How creating a pipeline for automatically analysing and sharing paradata facilitated the ability to take data-driven adjustments to improve data collections.

Paradata has proven key to building and managing dynamic data collections as we can use the paradata to identify and measure sample bias and measurement error. However, paradata contains large amounts of data. Extracting and cleaning paradata can be time consuming and use a lot of server capacity, therefore we need a robust and quick process for extracting, cleaning, storing, and using the data. Using only open-source programming languages, Python and R, with Google Cloud Storage services, we have created a pipeline that daily synchronises paradata from Blaise, cleans it and stores it in Google Cloud Buckets ready to be analysed to gain insightful information about the data collection. For instance, using Audit Trail Data, we can gain information about the best time to contact different demographic groups. Likewise, using Dial History Data, we can understand if certain groups should be higher prioritised in the CATI-dashboard Daybatch settings to adjust for sample biases. To make the results available to all stakeholders, irrespective of coding experience, the paradata analysis results are shared on an internal website that is updated daily. By creating an automatic pipeline that allows users to evaluate the data collection process without coding, we have made paradata more available to colleagues and stakeholders. Thus, we have made it easier to take data-driven decisions to adjust for bias and measurement error. As our tools, R and Python, are open-source, aspects of our pipeline can be implemented without any cost. Similarly, we hope that sharing the journey of how we created the pipeline and the benefits we saw from creating this pipeline can be useful for other Blaise users.

By E. Kilicdogan & E. Alstad

Statistics Norway