Automatically Importing, Appending and Merging the Demographic and Health Surveys

Damian C. Clarke*

September 7, 2012

Abstract

I provide here a quick guide to producing multi-country DHS files from the 1500+ surveys provided on-line. This guide is provided as a companion to the program DHS_Import.py and DHS_Multicountry.do which completely automate the creation of a comprehensive database of all publicly-available DHS files. In order to create these databases the user simply requires an internet connection, and the ability to run the two provided files; the first using the free program Python and the second using the statistical program Stata. The programs provided are written to include all surveys available at the date of use, even if those surveys were not available when the program was written.

1 Introduction

The Demographic and Health Surveys (DHS) are a set of surveys collected in developing countries which provide data relating to maternal health and fertility, child health, as well as specific modules on HIV, anemia and maternal mortality (among other variables). These files are publicly available on-line at http://www.measuredhs.com, along with a full description of their contents, methodology and use.

Currently the DHS files are stored by survey country and year, resulting in approximately 1600 separate files over the period 1990-2011. Hence, a user interested in working with cross-country data, must either manually 'point and click' to download these and then merge the downloaded files, or write a script to download and unzip the files and then merge the resulting

*Contact mail damian.clarke@economics.ox.ac.uk

files. I provide here such a routine which allows for the automatic importing, appending and merging of all publicly available DHS datasets. This routine consists of two files, a file written using the computer language Python which downloads, unzips and saves all files, and then a stata do file to process the DHS files. The reason I write the downloading script in Pyton rather than the more commonly used (by economists) ado language for Stata is due to Python's flexibility, as well as its simplicity and the fact that it is free and installed on many operating systems 'out of the box'[1]. In order to use this program it is not necessary for the user to know any Python, although a familiarity with Stata is required as all resulting datasets are produced in Stata's `.dta` format.

2 Files

Figure 1 provides a description of the DHS datasets including optimal merging direction. There are seven distinct datasets provided by the DHS (excluding the Wealth Index data and Height and Weight data). Each of these datasets merges directly with the Individual Women's Recode (IR), with the exception of the Male Recode (MR) which only merges to the IR in the case that a male and female live together and both report sharing a relationship. In this case, the merged database is available as the couple's recode (CR).

Each file link is joined by a unique merge code and the "DHS guarantees that their files can be matched seamlessly whenever a relationship is possible". The DHS provides a description of how to join each mergeable set of files on its website[2], however this table assumes that the user is working with data from a single country. Hence, to ensure that data still merges seamlessly, I add two new variables to each file; _cou (country) and _year (year) (see lines 27–29 of the file DHS_multicountry.do). In this way, files are merged as described in table 1.

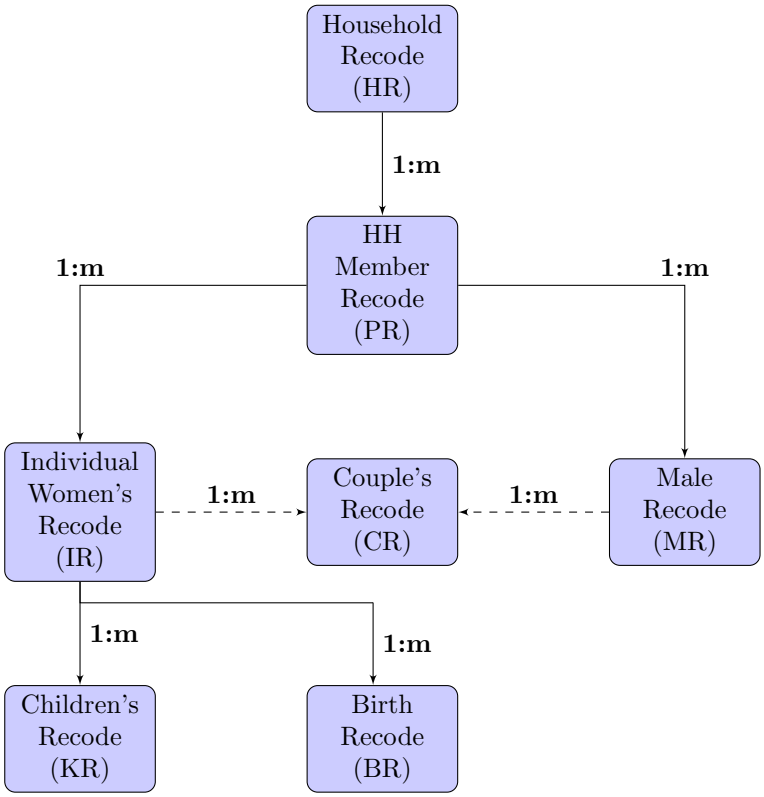
Table 1: DHS Merge variables

	Secondary Files	
	Match for household	Match for women
Base	_cou + _year + HV001 + HV002	_cou + _year + V001 + V002 + V003
Women	_cou + _year + V001 + V002	
Children	_cou + _year + V001 + V002	_cou + _year + V001 + V002 + V003
Men	_cou + _year + MV001 + MV002	Couples: _cou + _year + MV001 + MV002 + MV034i

[1]For example Python is included by default on ubuntu. Windows users without Python installed can install version 2.7.3 (suggested) here.

[2]`http://www.measuredhs.com/data/Merging-Datasets.cfm`

Figure 1: Links between DHS Recodes



Given that DHS files exist for each recode type (figure 1) and each country/year pair available, the file DHS_multicountry.do creates a cross-country dataset for each file type by merging each recode of a given type together. Thus, the resulting files will be HR_World, PR_World, IR_World and so on. Given that each of the mergeable 'World' files still has a unique merge code (as described in table 1), these can then be merged according to the user's needs. Using all files available at the date this document was produced, the size of the resulting files are listed in table 2. Whilst the sizes of these datasets is relatively large for work on a personal computer, these can easily be shrunk to a manageable size by either limiting the quantity of variables included, or by splitting the dataset into various smaller datasets. The datasets listed in table 2 include all cross-country variables, many of which will not be relevant in any given analysis. Due to this fact, the user can specify in the program which variables they wish to keep before appending the individual country files (line 49 of DHS_Multicountry.do).

Table 2: DHS Merge variables

File Name	Size
HR_World	21.4 GB
PR_World	9.2 GB
IR_World	16.9 GB
MR_World	1.2 GB
KR_World	2.9 GB
BR_World	9.5 GB

3 Creating the Cross-Country DHS Files

3.1 Running the Programs

The programs provided along with this document; DHS_Import.py and DHS_Multicountry.do, are operating-system independent, and require only that the user possess access to Python (freely available) and Stata (available at most universities). These programs have been successfully tested on Ubuntu 12.04 and Windows 7, and should work without difficulty on other Linux and Windows systems, plus MacOS. To run the programs, the user must download the two program files, plus an auxiliary text file (DHS_Countries.txt) which lists the full set of DHS survey countries. These three files should be saved in the same working directory.

In order to import and build the DHS data, the user should first run the Python file to download and unzip all datasets. This is done as follows (assuming that the user has already installed Python): firstly download the file from `https://sites.google.com/site/damiancclarke/computation` and save in any directory, also ensure that you are logged into the DHS data download website using your user name and password (available via on-line application). Secondly, open the command line (also known as the terminal) of your computer[3]. Finally, at the command line change to the working directory where you saved DHS_Import.py (using the command `cd`), and enter the following:

```
python DHS_Import.py file1 file2
```

where the argument `file1` represents the directory where you want to download the unzipped `.dta` files, and `file2` the folder where your operating system by default downloads files from the web (this may be something like "∼/Downloads" on Linux-based operating systems or "C/Downloads" on Windows). Once having entered this command the program will proceed to download all available datasets, and save them in the directory specified `file1`. Given that the speed at which the program runs depends upon the strength of the internet connection, it is

[3]In ubuntu this can be accessed by simply typing `Ctrl + Alt + T`, in Windows by typing `cmd` at the Run option of the Start menu, and in Mac via the Applications folder (`Applications→Utilities→Terminal`)

recommended at this stage to have access to a stable connection. It is also highly recommended that download settings be changed (if necessary) to ensure that the user need not authorize each individual download in a pop-up dialog box.

Once having downloaded all current DHS files, these are appended together into a final set of usable databases using the Stata file DHS_Multicountry.do. This file requires very few changes: the user simply need specify the values for the global and local variables on lines 14-16. The first global variable should be set to the directory in which the Python program was saved, the second global to the argument `file1` from the Python command, (the location of the data) and finally the local "clean" must be specified. For this local the user should enter `yes` if they wish for files added by the program to be removed upon completion (this is advisable where the user is concerned about the use of hard disk space on their computer).

3.2 Resulting Files

These programs will result in a full set of DHS databases as described in table 2, along with three raw text files. The first of these text files contains one line for each country (and country code) and year that the survey is run (eg CO Colombia 1990). The second lists all individual databases downloaded (for each country/year pair multiple surveys exist), and finally, a similar file is produced for use by the Stata `.do` file. These files may be of interest to the user should they require summary statistics of years and places surveyed in the DHS.

3.3 Final Points

These programs are provided for open use and can be modified in any way that the user desires. Any comments, queries or suggestions are welcome. The Stata `.do` file provided here does not include commands to merge datasets once they have been produced, as the types of merges required and the variables included will vary widely depending upon the interest of the user. I do however have `.do` files which program these merges, and am happy to make these available to interested users. For these files, or for any other details regarding these programs, please contact the author at damian.clarke@economics.ox.ac.uk.