ASHE Synthetic Data Pilot Update

Ian Banda, Iain Dove

Table of Contents

Presentation Date : 23 June 2023 | Need help? Contact the team.

What did we do?

ONS ran a synthetic data pilot from 6th February - 24th March using synthetic data created from ASHE 2020 dataset.

Accredited researchers were given access through the UK Data Service.

We had researchers from the institutions listed below;

- Muslim Council of Britain (MCB)
- ADR Wales
- University of Warwick
- Bank of England

- NISRA
- University of Edinburgh

User feedback received through this pilot will help the ONS to maximise benefits of data access and further explore the feasibility of synthesising more data in future.

Governance/ Data Access

Data owner approval was needed to begin the synthesis of the data and share externally

Potential users were reached out through newsletters and webinars - all were already Accredited Researchers

Once volunteers had registered their interest an intro session was held along with a user guide sent out

Following this, user agreements were signed

Users applied for access to the data via the UK Data Service

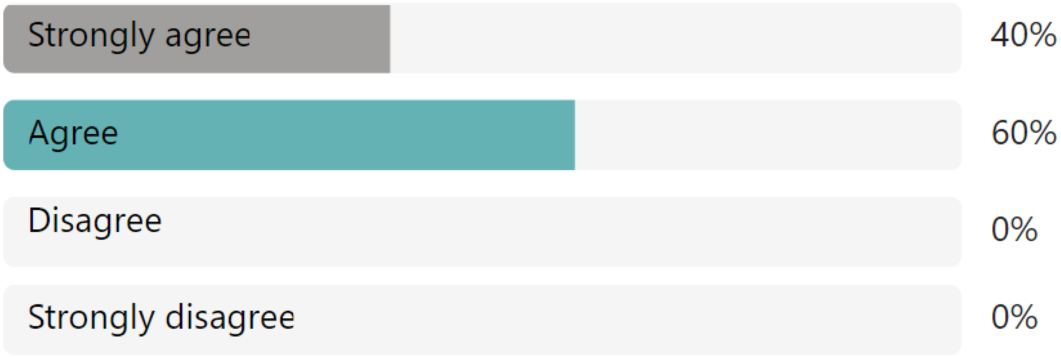
A survey was sent out after the pilot for feedback

What did we ask the participants to test / explore?

- How much work can be done in the synthetic before getting access to the real data?
- Ease of access
- Thoughts on the general utility of the data

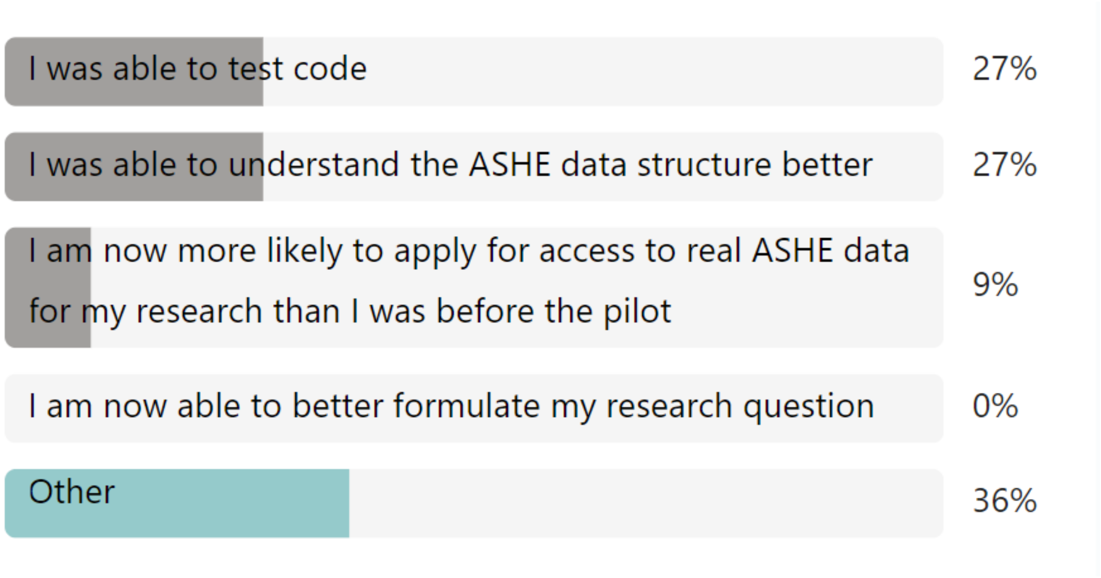
Results

I found access to the synthetic data useful



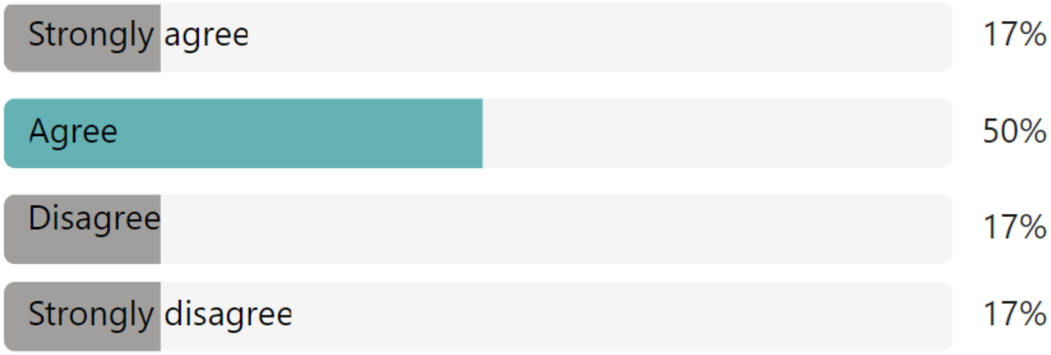
All participants agreed access to the synthetic data was useful.

Key uses



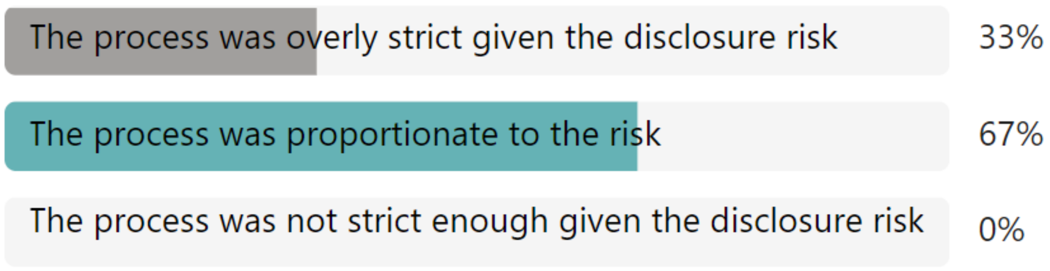
It was useful for several reasons including testing code, having a better understanding to the structure of ASHE data and one participant said they would be more likely to apply to use real ASHE data for their research.

The process to access the synthetic data was simple

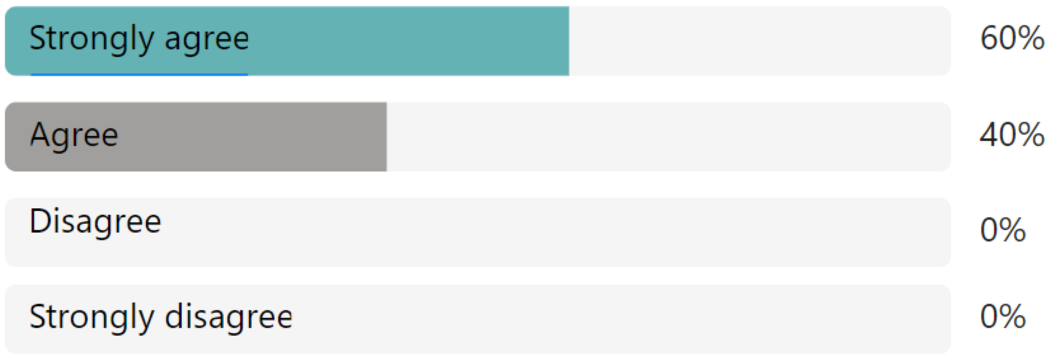


There were mixed reviews on the process to gaining access. Although four out of six participants agreed that the process to accessing the data was simple, the below chart indicates that participants had opposing thoughts on whether this process was proportionate to the level of identification risk.

The process to access the synthetic data was proportionate to the level of identification risk



I would be interested in using low fidelity synthetic versions of other SRS datasets in the future



> "I think this is a really good innovation, and will really help improve accessibility and understanding of microdata in the future".

What suggestions do you have for improving the offer of synthetic data by ONS in future?

> "It was helpful to have 2 years of data included in the dataset, but a longer time series would help even more, as we are often interested in time series work with the data".
>
> "In future iterations potentially, lookup tables could be included alongside the synthetic data so you could join them into analyses?"
>
> "I would favour just clearly labelling the dataset itself as synthetic, rather than all the individual variables within"
>
> "More user friendly file structure and data manual would help"
>
> "I'm a big fan of keeping low fidelity with transparent caveats"

> “Having low fidelity whilst keeping the relationships of some variables present could provide more utility whilst keeping the disclosure risk very limited”

Limitations

Our sample was small, more participants would have provided us with a larger sample to gain wider feedback from a range of users.

However, the smaller sample enabled us to keep regular contact with the participants, gain feedback throughout the pilot, and resolve any issues quickly.

Recommendations

The results suggested that low fidelity data was useful, so we are going to continue to focus on low fidelity data.

We will review the governance options and find a long term solution.

We have interests in synthesizing other datasets depending on data owner approval.