The Explainability of Downsampling

Morgan Frodsham & Matthew Forshaw
School of Computing, Newcastle University

Introduction

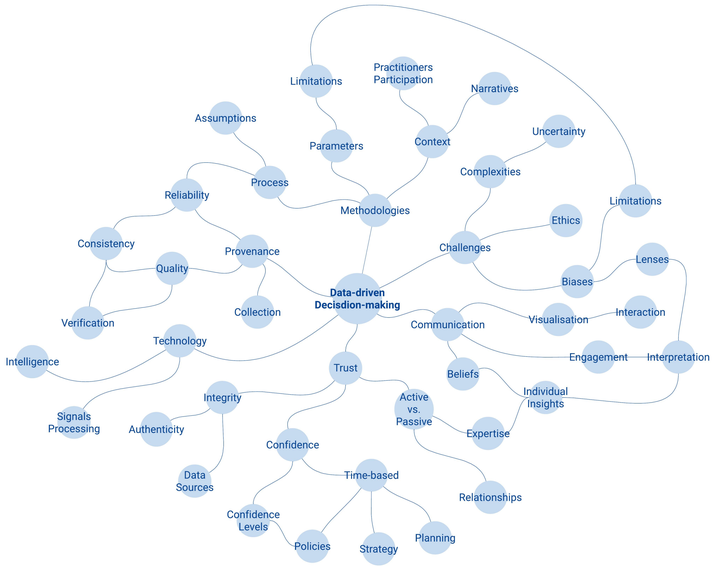
HM Government is committed to making data-driven decisions that engender public trust[1]. Achieving this requires transparency of how data is used in government decision-making[2].

This research aims to:

1. Highlight that government decision-makers need to trust how the data is used.
2. Share how transparency can be created in the data processing pipeline, focused on downsampling.

User Research

Sixteen officials were interviewed; nine decision-makers and seven data practitioners. Themes are shared below:

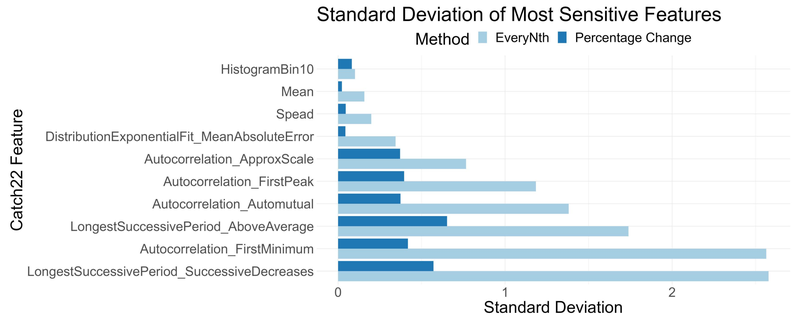


"We should put a lot more emphasis on metadata – how to appropriately describe the level of uncertainty and provenance of the data."

"Your trust changes to more being trust in the source of the information and how that information was generated."

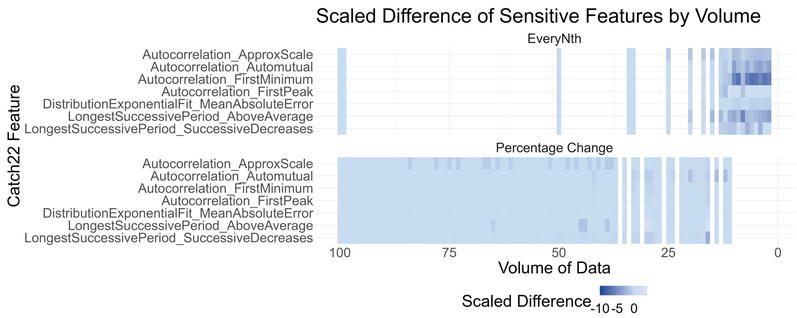
Downsampling Sensitivity

The time series features[3] most sensitive to downsampling were identified by calculating the standard deviation of difference (scaled) between the original and processed data.



Downsampling Impact

The EveryNth and Percentage Change downsampling algorithms were applied to the most sensitive features.



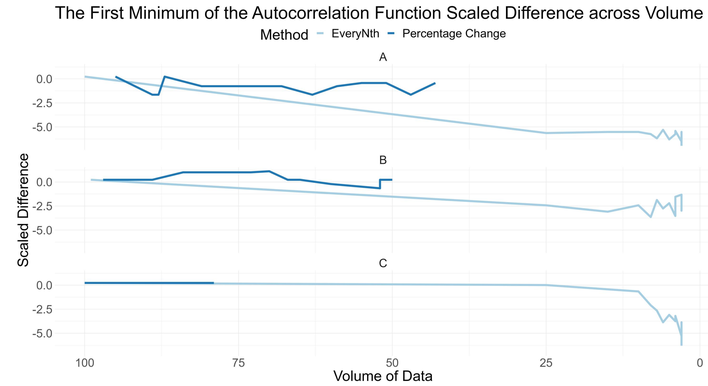
The impact of downsampling changes as more data is discarded; this is acute for EveryNth and subtle, but observably inconsistent for Percentage Change.

References

1. Cabinet Office, “Government transformation strategy: Better use of data.” HM Government, 2017;
2. Department for Digital, Culture, Media & Sport, “National data strategy.” HM Government, 2020;
3. C. H. Lubba, B. Fulcher, T. Henderspn, B. Harris, O. r. TL, and O. Cliff, “catch22: CAnonical time-series CHaracteristics.” R Journal, 2022.

Feature Variation

The impact of the downsampling algorithms on the most sensitive features varies across different time series.



This difference suggests that it is possible to select the features that best preserve each time series.

Discussion

1. User research suggests that transparency of how data is used, and its limitations, engenders trust for decision-makers and data practitioners.
2. This research identifies seven time series features that appear to be especially sensitive to the impacts of downsampling.
3. These features offer a new approach for identifying and communicating the downsampling algorithms and parameters that preserve representativeness with the lowest demand on computing resources.

Future work applying this research to alternative time series is needed to further test the approach of this research. Conducting usability research with decision-makers and data practitioners will also help refine these insights.

M.C.M.FRODSHAM2@newcastle.ac.uk