DSC-158 Producing synthetic data for Census 2021 rehearsal

2019-02-08

The Campus synthetic data project team is working with PPP Transformation on generating multiple datasets for testing the load balancing and functions used in the processing pipeline of the ONS Census. Broadly, the project involves modifying the variables of the 2011 Census dataset to match the format of the 2021 Census and producing plausible synthetic data for use in testing.

Team members

Data Scientists

- Ioannis Kaloskampis
- Arturas Eidukas
- David Pugh
- Skevi Pericleous
- Jonathan Rees
- Chaitanya Joshi
- Jasmine Latham

Delivery managers

- Sharon Hill
- Lucy Inker-Davies

The need

The general aim of this project is to develop a dataset representative of the expected 2021 Census and Census Coverage Survey (CCS) data from the 2011 pre-Resolve Multiple Responses (RMR) Census data currently available in DAP. This will allow load testing, and some unit testing of program utility associated with the 2019 and 2021 Collection and Processing Rehearsals, and the 2021 Census itself to continue with significantly reduced risk.

Impact

- The main impact of the project is that it supports the Census Rehearsal. Specifically, the generation of synthetic data for this project will mean that the Census team will be able to develop their processing pipeline in advance of receiving the Census data. This will give them more time to prepare and enhance their processing steps.
- Synthetic data generation is currently one of the most challenging problems and is of great interest to the scientific world as a means to: a) provide a dataset for development when the real data are not available, b) share sensitive data by limiting the risk of disclosure.
- Several teams within ONS have expressed an interest for similar tasks. The project will allow the Campus to develop capability for such tasks.

Data science

- Synthetic data techniques applied to the Census data within DAP.
- Efficient handling of Big Data within DAP and data engineering with Pyspark.
- The project utilises the ONS DAP development platform to handle sensitive big data.
- The ONS Census 2011 dataset is used to generate synthetic data.

Broadly, the Census team has a processing pipeline and the Data Science Campus is providing them with data to test some of the stages of their pipeline. The project has four stages:

- Stage 1: Datasets to test the Resolve Multiple Responses (RMR) stage.
- Stage 2: Datasets to test the Data Capture & Coding Requirements Specification (DCCRS) stage.
- Stage 3: Datasets to test the Census Coverage Survey (CCS) and Consolidated stage.
- Stage 4 (optional): Fully synthetic dataset.

The stages of the project are given in the diagram below:

Stakeholders

- Census 2021 Processing team

Code and outputs

- Tables in CSV format as specified by the Census 2021 Processing Team.
- Code in Pyspark within DAP.

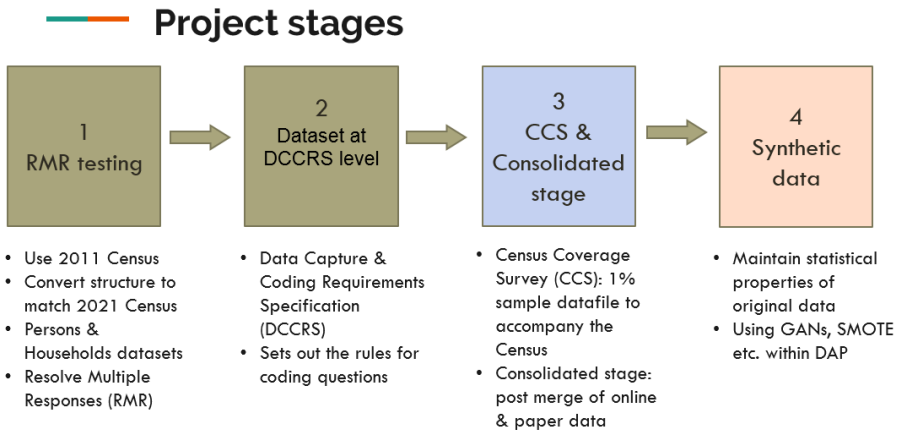

Figure 1: image

Related and existing work

The project utilises techniques discussed in the following reports, published by the Synthetic data team of the Data Science Campus:

a. Quick overview: https://datasciencecampus.ons.gov.uk/synthetic-data-for-public-good-and-art/

b. Technical report - System description & Results https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/

c. Technical report - GANs https://datasciencecampus.ons.gov.uk/projects/generative-adversarial-networks-gans-for-synthetic-dataset-generation-with-binary-classes/

Github repo for project management

https://github.com/datasciencecampus/census_synthetic

Delivery

- [x] **January 2019** Project started
- [x] **April 2019** Delivery of Synthetic data for Stage 1.
- [x] **June 2019** Delivery of Synthetic data for Stage 2.
- [ ] **Mid-October 2019** Delivery of Synthetic data for Stage 3.

Further information

Please contact datasciencecampus@ons.gov.uk for more information.

Updates

- No updates yet.