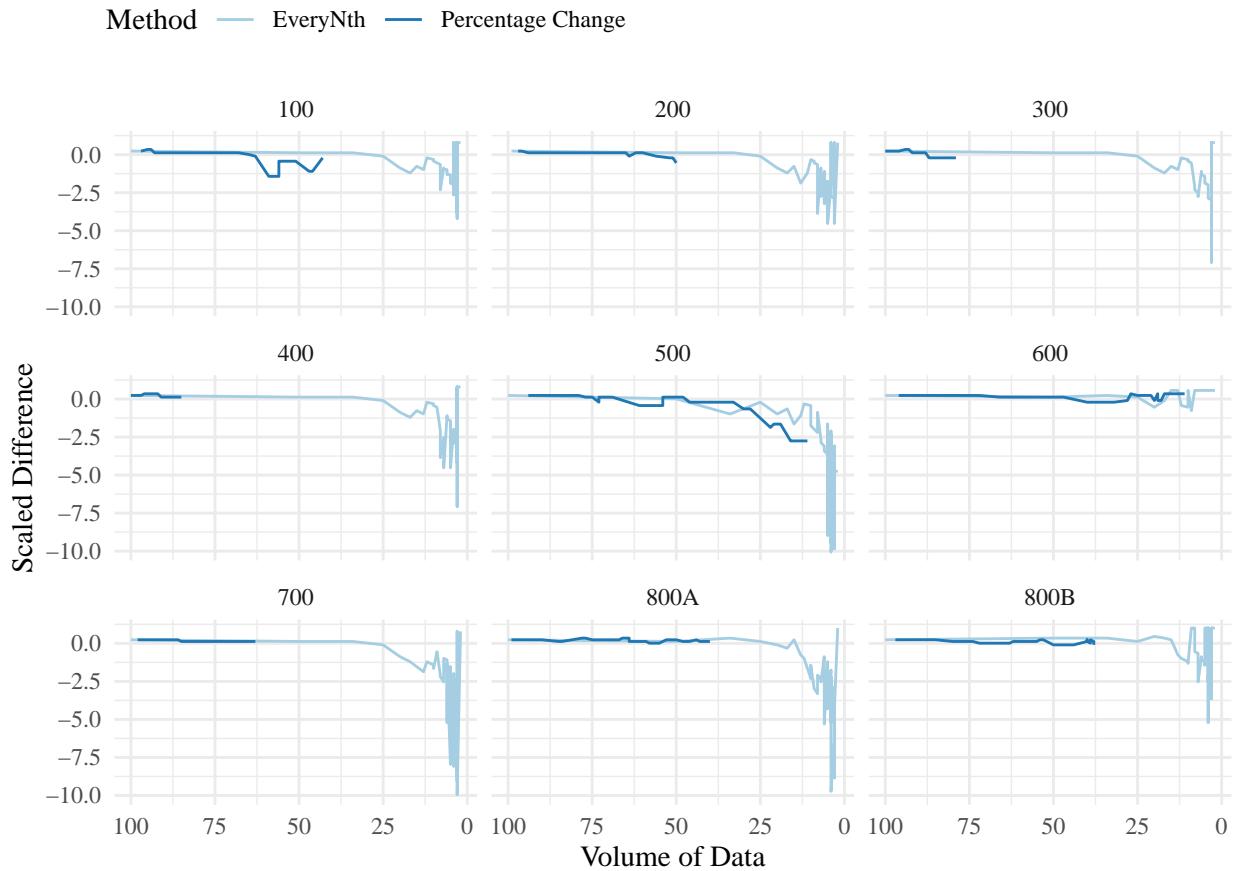
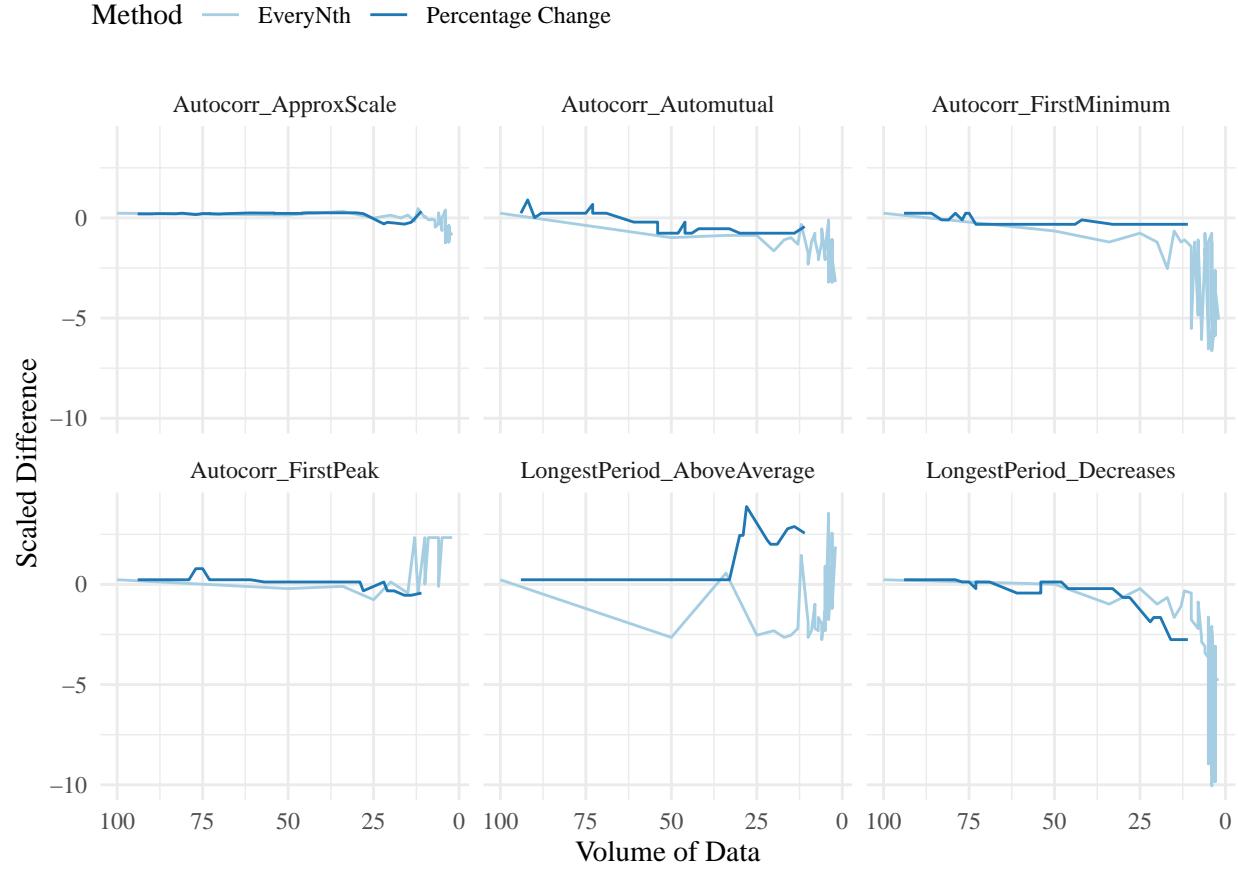


Figure 5 presents the impact of discarding data for both downsampling algorithms. The volume of data retained by the *Percentage Change* algorithm varies by the type of time series whereas the volume of data retained by the *EveryNth* algorithm trends towards zero. The line graphs presented in Figure 5 highlight how different downsampling algorithms may more or less preserve the integrity of different time series. For example, time series '300' and '500' appear to be more impacted by the *Percentage Change* algorithm than other time series. How the most sensitive `catch22` time series features vary across the different time series types is shared in Annex G.

This variation of downsampling algorithms across time series types causes the author to question whether the downsampling sensitivity of `catch22` features may vary across time series type too. An example of this dynamic is visualised by the line graphs in Figure 6 on the next page, 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.

Figure 6 highlights that, although `LongestPeriod_Decrease` is identified as the most sensitive across both downsampling algorithms, `Autocorr_Automutual` (`IN_AutoMutualInfoStats_40_gaussian_fmmi`) might better indicate the impact for time series '500.' `Autocorr_Automutual` is the minimum of the automutual information function; this `catch22` feature outputs a measure of "*autocorrelation in the time series, as the minimum of the automutual information function*" [48]. `Autocorr_FirstMinimum` (`CO_FirstMin_ac`) and `Autocorr_FirstPeak` (`PD_PeriodicityWang_th0_01`) also seem to show early impacts of the *Percentage Change* algorithm.

Figure 6: Variation of Catch22 Features for Time Series 500



Interestingly, Figure 6 shows that the *Percentage Change* algorithm has the opposite impact on *LongestPeriod_AboveAverage* (`SB_BinaryStats_mean_longstretch1`) than the *EveryNth* algorithm.

A comparison of the most sensitive features for each time series type is shared in Annex H. It would be beneficial to explore this further as this variation may offer data practitioners a method for identifying subsets of features that best indicate the preservation of different time series.

6 CONCLUSION

Increasing volumes of time series data are being widely generated and used by industry and research [10]. Downsampling is a vital data processing technique for reducing this volume, addressing limitations like processing time, computing costs, storage capabilities and sustainability ambitions [10]–[12]. However, downsampling can perturb the statistical properties [45] and visual perception [49] of time series, potentially impacting the insights that inform data-driven decisions.

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. Decision-makers shared that there is a spectrum of transparency desired for the data processing pipeline, which depends on the question being addressed by the data; its 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. This user research indicates that decision-makers need transparency of how the data is used to trust the insights informing data-driven decisions.

There are seven `catch22` time series features [38] that appear to be more sensitive to the impacts of downsampling. `Catch22` features are statistical properties of time series so they are more or less affected by

downsampling depending on which data points are preserved from the original time series. The impact of the *everyNth* and *Percentage Change* downsampling algorithms on the most sensitive features appears to vary across different time series types. The downsampling sensitivity of `catch22` features also appears to vary across time series. These results suggest that different subsets of `catch22` features could be selected to best indicate the preservation of time series after downsampling. Tailored subsets of features enable data practitioners to design and build interactive visualisations for decision-makers to explore the limitations of the data they are considering.

The varying sensitivity of `catch22` features to different downsampling algorithms is previously unstudied. This research offers a new visualisation methodology for data practitioners to communicate the impact of downsampling on different decisions and create meaningful transparency about the downsampling choices they are making throughout the data processing pipeline. This could help decision-makers to trust the data informing their decisions, supporting decision-makers to understand the limitations of the data available and the thresholds at which the data insights can or cannot be relied upon for each decision.

7 FUTURE WORK

There are many avenues for future work to better understand the findings, and address the limitations, of this research.

6.1 User research

Further user with a wider, more representative group of decision-makers and data practitioner is needed to develop statistically significant findings for the UK Government. alongside this, additional analysis on the transcripts and recordings of the user research would improve the analysis. For example, using other natural language processing techniques, such as topic modelling, to better identify the connections between the key themes identified. Usability research with decision-makers and data practitioners would also beneficial. Usability would contribute to the refinement of visualisations so that they better communicate the impact of downsampling on different time series. How these visualisations perturb the decision-makers' and data practitioners' perception of downsampling impact is an important avenue of further investigation [49].

6.2 DownSampling Sensitivity

It would be beneficial to use the data pipeline developed by C. H. Lubba et. al [38] to generate other subsets of time series features for distinct tasks in different domains. Examining the downsampling sensitivity of new subsets of `catch22` features is likely to generate the information needed to understand why certain features are more sensitive to downsampling across different types of time series. It could also be interesting to explore the impact of data discarding techniques, such as downsampling, through classifications of missing data [28]. Such classifications may offer alternative features for evaluating and visualising the impact of discarding data.

6.3 DownSampling Impact

Further iterations of downsampling on the synthetic time series used in this research is an important step to address the limitation of this research. These iterations will enable a better comparison of the *EveryNth* and *Percentage Change*. Equally, repeating the analysis methodology of this research for other downsampling algorithms, such as LTTB [7], [11], [36], as well as applying utilising the real-world time series shared by the Alan Turing Change Point Dataset [45] set will allow a deeper investigation into the impact of downsampling. These avenues of future work are also vital for generating the information needed to fully realise the ambitions of this research - developing comparative and interactive visualisations to better communicate the impact of downsampling on time series. These visualisations should also be evaluated with users through usability research [49].

6.4 Feature Variation

The future work for understanding downsampling sensitivity and impact could be combined to develop thresholds and benchmarks for different time series. This avenue for further investigation could enable the most relevant and sensitive time series features to be embedded for proactive downsampling that reduces the demand on computing resources. For example, embedding the most sensitive features in smart sensors monitoring a particular type of time series would enable downsampling at the point of collection, reducing demand on computing resources for storage. Such tailoring of downsampling could also facilitate more nuanced discussions between decision-makers and data practitioners on the thresholds of downsampling for different decision types to ensure that insights from downsampled data are used within the bounds of their limitations.

8 Annex A: Acknowledgement of User Research

In line with the signed consent forms, the user research transcripts and unprocessed findings cannot be shared beyond the researcher and their supervisor. Figure 7 shares the supervisor's email confirmation acknowledging the user research conducted.

Figure 7: User Research Confirmation from Supervisor

Matthew Forshaw
To: Morgan Frodsham (PGT)

Good afternoon Morgan,

Thank you very much for providing these.

I can confirm we have receipt of sixteen consent forms, and you have made available fifteen full transcripts from your discussions with all but one participant who did not give permission for their transcript to be shared.

You are very welcome to include a copy of this email in the Annexes of your thesis.

Best wishes,
Matt

Dr Matthew Forshaw (he/him)
Reader in Data Science, Newcastle University (0.5FTE) and Co-Lead for the Scalable Research Group
Senior Advisor for Skills, The Alan Turing Institute (0.5FTE)

Email: matthew.forshaw@ncl.ac.uk
URL: <https://www.ncl.ac.uk/computing/staff/profile/matthewforshaw.html>

I work a flexible working pattern so may send emails out of 'normal' working hours. Please be assured that I do not expect a response outside of your own working hours.

[Reply](#) | [Forward](#)

9 Annex B: Data Practitioner Quotes

Interesting (anonymised) quotes that highlight the sentiment of data practitioners from user research interviews are shared below:

- “*You come at it from a particular viewpoint with your particular expertise, somebody with different background, somebody with different expertise might come at the problem in a different way.*”
- “*... my impression is that a lot of people make decisions from data without actually understanding what that data represents or simultaneously, and what the theoretical assumptions on our data and data collection is.*”
- “*... getting the point across of how downsampling and other techniques effects that data getting that balance right of explaining the assumptions and explaining what what's been done without getting in the technical details...*”
- “*I would probably think about the impact of the decision and that would inform me about whether I'm comfortable making a decision on that with that data.*”
- “*... we have to understand what things you may think of the data tells you, but it doesn't actually tell you because that's where you often get caught out.*”
- “*... making the assumption of what people how people will think the data will be interpreted versus how you can interpret the data is a really, really useful because you go, OK, people gonna want to say this about this data, but this is what actually sets.*”
- “*... if you have done the due diligence, generally decision makers trust you to do the due diligence.*”
- “*The best that you can say in general is that you have to understand what the information requirement is and what the information processing is that happens, and then you can make an estimate of the effect that downsampling, or indeed any information reduction technique has on the on the decision.*”
- “*... ensure that people are thinking about what the processes that they are going through in order to make the decision, not so, not just taking the recommendation but actually understanding how it's got there and where the data came from and how that was achieved.*”
- “*I think [helping decision-makers trust data] is quite dependent on the personality of who you're dealing with at the time and what they're willing to kind of expose of their own understanding and lack of understanding or motivation.*”
- “*Data is collected for a purpose then being used for another purpose and they might not exactly map on top of one another, which I think is quite challenging. Often we end up getting data that was collected by somebody else for something else, and we've got to shoehorn it into what we are interested in.*”
- “*I haven't engaged with stakeholders on a level like that to understand whether they are aware of the impacts of decisions like that on any end results and how much of a difference that can make... but it is very much a stakeholder problem because it impacts what you will get out of it.*”

10 Annex C: User Research Concept Maps

Key word counts alongside the qualitative thematic analysis inform concept maps which visualise the key themes and key words. The concept maps for each decision-maker and data practitioner are shared in this section.

Figure 8:

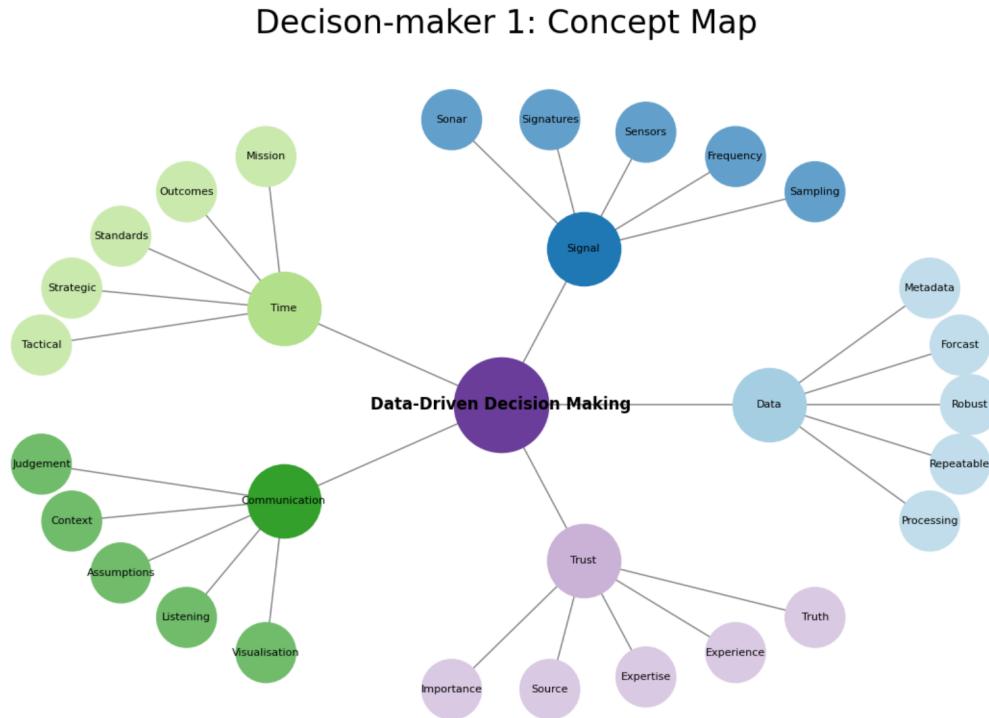


Figure 9:

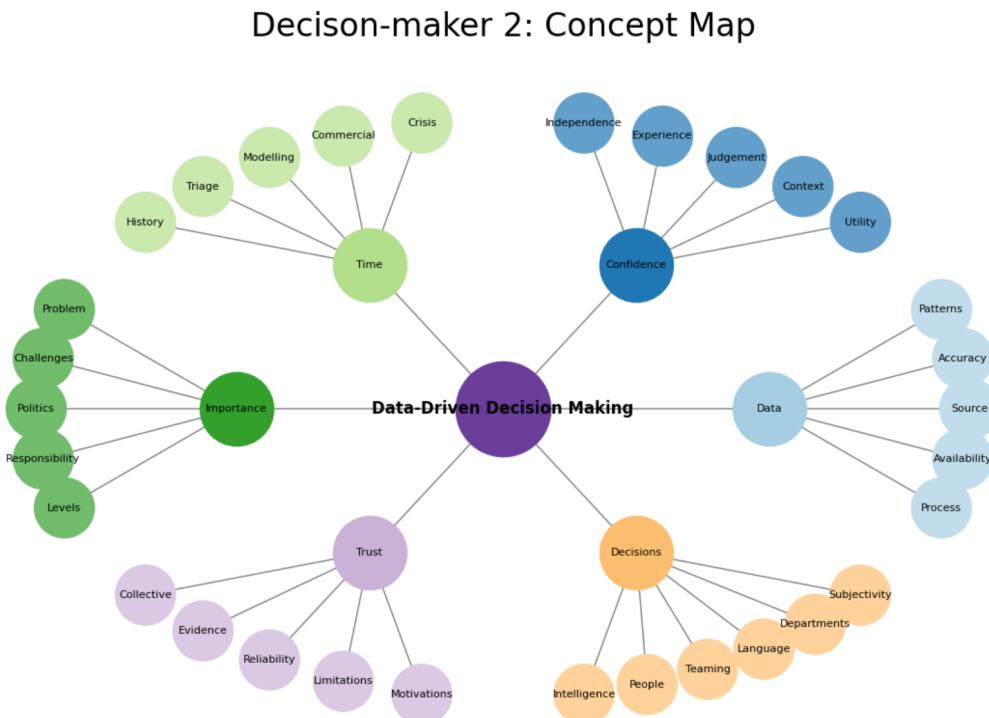


Figure 10:

Decision-maker 3: Concept Map

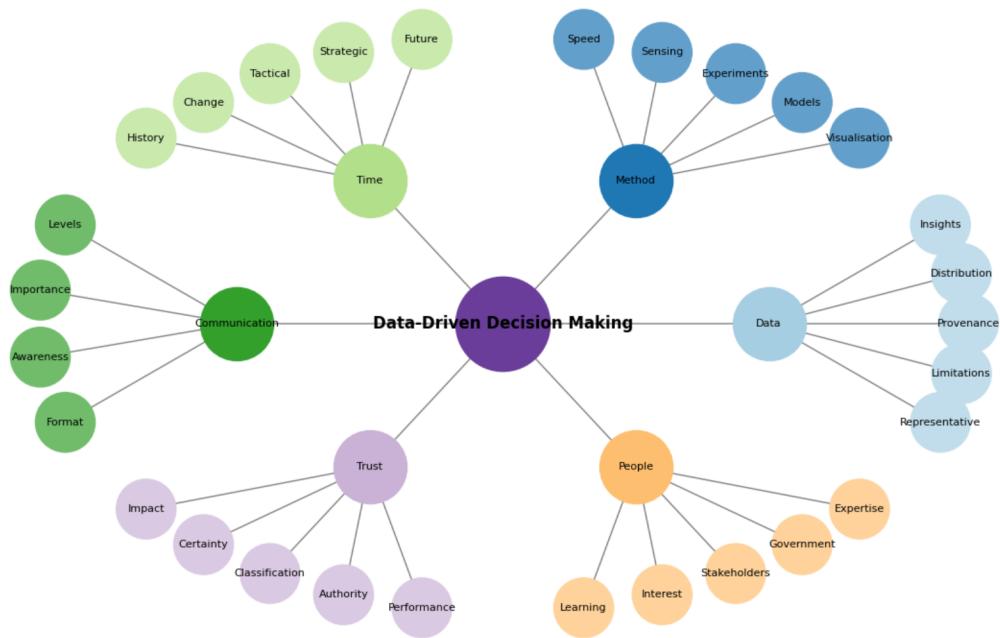


Figure 11:

Decision-maker 4: Concept Map

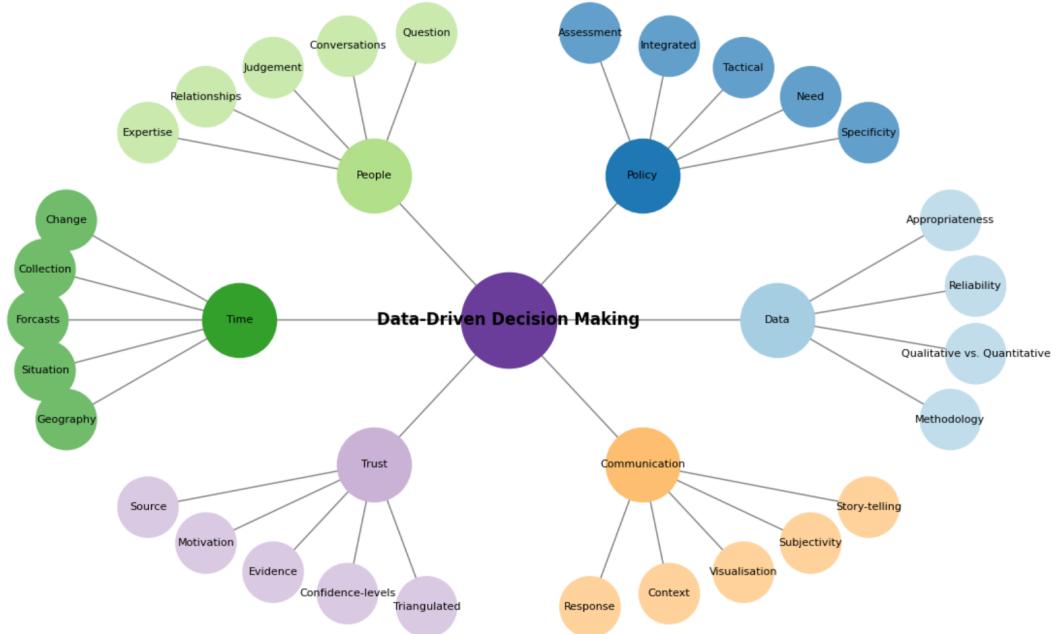


Figure 12:

Decison-maker 5: Concept Map

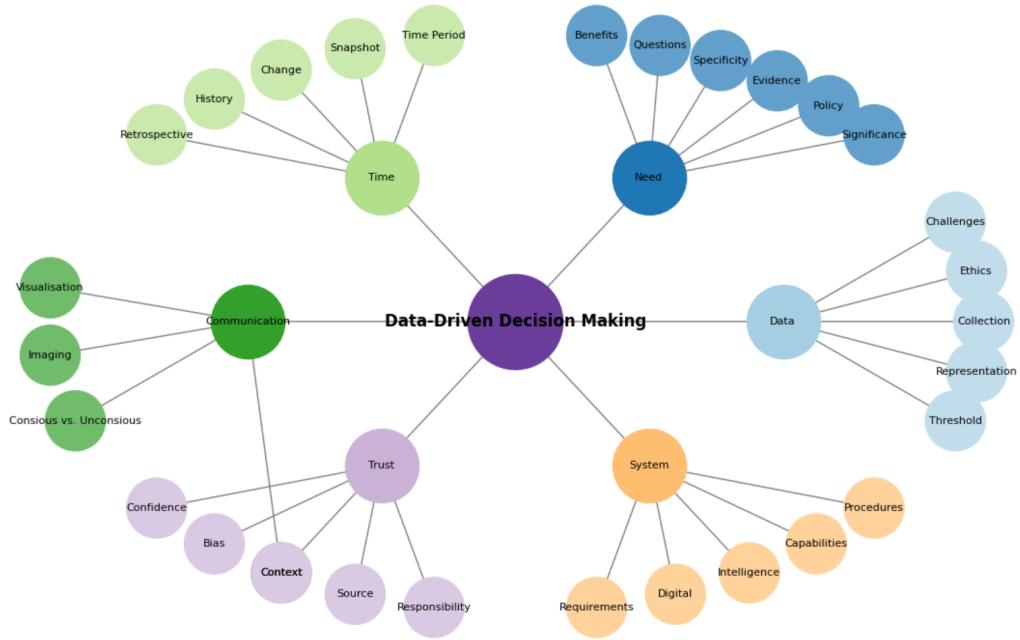


Figure 13:

Decison-maker 6: Concept Map

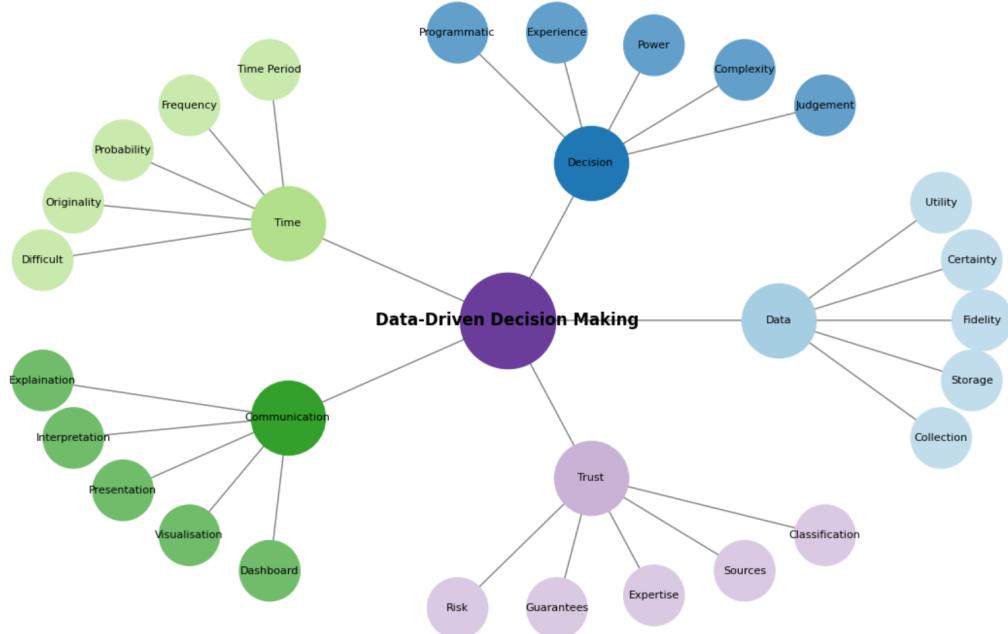


Figure 14:

Decison-maker 7: Concept Map

Figure 15:

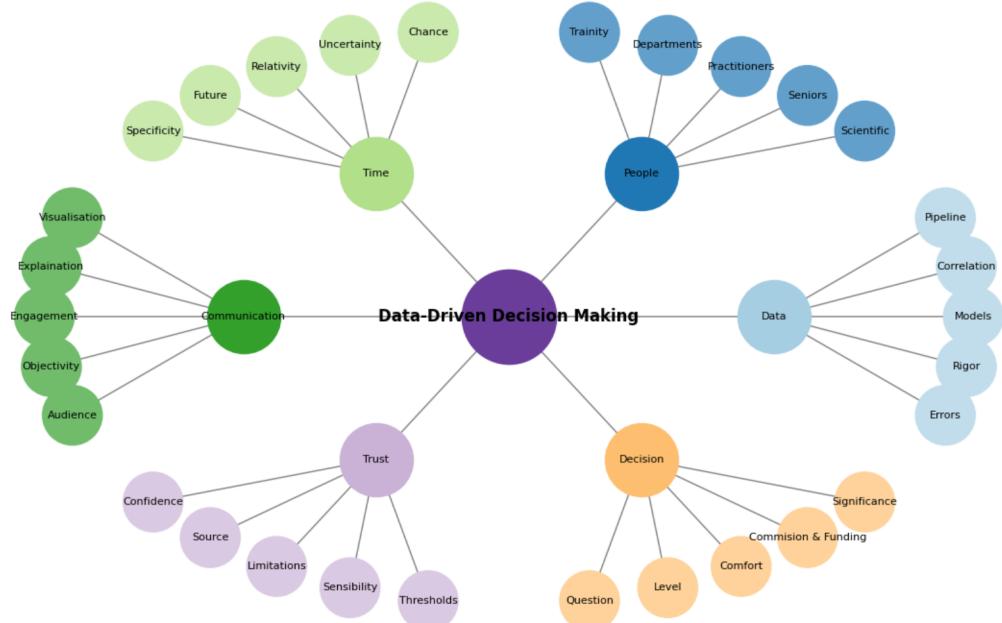
Decison-maker 8: Concept Map

Figure 16:

Decision-maker 9: Concept Map

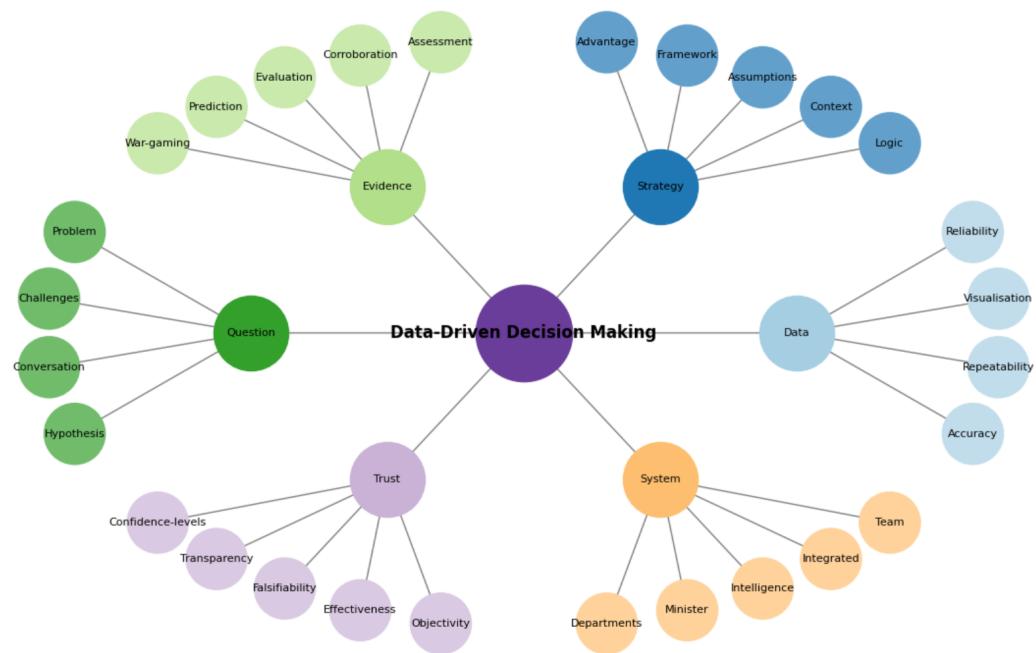


Figure 17:

Data Practitioner 1: Concept Map

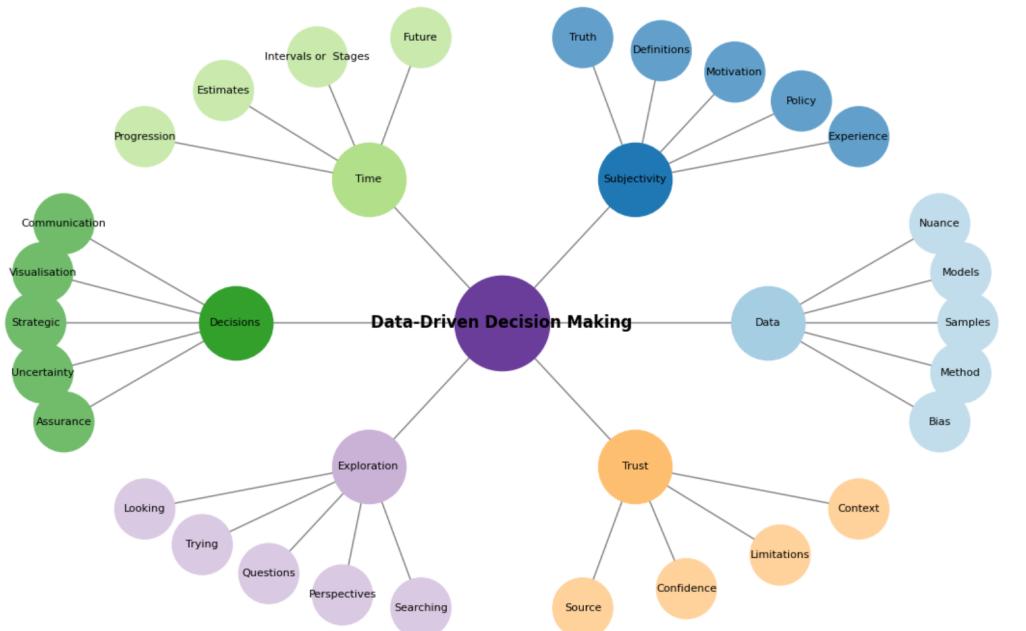


Figure 18:

Data Practitioner 2: Concept Map

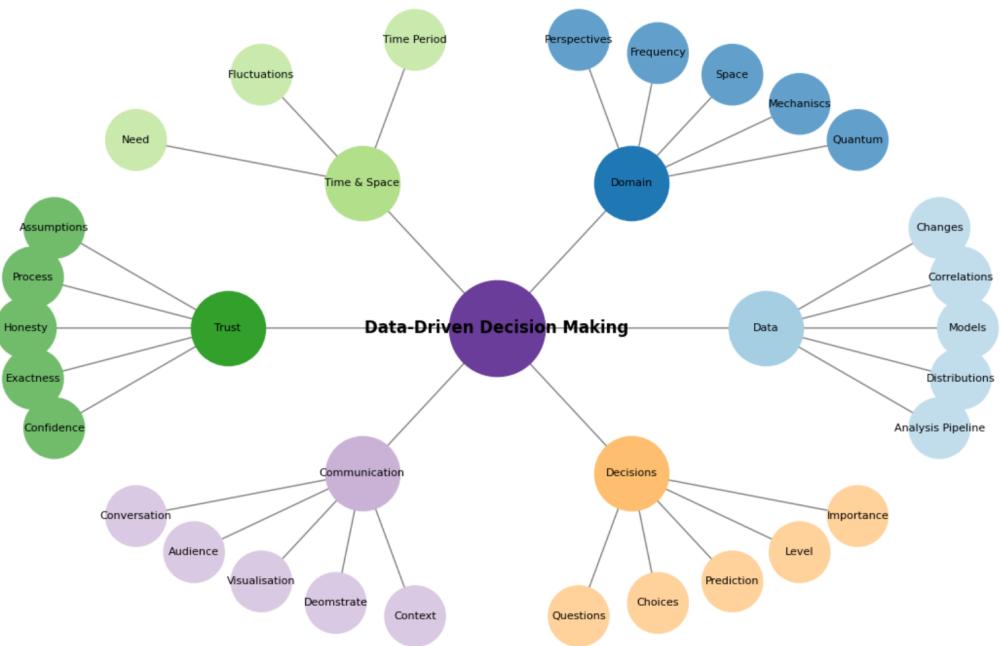


Figure 19:

Data Practitioner 3: Concept Map

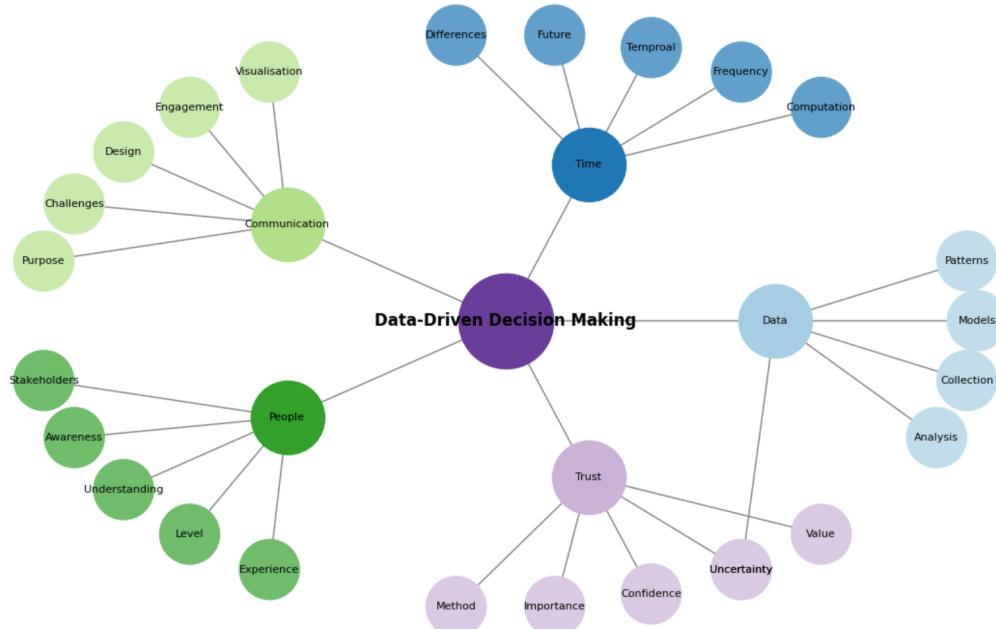


Figure 20:

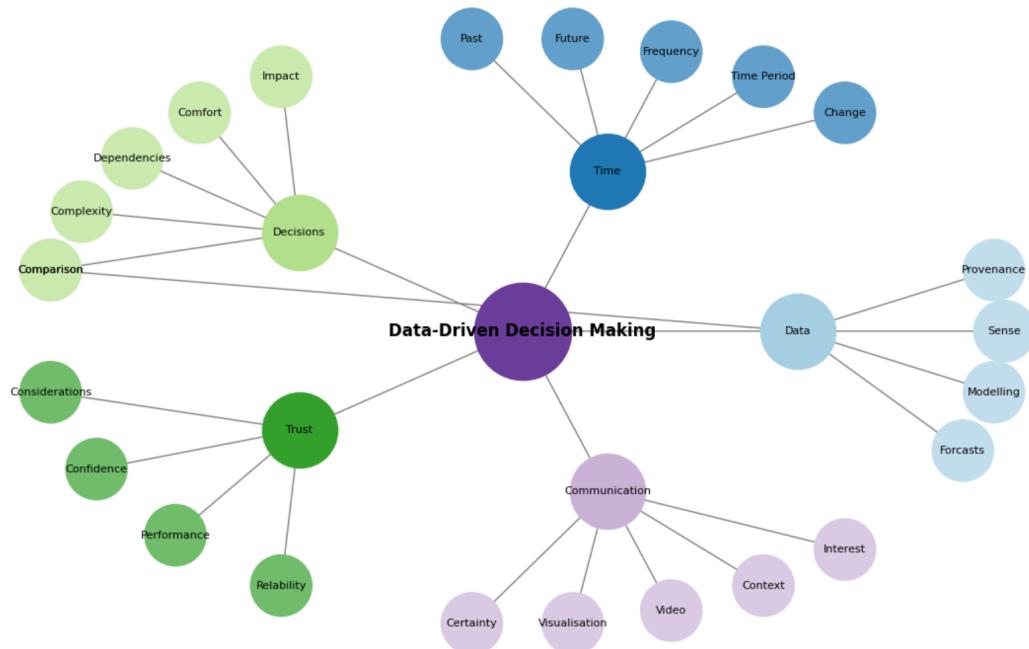
Data Practitioner 4: Concept Map

Figure 21:

Data Practitioner 5: Concept Map

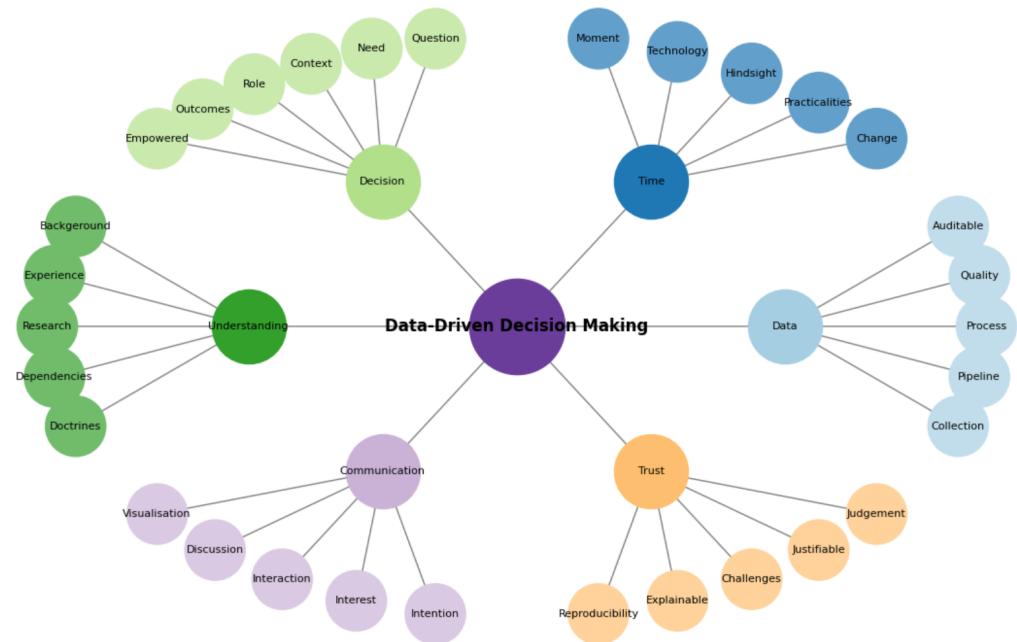
Figure 22:

Data Practitioner 6: Concept Map



Figure 23:

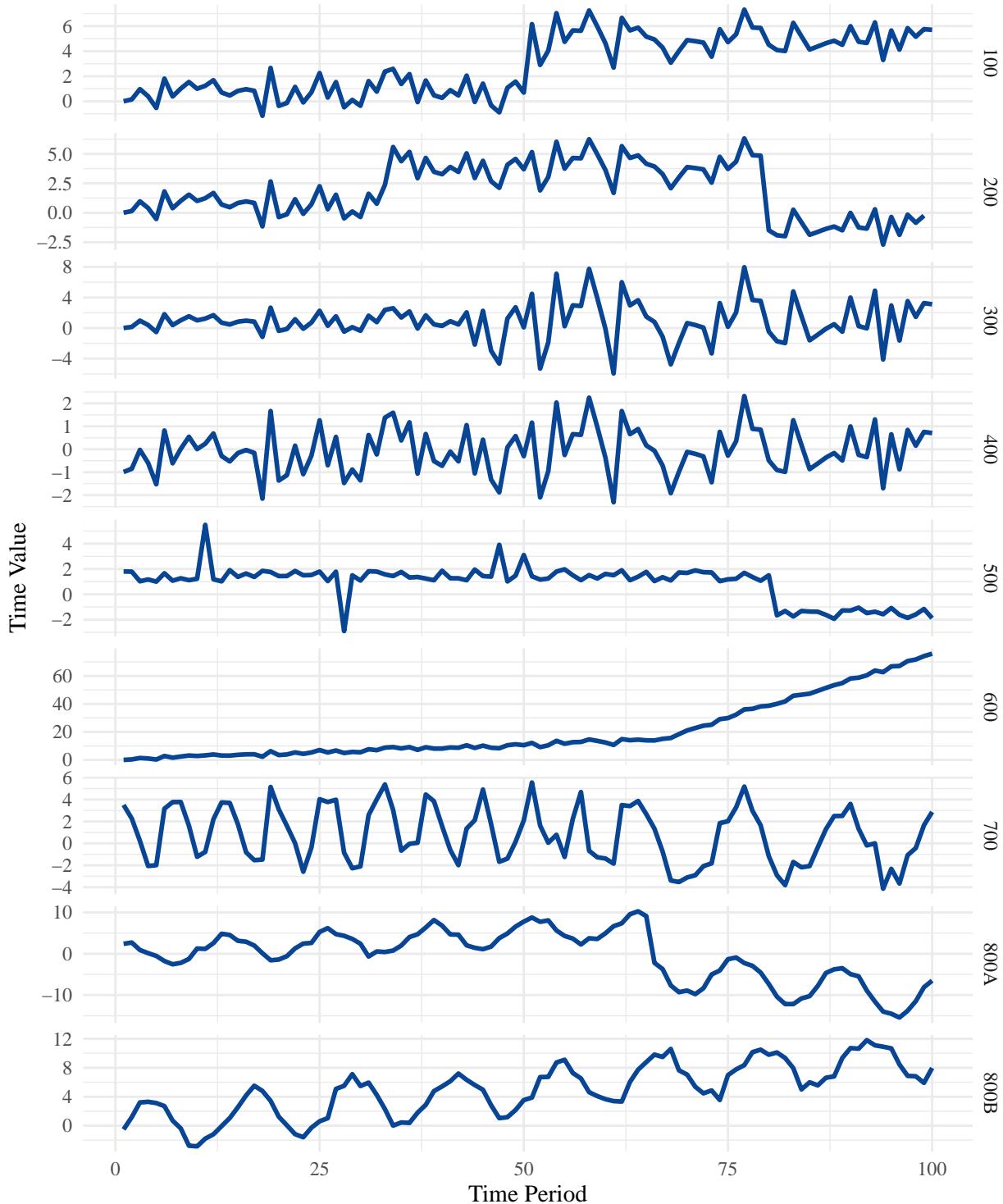
Data Practitioner 7: Concept Map



11 Annex D: Original Nine Synthetic Time Series

Each line plot in Figure 24 below visualises one of the nine original synthetic time series from the Alan Turing Institute [45].

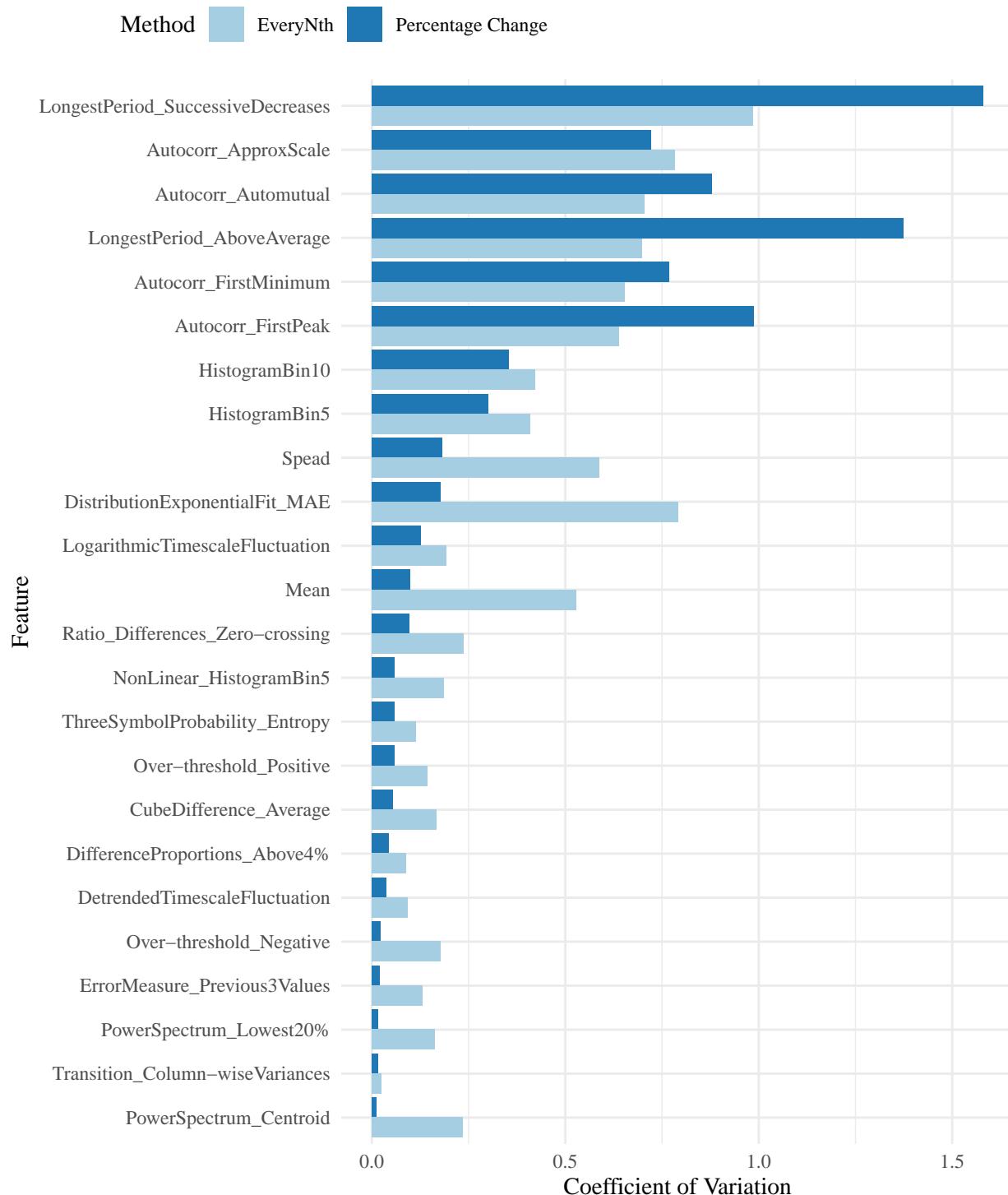
Figure 24: Annotated Change Nine Synthetic Time Series



12 Annex E: Coefficient of Variation for all Catch22 Features

Figure 25 visualises the absolute coefficient of variation values for each catch22 time series feature.

Figure 25: Coefficient of Variation for Catch22 Features



13 Annex F: The Statistical Properties of Catch22 Features

Table 3: Catch22 Feature Overview [47]

Category	Original Feature Name	Other Feature Name	Description
Distribution shape	DN_HistogramMode_5	HistogramBin5	5-bin histogram mode
Distribution shape	DN_HistogramMode_10	HistogramBin10	10-bin histogram mode
Extreme event timing	DN_OutlierInclude_p_001_mdrmd	Over-threshold_Positive	Positive outlier timing
Extreme event timing	DN_OutlierInclude_n_001_mdrmd	Over-threshold_Negative	Negative outlier timing
Linear autocorrelation	CO_flecac	Autocorr_ApproxScale	First crossing of the ACF
Linear autocorrelation	CO_FirstMin_ac	Autocorr_FirstMinimum	First minimum of the ACF
Linear autocorrelation structure	SP_Summaries_welch_rect_area_5_1	PowerSpectrum_Lowest20%	Power in lowest 20% frequencies
Linear autocorrelation	SP_Summaries_welch_rect_centroid	PowerSpectrum_Centroid	Centroid frequency
Linear autocorrelation	PD_PeriodicityWang_th001	Autocorr_FirstPeak	Wang's periodicity metric
Simple forecasting	FC_LocalSimple_mean3_stderr	ErrorMeasure_Previous3Values	Error of 3-point rolling mean forecast
Incremental differences	FC_LocalSimple_mean1_tauresrat	Ratio_Differences_Zero-crossing	Change in autocorrelation timescale after incremental differencing
Incremental differences	MD_hrv_classic_pnn40	DifferenceProportions_Above4%	Proportion of high incremental changes in the series
Symbolic	SB_BinaryStats_mean_longstretch1	LongestPeriod_AboveAverage	Longest stretch of above-mean values
Symbolic	SB_BinaryStats_diff_longstretch0	LongestPeriod_SuccessiveDecreases	Longest stretch of decreasing values

Table 4: Catch22 Feature Overview Continued [47]

Symbolic	SB_MotifThree_quantile_hh	ThreeSymbolProbability_Entropy	Entropy of successive pairs in symbolized series
Symbolic	SB_TransitionMatrix_3ac_sumdiagcov	Transition_Column-wiseVariances	Transition matrix column variance
Non-linear autocorrelation	CO_HistogramAMI_even_2_5	NonLinear_HistogramBin5	Histogram-based automutual information (lag 2, 5 bins)
Non-linear autocorrelation	CO_trev_1_num	CubeDifference_Average	Time reversibility
Non-linear autocorrelation	IN_AutoMutualInfoStats_40_gaussian_fmmi	Autocorr_Automutual	First minimum of the AMI function
Self-affine scaling	SC_FluctAnal_2_rsrangefit_50_1_logi_prop_r1	LogarithmicTimescaleFluctuation	Rescaled range fluctuation analysis (low-scale scaling)
Self-affine scaling	SC_FluctAnal_2_dfa_50_1_2_logi_prop_r1	DetrendedTimescaleFluctuation	Detrended fluctuation analysis (low-scale scaling)
Other	CO_EMBED2_Dist_tau_d_expfit_meandiff	DistributionExponentialFit_MAE	Goodness of exponential fit to embedding distance distribution
Other	DN_Mean	Mean	Mean
Other	DN_Spread_Std	Spread	Standard Deviation

It is interesting to note that the seven most sensitive `catch22` features are in three categories: `symbolic`, `linear autocorrelation`, and `other`. Further investigation to understand why would be beneficial.

14 Annex G: Variation of Catch22 Features across Time Series Types

Figure 26 presents feature with the highest coefficient of variation across the 900 time series is the longest sequence of successive steps that decrease (`LongestPeriod_Decreases` or `SB_BinaryStats_diff_longstretch0`). This catch22 feature “calculates the longest sequence of successive steps in the time series that decrease” [48]. Initial visual analysis suggests that time series ‘100,’ ‘300,’ and ‘500’ are most affected by this feature.

Figure 26: Variation of the Longest Sequence of Successive Steps that Decrease

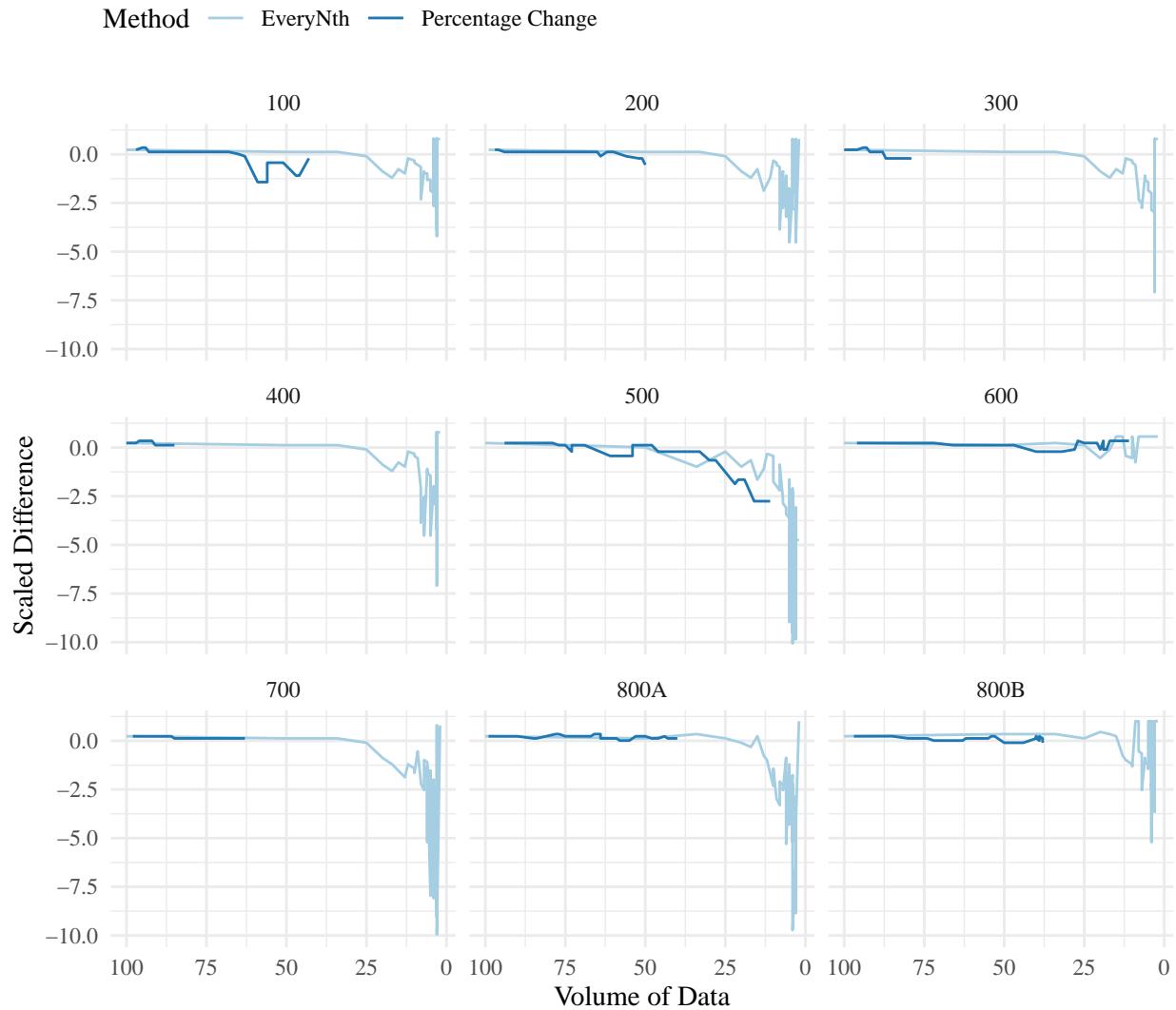


Figure 27 presents feature with the second highest coefficient of variation is the longest sequence of successive steps that decrease (`Autocorr_FirstMinimum` or `C0_FirstMin_ac`). This catch22 feature represents “*calculates the longest sequence of successive steps in the time series that decrease*” [48]. Initial visual analysis suggests that time series ‘100,’ ‘200’ and ‘500’ are more affected by this feature.

Figure 27: Variation of the First Minimum of the Autocorrelation Function

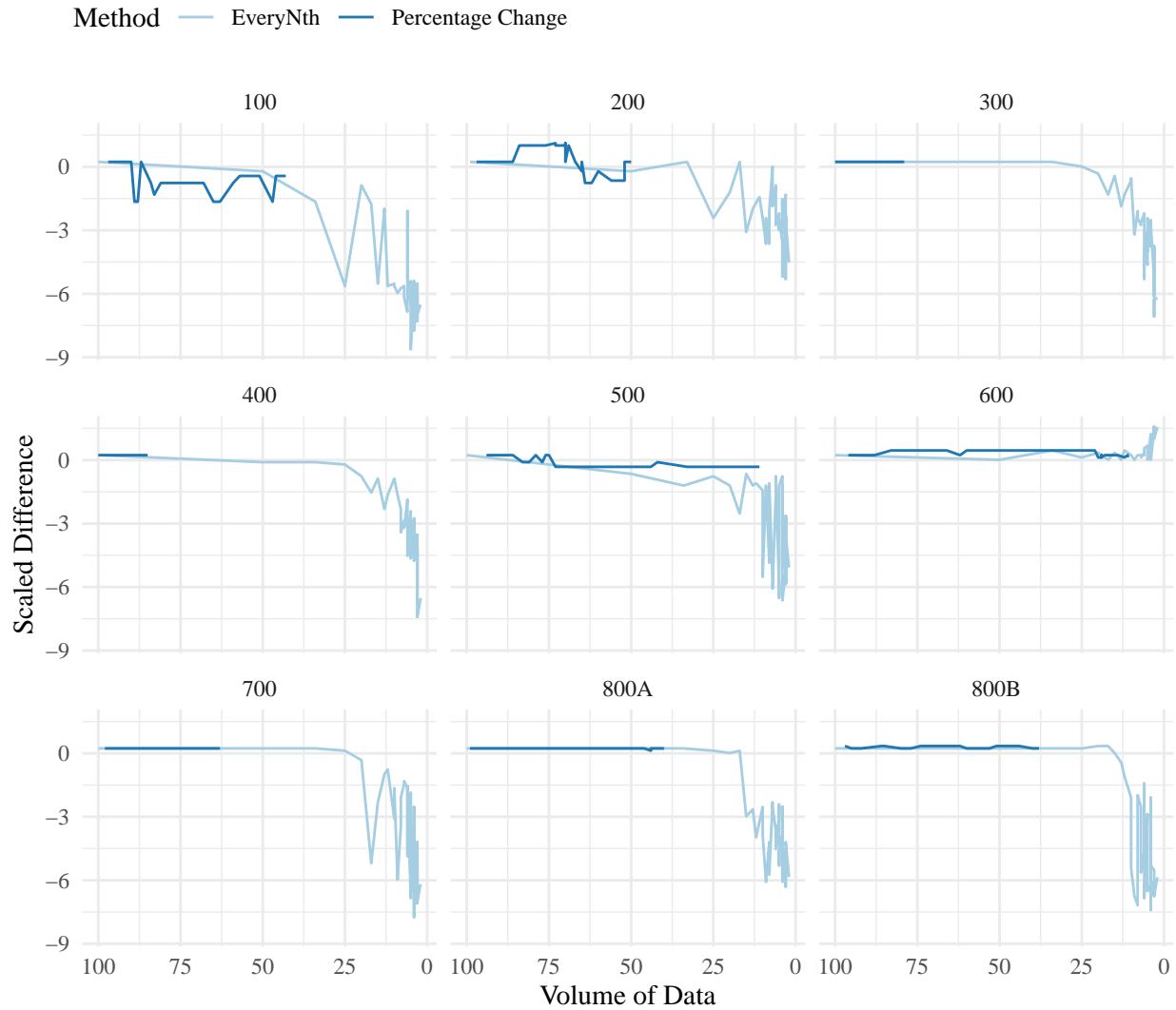


Figure 28 presents the feature with the third highest coefficient of variation is the minimum of the automutual information function (`Autocorr_Automutual` or `IN_AutoMutualInfoStats_40_gaussian_fmmi`). This catch22 feature outputs a measure of “*autocorrelation in the time series, as the minimum of the automutual information function*” [48]. Initial visual analysis suggests that time series ‘100,’ ‘200,’ ‘500,’ and ‘600’ are more affected by this feature.

Figure 28: Variation of the Minimum of the Automutual Information Function

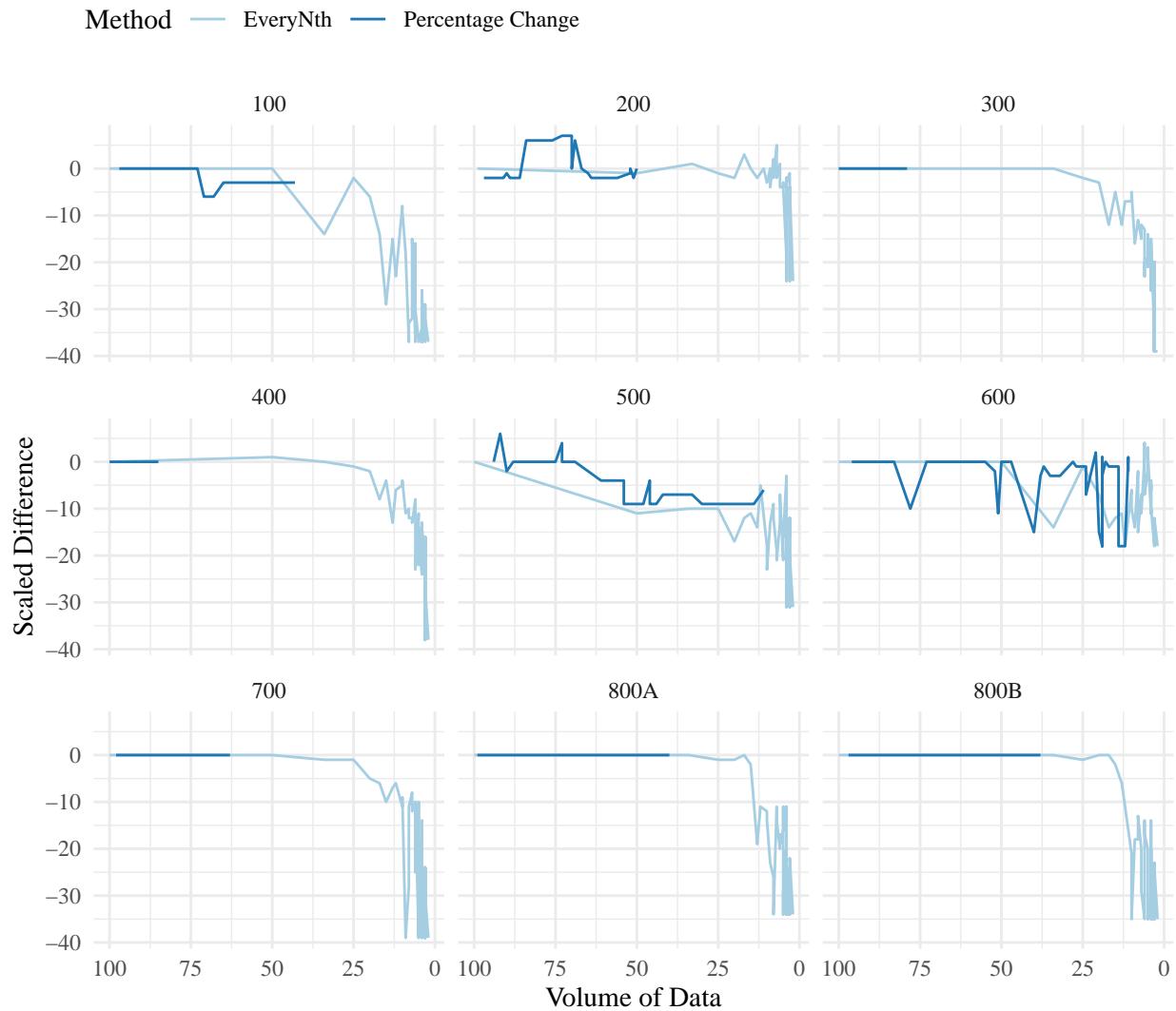


Figure 29 presents the feature with the fourth highest coefficient of variation is the longest sequence of successive values greater than the mean (`LongestPeriod_AboveAverage` or `SB_BinaryStats_mean_longstretch1`). This catch22 feature “calculates the longest successive period of above average values” [48]. Initial visual analysis suggests that time series ‘200,’ ‘500,’ and ‘800B’ are more affected by this feature.

Figure 29: Variation of the Longest Sequence of Successive Values Greater than the Mean

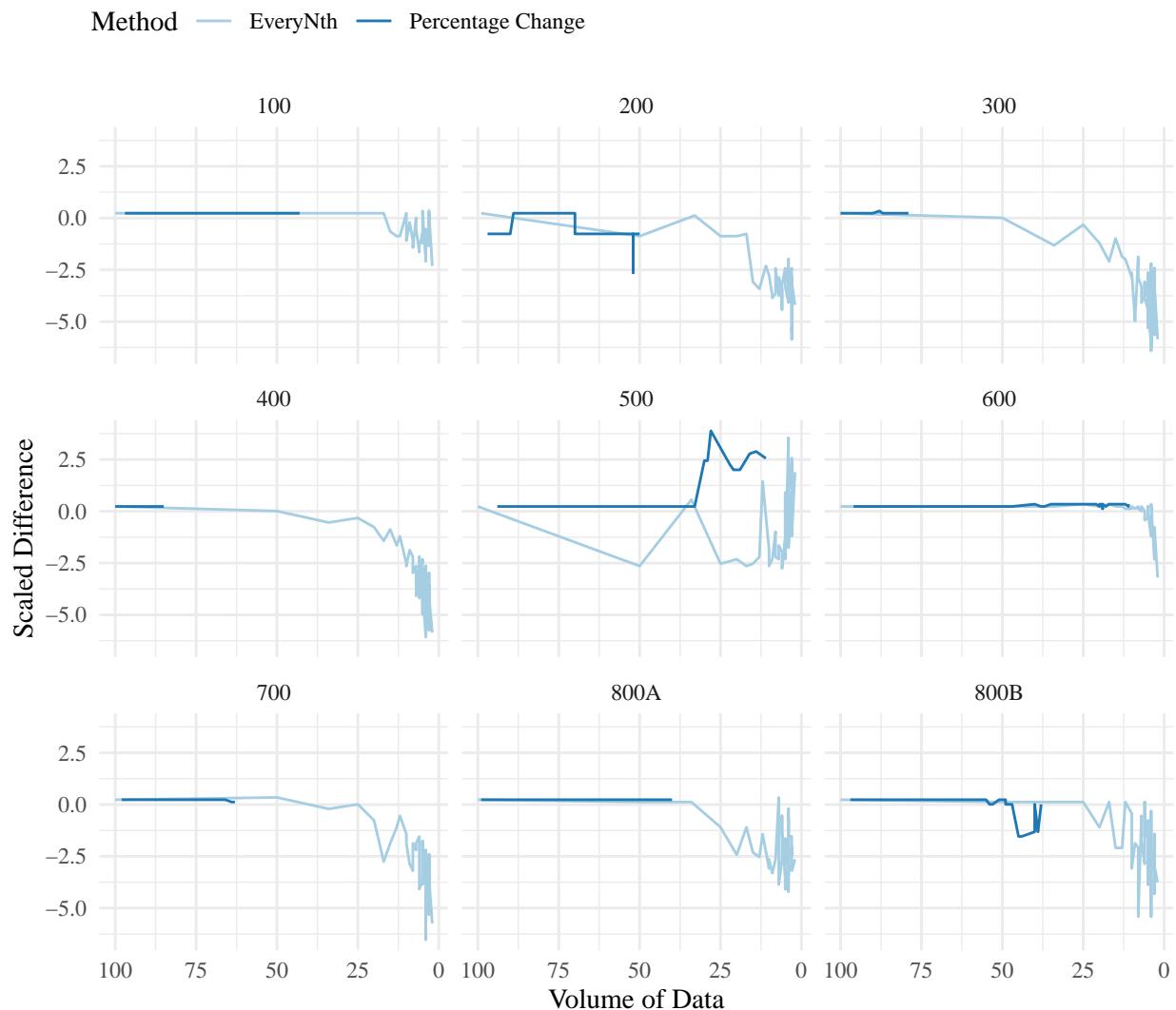


Figure 30 presents the feature with the fifth highest coefficient of variation is the approximate scale of autocorrelation (`Autocorr_ApproxScale` or `C0_f1ecac`). This catch22 feature “computes the first $1/e$ crossing of the autocorrelation function of the time series” [48]. It is similar to the first minimum of the autocorrelation function. Initial visual analysis suggests that time series ‘100,’ ‘200,’ ‘500,’ ‘800A’ and ‘800B’ are more affected by this feature.

Figure 30: Variation of the Approximate Scale of Autocorrelation

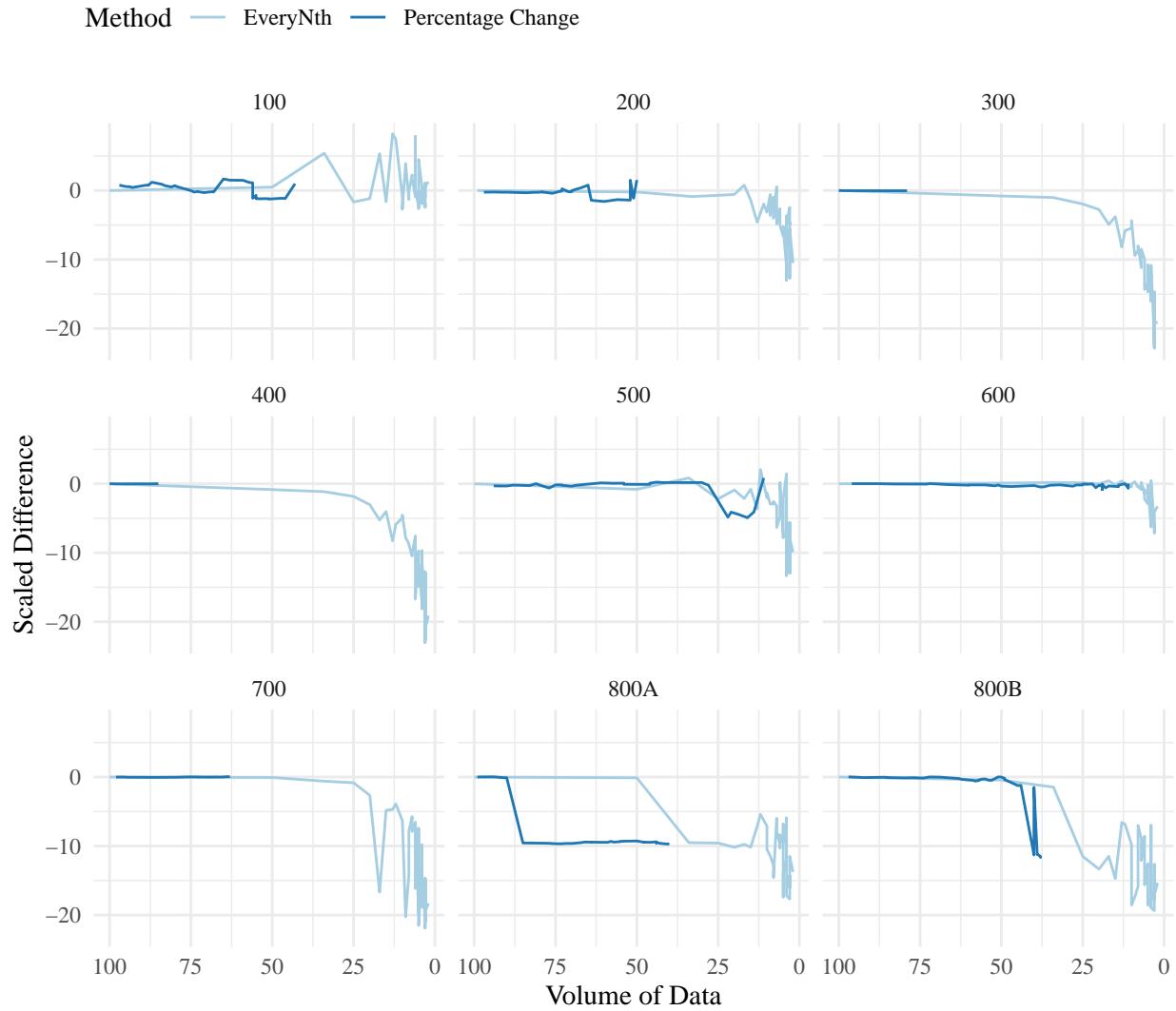


Figure 31 presents the feature with the sixth highest coefficient of variation is the first peak in the autocorrelation function (`Autocorr_FirstPeak` or `PD_PeriodicityWang_th0_01`). This catch22 feature “*returns the first peak in the autocorrelation function satisfying a set of conditions (after detrending the time series using a single-knot cubic regression spline)*” [48]. In general, the feature returns higher values for slower time series oscillation. Initial visual analysis suggests that time series ‘100,’ ‘200’ and ‘500’ are more affected by this feature.

Figure 31: Variation of the First Peak in the Autocorrelation Function

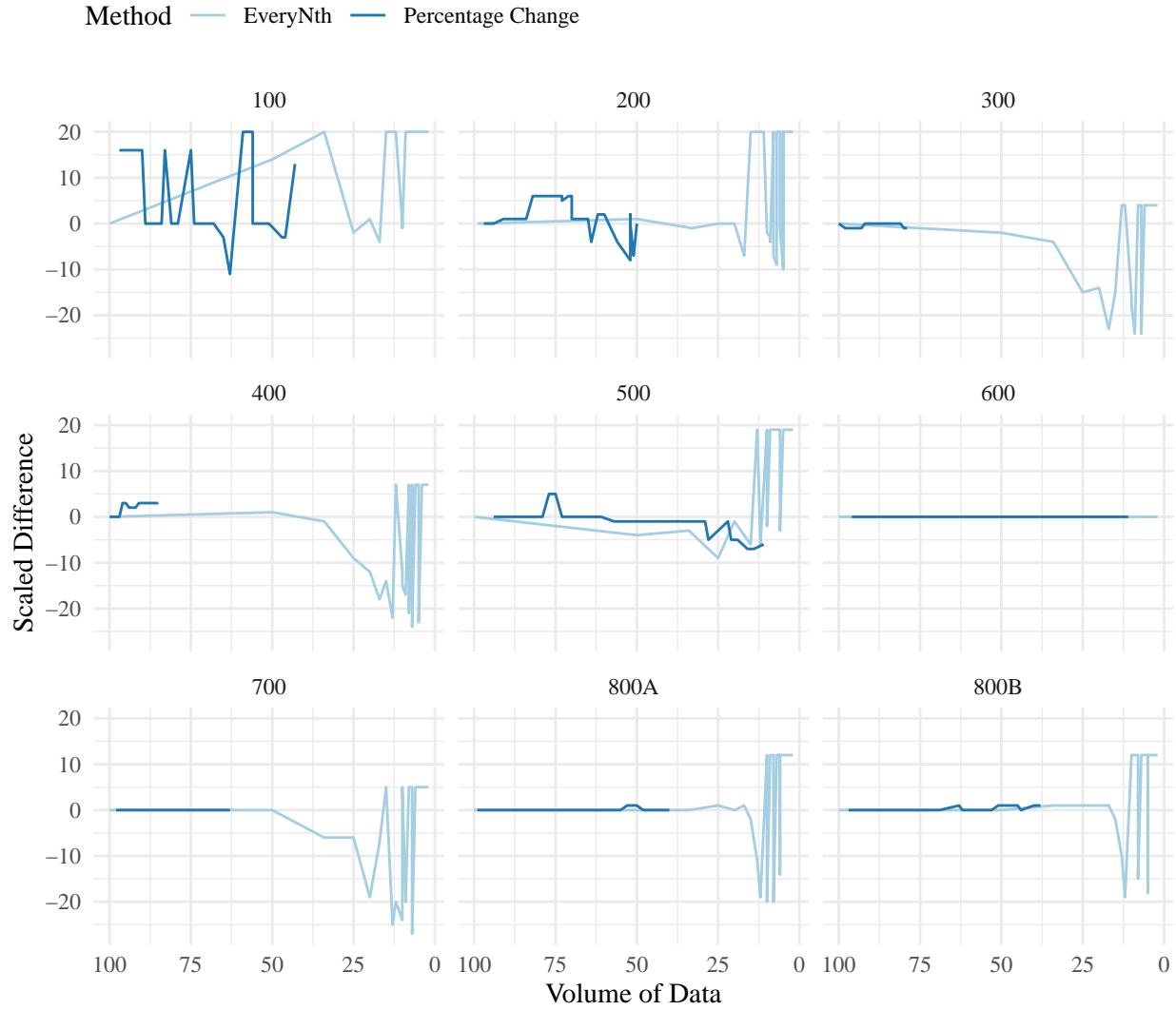
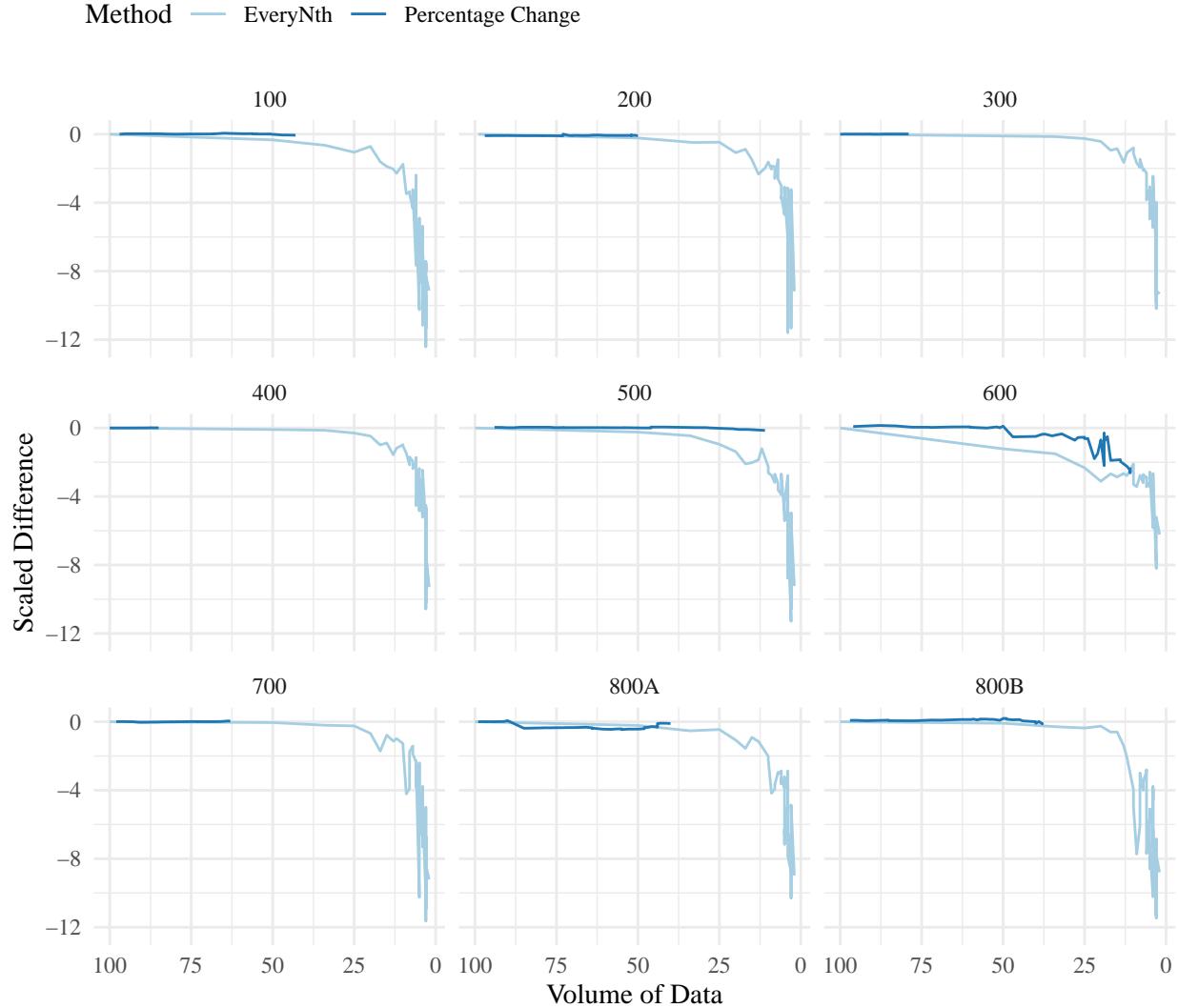


Figure 32 presents the feature with the seventh highest coefficient of variation is the mean absolute error of an exponential fit to a probability distribution (`DistributionExponentialFit_MAE` or `C0_EMBED2_DIST_TAU_D_EXPFIT_MEANDIFF`). To calculate the mean absolute error, this `catch22` feature “represents the time series in a two-dimensional time-delay embedding space... It then computes successive distances between points in this 2D embedding space and analyzes the probability distribution of these distances” [48]. Initial visual analysis suggests that time series ‘600’ and ‘800’ are more affected by this feature.

Figure 32: Variation of the Mean Absolute Error of an Exponential Fit



15 Annex H: Most Sensitive Catch22 Features for each Time Series Type

Figure 33 visualises how the six most sensitive features for time series 100. Initial visual analysis suggests that `Autocorr_FirstMinimum`, `Autocorr_FirstPeak` and `LongestPeriod_Decreases` may best indicate the impact of downsampling for this time series.

Figure 33: Variation of Catch22 Features for Time Series 100

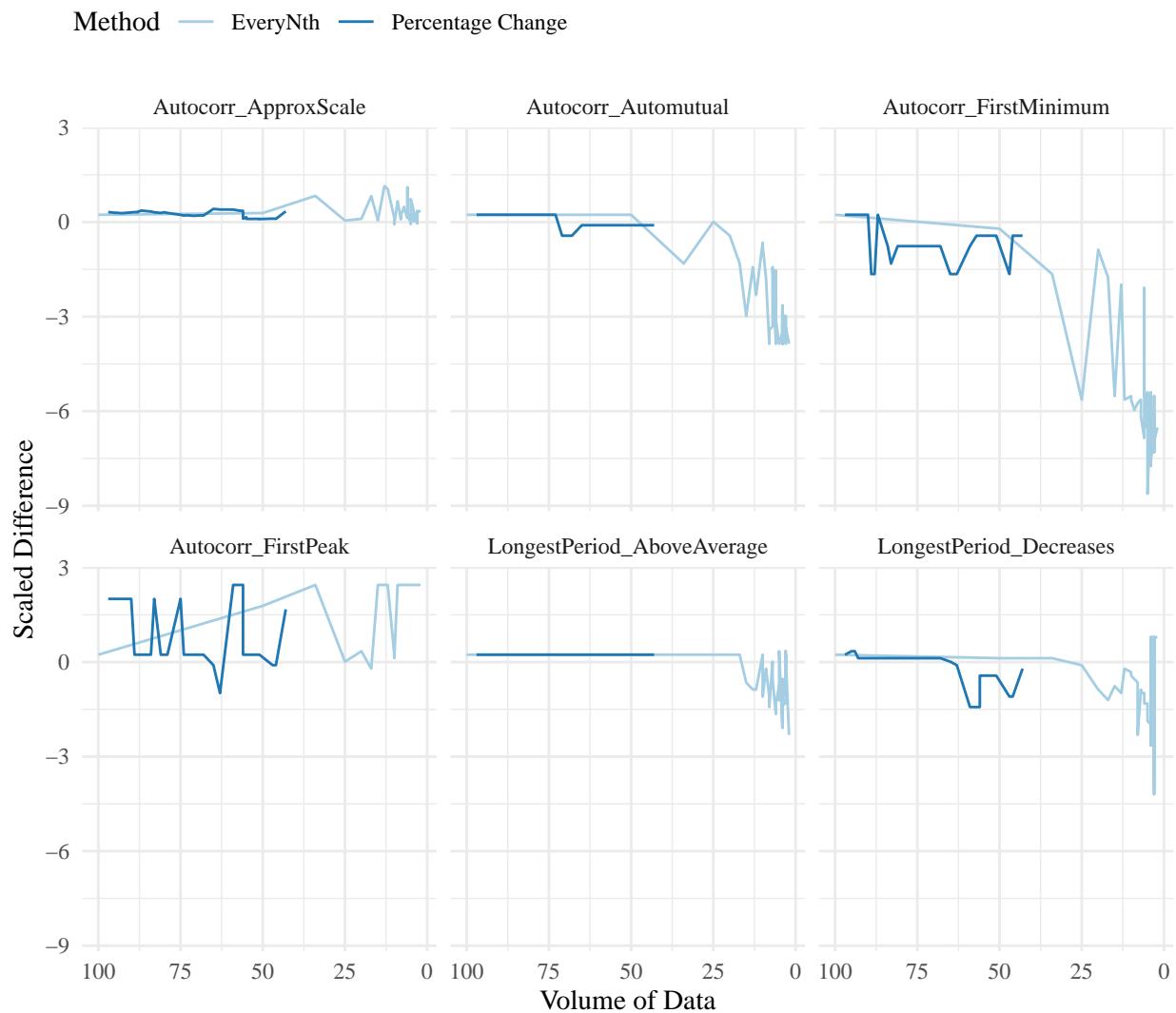


Figure 34 visualises how the six most sensitive features for time series 200. Autocorr_Automutual, Autocorr_FirstMinimum, Autocorr_FirstPeak and LongestPeriod_AboveAverage may best indicate the impact of downsampling for this time series.

Figure 34: Variation of Catch22 Features for Time Series 200

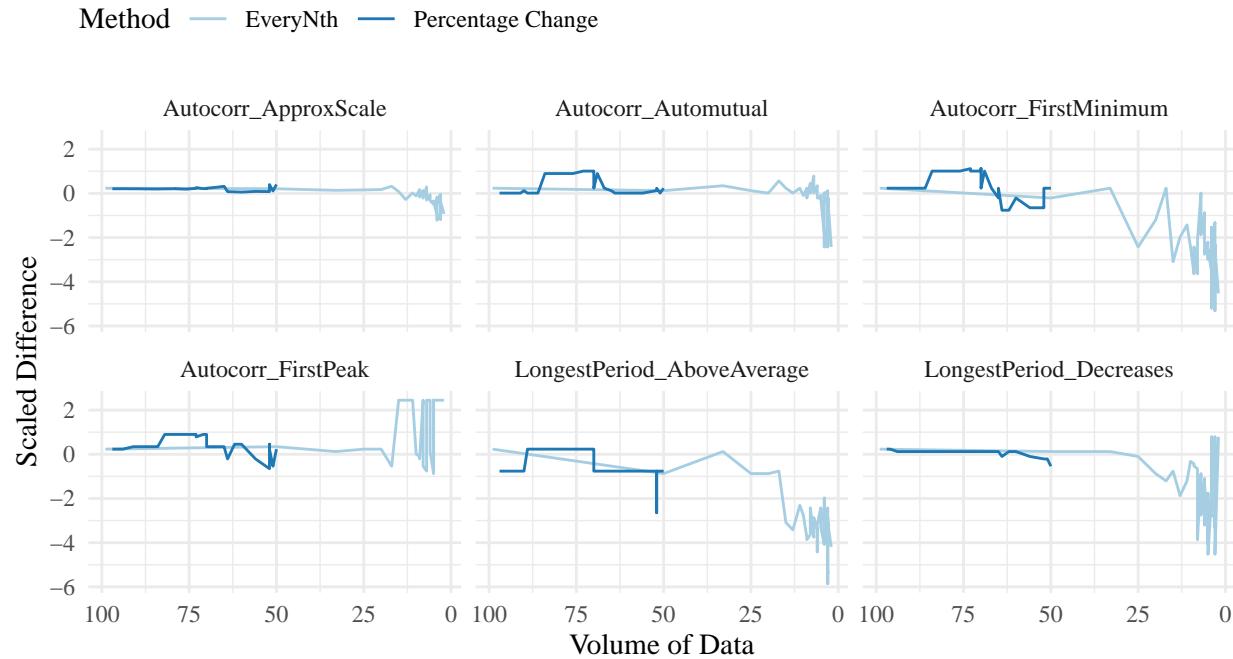


Figure 35 visualises how the six most sensitive features for time series 300. Autocorr_FirstPeak, LongestPeriod_AboveAverage and LongestPeriod_Decreases may best indicate the impact of downsampling for this time series.

Figure 35: Variation of Catch22 Features for Time Series 300

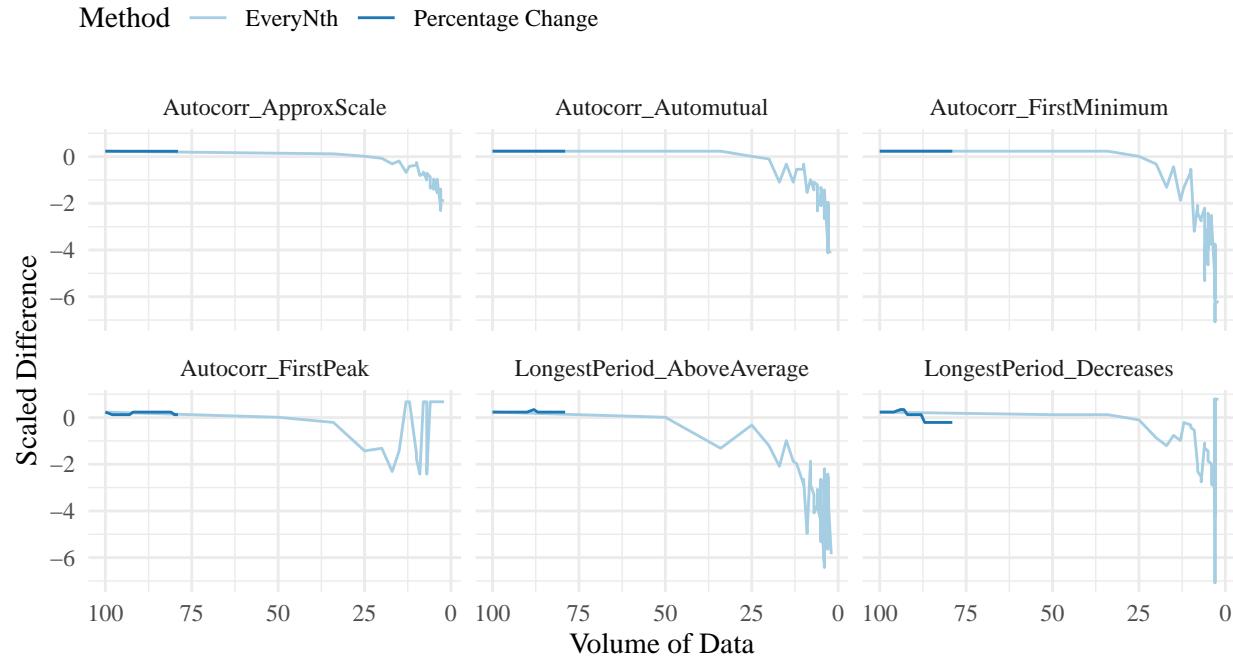


Figure 36 visualises how the six most sensitive features for time series 400. `Autocorr_FirstPeak` and `LongestPeriod_Decreases` may best indicate the impact of downsampling for this time series.

Figure 36: Variation of Catch22 Features for Time Series 400

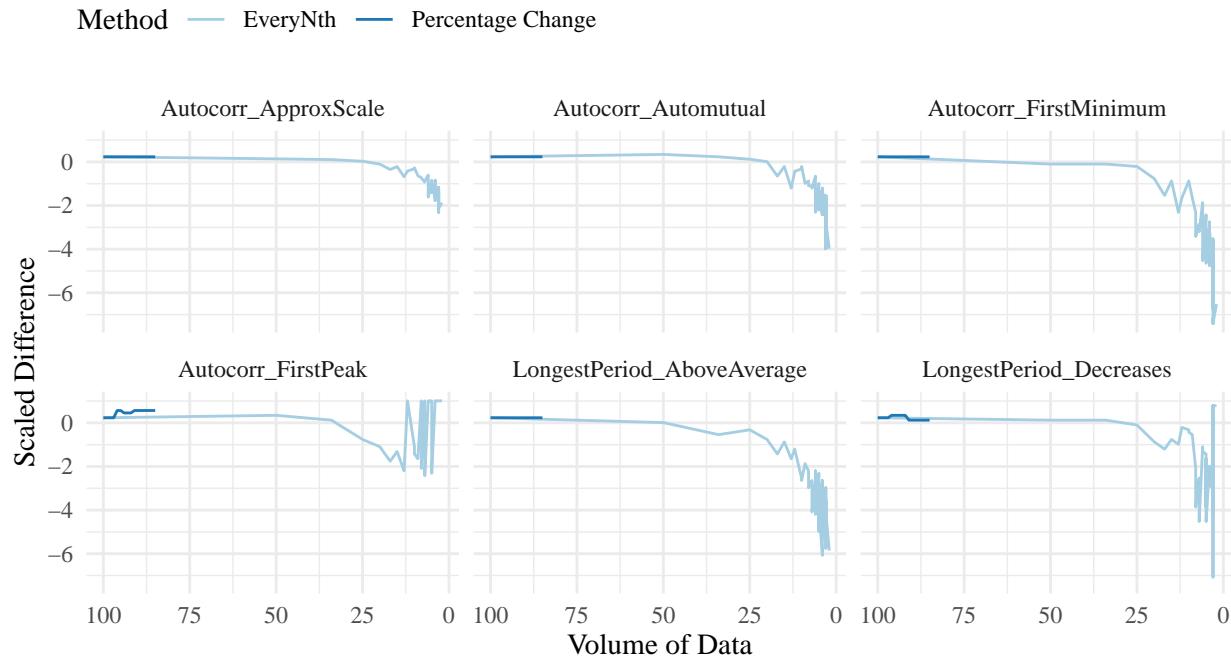


Figure 37 visualises how the six most sensitive features for time series 500. `Autocorr_Automutual`, `Autocorr_FirstMinimum`, and `Autocorr_FirstPeak` may best indicate the impact of downsampling for this time series.

Figure 37: Variation of Catch22 Features for Time Series 500

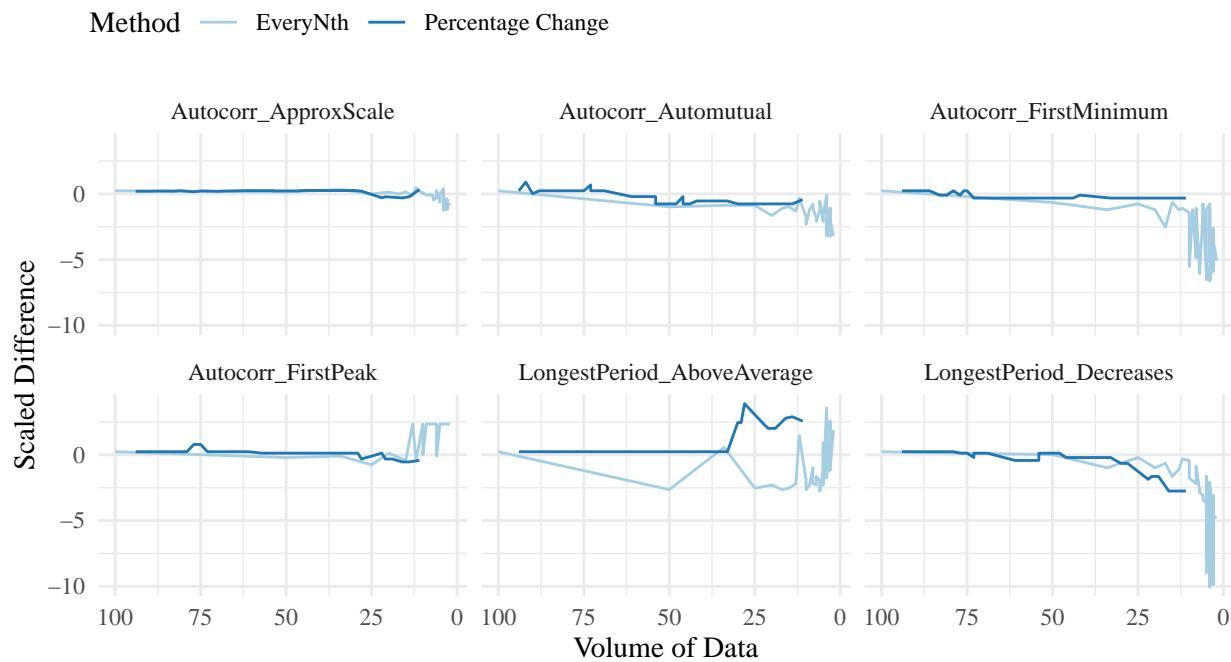


Figure 38 visualises how the six most sensitive features for time series 600. `Autocorr_Automutual` and `Autocorr_FirstMinimum` may best indicate the impact of downsampling for this time series.

Figure 38: Variation of Catch22 Features for Time Series 600

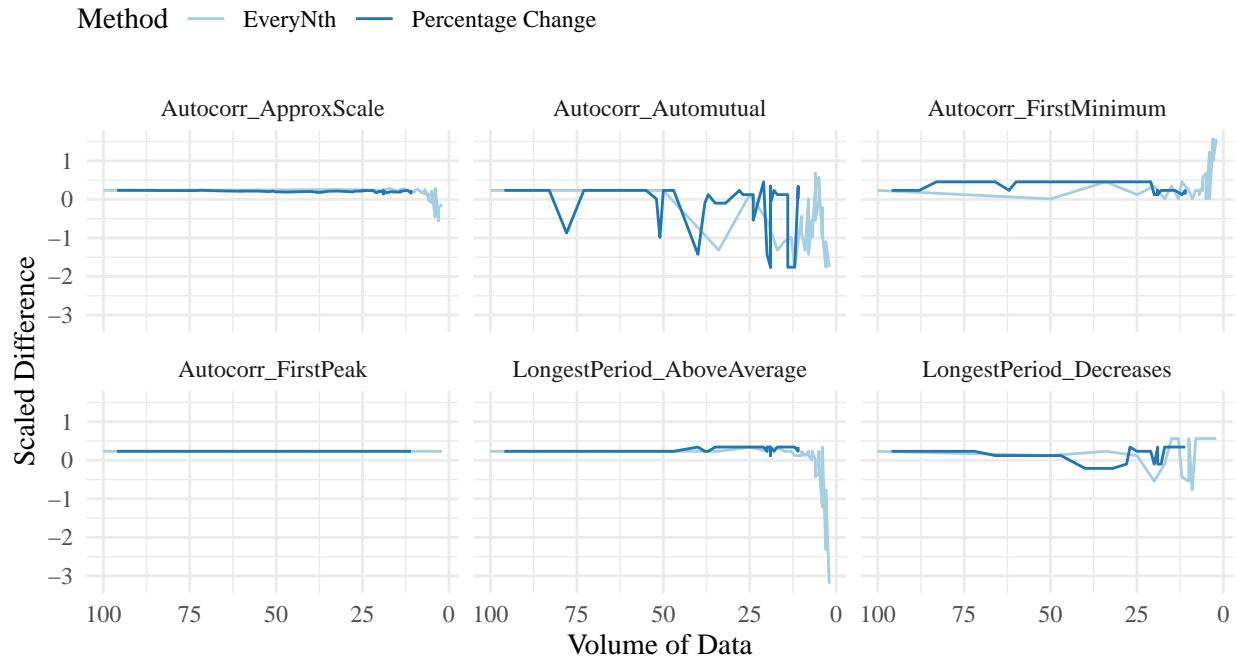


Figure 39 visualises how the six most sensitive features for time series 700. `LongestPeriod_AboveAverage` and `LongestPeriod_Decreases` may best indicate the impact of downsampling for this time series.

Figure 39: Variation of Catch22 Features for Time Series 700

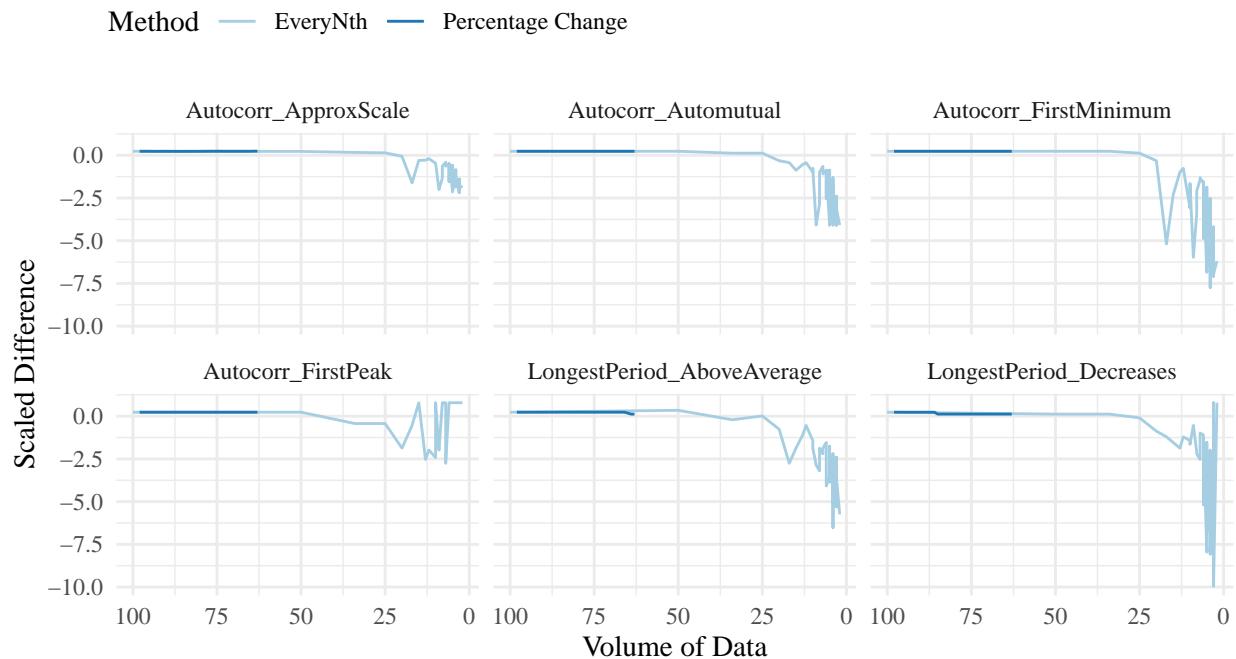


Figure 40 visualises how the six most sensitive features for time series 800A. `Autocorr_ApproxScale` and `LongestPeriod_Decreases` may best indicate the impact of downsampling for this time series.

Figure 40: Variation of Catch22 Features for Time Series 800A

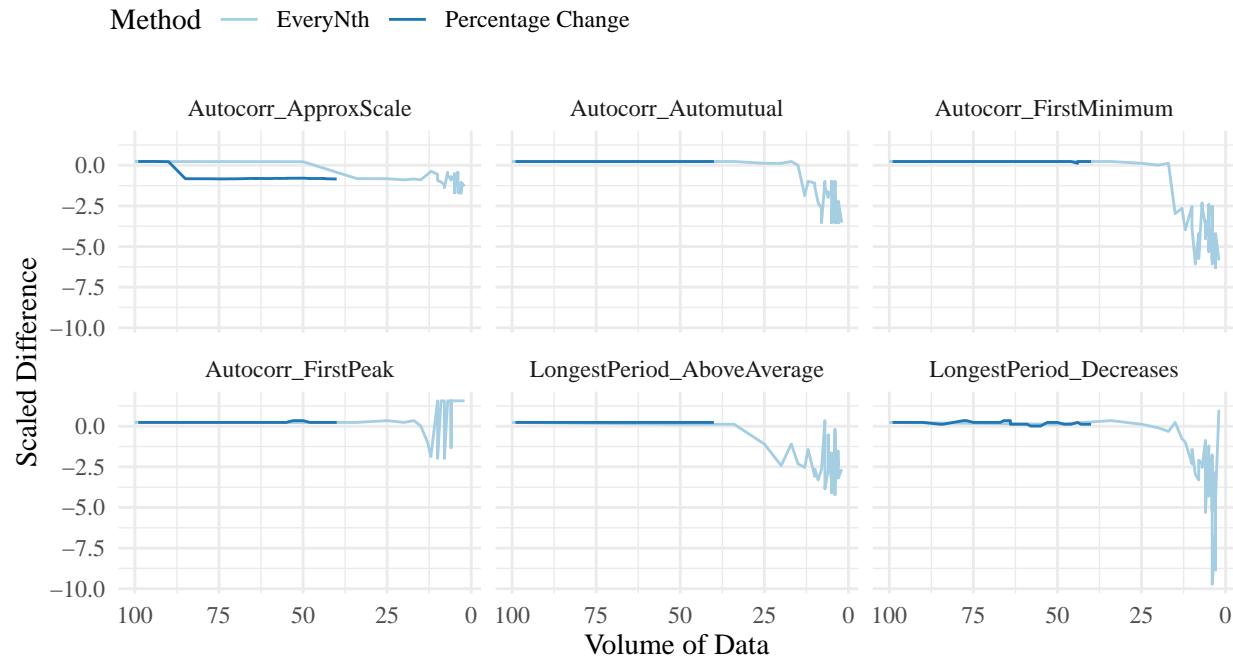
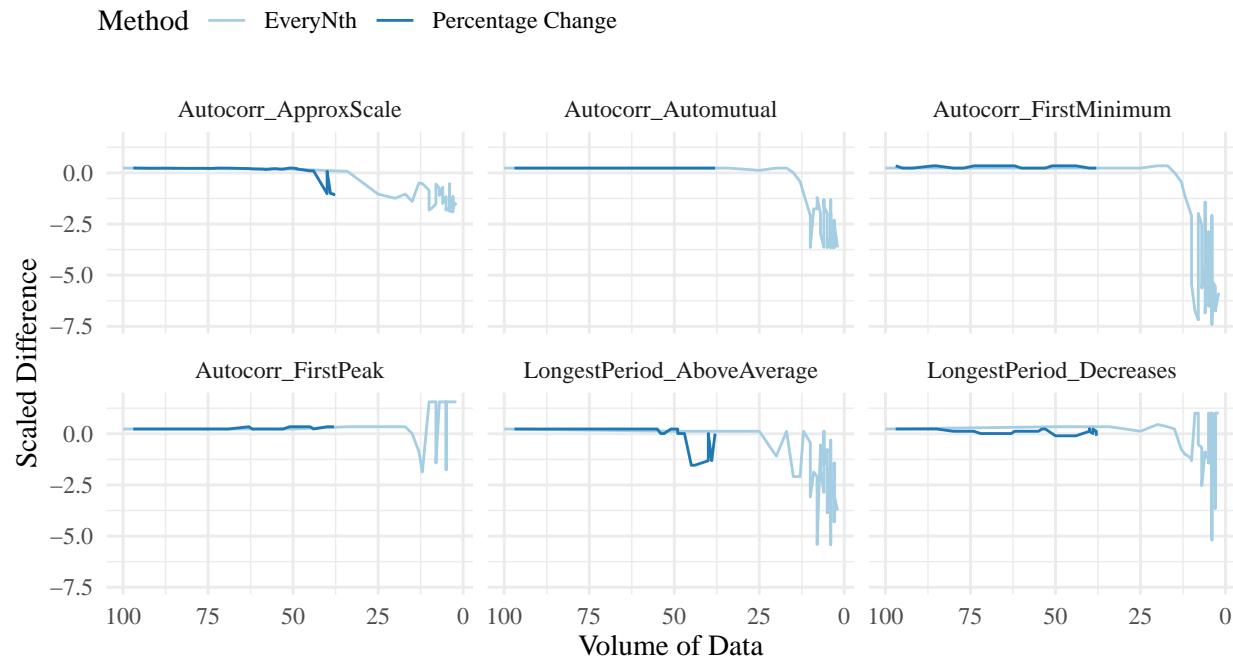


Figure 41 visualises how the six most sensitive features for time series 800A. `Autocorr_ApproxScale`, `Autocorr_FirstMinimum`, `LongestPeriod_AboveAverage` and `LongestPeriod_Decreses` may best indicate the impact of downsampling for this time series.

Figure 41: Variation of Catch22 Features for Time Series 800B



REFERENCES

- [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] 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.
- [7] J. Donckt, J. Donckt, M. Rademaker, and S. Hoecke, “MinMaxLTTB: Leveraging MinMax-preselection to scale LTTB.” <https://arxiv.org/abs/2305.00332>, 2023.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] The Shift Project, “Implementing digital sufficiency,” 2020.
- [13] 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.
- [14] 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.
- [15] B. Bechtold, “Beyond the barrel: How data and analytics will become the new currency in oil and gas.” <https://gblogs.cisco.com/ca/2018/06/07/beyond-the-barrel-how-data-and-analytics-will-become-the-new-currency-in-oil-and-gas/>, 2018.
- [16] D. Vaughan, *The challenger launch decision: Risky technology, culture, and deviance at NASA*. Chicago; London: The University of Chicago Press, 1996.
- [17] S. Moritz and T. Bartiz-Beielstein, “imputeTS: Time series missing value imputation in r,” vol. 9.1. <https://github.com/SteffenMoritz/imputeTS>; R Journal, 2017. doi: 10.32614/RJ-2017-009.
- [18] C. H. Lubba, B. Fulcher, T. Henderspn, B. Harris, O. r. TL, and O. Cliff, “catch22: CAnonical time-series CCharacteristics.” <https://github.com/DynamicsAndNeuralSystems/catch22>; R Journal, 2022. doi: 10.5281/zenodo.6673597.

- [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.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] J. Hao, Q. Shi, Y. Ye, and W. Zeng, "TimeTuner: Diagnosing time representations for time-series forecasting with counterfactual explanations." 2023. doi: 10.48550/arXiv.2307.09916.
- [26] 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.
- [27] S. Moritz and T. Bartiz-Beielstein, "imputeTS: Time series missing value imputation in r." <https://cran.r-project.org/web/packages/imputeTS/vignettes/imputeTS-Time-Series-Missing-Value-Imputation-in-R.pdf>; R Journal, 2017.
- [28] S. van Buuren, *Flexible imputation of missing data*. /url- [29] 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)*, <https://www.usenix.org/conference/atc20/presentation/visheratin>; USENIX Association, 2020, pp. 589–601.
- [30] 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.
- [31] HM Government, "Global britain in a competitive age: The integrated review of security, defence, development and foreign policy." <https://www.gov.uk/government/publications/global-britain-in-a-competitive-age/the-integrated-review-of-security-defence-development-and-foreign-policy>; GOV.UK, 2021.
- [32] 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.
- [33] A. Kohn, D. Moritz, and T. Neumann, "DashQL – complete analysis workflows with SQL." 2023. doi: 10.48550/arXiv.2306.03714.
- [34] 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.
- [35] J. Pettersson and P.-O. Gutman, "Automatic tuning of the window size in the box car back slope data compression algorithm," *Journal of Process Control*, vol. 14, no. 4, pp. 431–439, 2004, doi: 10.1016/j.jprocont.2003.07.006.
- [36] 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. doi: 10.1145/3589290.
- [37] 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.
- [38] 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.

- [39] 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.
- [40] 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.
- [41] 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.
- [42] 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.
- [43] jonodrew, "Commits." [urlhttps://github.com/MoFrod/downsampling_timeseries/commits/main](https://github.com/MoFrod/downsampling_timeseries/commits/main), 2023.
- [44] OpenAI, "ChatGPT 4: CodeInterpreter," vol. "August 3". 2023.
- [45] G. J. J. van den Burg and C. K. I. Williams, "An evaluation of change point detection algorithms." <https://github.com/alan-turing-institute/AnnotateChange>, 2022. doi: 10.48550/arXiv.2003.06222.
- [46] M. Forshaw, "Jettison MVP code." <https://github.com/MattForshaw/Jettison/tree/main/MVPCode>, 2023.
- [47] M. J. Watson, A. Liakopoulos, D. Brzakovic, and C. Georgakis, "A practical assessment of process data compression techniques," *Industrial & Engineering Chemistry Research*, vol. 37, no. 1, pp. 267–274, 1998, doi: 10.1021/ie970401w.
- [48] C. H. Lubba, B. Fulcher, T. Henderspn, B. Harris, O. r. TL, and O. Cliff, "catch22 features." <https://feature-based-time-series-analys.gitbook.io/catch22-features/>, 2023.
- [49] Y. Wu, N. Cao, D. Archambault, Q. Shen, H. Qu, and W. Cui, "Evaluation of graph sampling: A visualization perspective," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 401–410, 2017, doi: 10.1109/TVCG.2016.2598867.