MANCHESTER 1824

The University of Manchester

Cathie Marsh Centre for Census and Survey Research

SARs Custom Subset Tool

CCSR Working Paper 2006-04
Sam Smith
S.Smith@manchester.ac.uk

This discusses the background and implementation behind the SARs Custom Subset Tool, which allows users to download a customised subset of variables from the 2001 Samples of Anonymised Records.

SARs Custom Subset tool

Sam Smith*
www.ccsr.ac.uk/sars

1 Introduction

The Small Area Microdata[1] (SAM) from the 2001 Census is a 5% Sample of Anonymised Records (SAR[2]) with Local Authority geography, containing 2.98 million cases. The number of cases and 139 variables makes this file unwieldy, requiring significant amounts of memory (3Gb ideally) and CPU power to manipulate. As a result, it may be that the first experience many users get of the SAM will be when they are struggling just to open it in order to reduce the amount of detail to be able to conduct their research.

The SAM is extremely valuable in answering a wide range of research questions, each of which potentially requires a different subset of variables, at differing levels of detail. It has generally been considered that whilst producing "mini-SARs", subsets of specific variables, is possible, it is difficult to design a suite of datasets that cover adequately even a majority of research questions, and it does not match the cost in terms of increased complexity and confusion other than in specific teaching datasets[3].

A different approach is an enhanced data download process, offering variable selection and subsetting functionality to the user. While other tools exist[4], which have some subset and download functionality inbuilt as part of their much wider feature set and analysis, there is nothing that we could find which did this in a simple and efficient manner, specifically for people who do not need nor wish to go through a simplified data exploration process to find the data they already know they want.

Given the tradeoffs between detail in a dataset, and the fact that most users only wish to use specific variables or variable groups from a dataset at any one time, there is a

*s.smith@manchester.ac.uk

[1] `http://www.ccsr.ac.uk/sars/2001/sam/`

[2] `http://www.ccsr.ac.uk/sars/`

[3] the teaching dataset for the SAM is the records for England's North West Government Office Region

[4] Nesstar which we use, and Berkeley SDA which we looked at - `www.nesstar.com` and `http://sda.berkeley.edu`

potential benefit to offering the above as a service, to make the initial stages easier for the researchers. To avoid potential issues in future matching, we mandate that users get certain variables (PNUM and ID in the SAM).

2 Implementation details

To download any of the End User Licensed SARs[5] (of which the SAM is one), users log in using their Athens username, pick the dataset they require, select the format they want it in (generally SPSS or Stata) and download the zip archive to their PC.

For large datasets such as the SAM, when users first open it in SPSS or Stata they may drop variables or cases into which they have no interest in this analysis - dropping records with age under 16 when interested only in those who work full time.

We extend this process by adding an optional stage where users can select the variables they need prior to downloading, and give them a custom download of just the variables they ask for, in SPSS or Stata. Using a list of variables and their standard groupings[6] we present a webpage for which has checkboxes for each variable. All variables in the SAM are in one of 6 groups (individual, household, family, imputation information or imputation flags). These groupings are in the DDI metadata we already have for other processes[7], and it is a simple matter to reformat it for these purposes.

The variable selections are then submitted via a web form which validates that the variables exist in a given datafile, and presents the user's web browser with a page while the custom data file is created. In the background, the program creates a syntax file which stata then runs to read the SAM, perform the subset through the `keep` command on the variables that the user requested, and creates the output file. If the user requested SPSS format output, Stat/Transfer is then run on the output file. Due to the size of the SAM, this process may take a few minutes, so the web page mentioned above will auto-reload with until the process completes.

When the dataset has been created, it is placed in a zip archive, and the page presented to the user is replaced to show a download link. When that page next refreshes, the user may download their custom dataset.

Depending on the size of the datafile, the number of variables and the complexity of

[5] which is similar to all non-Special Licensed UKDA datasets

[6] which we have for other Nesstar already

[7] principally Nesstar, but this is widely reused as all our metadata systems are driven from the same DDI files.

processing being done, this whole process may take less than 30 seconds. The use of the refresh mechanism means that the time taken to process the data is not limited by timeouts which can be problematic through other methods.

2.1 Imperfections

This process currently works on a datafile level, rather than a dataset. While this is not a limitation for the SARs (where the dataset is a single datafile), it may be more of an issue for other datasets spread across multiple files, although there are a number of potential methods of handling this, depending on complexity and size of datasets.

R[8] was used for the initial implementation as a proof of concept and feasibility of the system, mainly due to licensing restrictions on Stata, and licensing is a large reason for not using SPSS at all. Where the data processing was done by R, the data structures available were limited to those inside R. While this does not affect the data itself in any way, and was fine for prototyping, variable labels were lost on export to SPSS.

For production use, we have switched to Stata to perform the analyses due to it being more memory efficient, faster and just as flexible. Stata is the main format used for data processing within the SARs team and hence is a format that we natively supply and support to users. While Stata does not offer SPSS output, SPSS can read Stata files; however, due to user demand, we use Stat/Transfer to convert to SPSS if the user requests an SPSS file.

3 Further potential applications

It is possible to look at requests and prohibit specific combinations of variables on a per request (or even per user basis[9]). As an example, if a dataset contained 5 "detailed" variables, researchers could be limited in the requests that would be accepted: "pick any 2 detailed variables". Similarly "problematic" variable combinations could be prohibited or routed to manual checks or other processes as required.

While the only current check on a variable is that it exists, that check is done outside and prior to the stata statistical engine being run. At this point it is possible to do arbitrary checks on who is requesting which variables, using any information that can

[8] `http://www.r-project.org`

[9] Using the incoming academic authentication system called Shibboleth which will replace Athens, it will be possible to also limit based on attributes of users

be available within the system or which can be available to it. These can be checks against a database of past downloads, and before the dataset is created, be set to route users to require additional, possibly manual, confirmation through other channels where applicable. Different checks could apply to different users or variables, and could take place at any point in the process at the start, based on the user, or the data they request before any data is accessed, or within Stata after the subset has been created based on that request.