Max-Planck-Institut für demografische Forschung
Max Planck Institute for Demographic Research
Konrad-Zuse-Strasse 1 · D-18057 Rostock · GERMANY
Tel +49 (0) 3 81 20 81 - 0; Fax +49 (0) 3 81 20 81 - 202;
http://www.demogr.mpg.de

MPIDR TECHNICAL REPORT 2015-001
APRIL 2015

Merging, exploring, and batch processing data from the Human Fertility Database and Human Mortality Database

Jon Minton (jonathan.minton@glasgow.ac.uk)

For additional material see www.demogr.mpg.de/tr/

This technical report has been approved for release by: Vladimir Shkolnikov (shkolnikov@demogr.mpg.de), Head of the Laboratory of Demographic Data.

Technical reports of the Max Planck Institute for Demographic Research receive only limited review. Views or opinions expressed in technical reports are attributable to the authors and do not necessarily reflect those of the Institute.

Merging, exploring, and batch processing data from the Human Fertility Database and Human Mortality Database

Dr Jon Minton[1]

20 May 2015

Introduction

As well as picking and choosing particular countries and variables to download from the Human Fertality Database (HFD) and the Human Mortality Database (HMD), it is also possible to initiate a bulk download of all data from either database. In both cases, the user is presented with a compressed file in .zip format, containing a number of separate variables in a large number of separate files. In the case of the Human Mortality Database, the same variable is contained in separate files in separate directories for each of the countries for which the variable have been collected. Although, when interested in a single country's records, it can be relatively straightforward to identify which file, from which directory, the appropriate data should be loaded from, some analyses would require accessing a large number of files, and so the amount of effort required to manage the data in these cases would be much increased. For example, if a researcher wants to combine population and death counts from a number of countries, and further form this to see how a variable differs in one country compared with a number of others, then the number of separate files that would need to be accessed could be very large.

If, instead, relevant data from a number of separate countries, and for a number of separate variables of interest, were available, in a consistent format, within a single file, then the amount of effort required to perform such comparative analyses can be much reduced.

The first part of this technical report will describe how to use two functions that I have developed, in R, in order to automatically merge a large number of files from the HFD, and HMD respectively, into just two datafiles, which can be then saved as plain text files, ready to use for further analyses. The second part of this technical report presents a motivating example, showing the benefits of having such data in a smaller number of files, by showing how the process of producing shaded contour plots (SCPs) for each of the countries and variables included can then be automated, meaning that potentially hundreds of different analyses can be produced and updated very quickly.

Both the functions for data gathering presented in the first half of the technical report, and the functions which generate images as presented in the second half of the technical

[1] Affiliation: School of Social & Political Sciences,
College of Social Sciences,
University of Glasgow, UK
Email: jonathan.minton@glasgow.ac.uk

report, make extensive use of the 'plyr' package, and the associated 'split-apply-combine' paradigm. [1,2] Readers are encouraged to learn more about the plyr package, and the related dplyr package, and to explore the contents of the functions used here, in order to understand how they work, and how the functions and approach described here can be applied to a much wider range of data management and analyses tasks in demographic analyses. [3] However even if the functions are thought of as 'black boxes' and their contents are not explored, it is hoped that the outputs they produce will be valuable to researchers in cutting down the amount of time they need to spend on data management tasks rather than substantive analyses.

Preparation

Tidy Data Principles

Hadley Wickham defines a dataset as 'tidy' if:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.[4]

Each row within a 'tidy data' format table contains two types of variable:

1. **'Where'** variables, defining the 'location' of the observation; and
2. **'What'** variables, defining the value measured at the 'location'.

In the case of the demographic data of interest in this exercise, 'location' variables include:

1. Age;
2. Year;
3. Country;
4. Sex (for the HMD).

Each row therefore begins with the three (HFD) or four (HMD) location variables, followed by a number of observation variables for that particular location. These could include:

1. Death counts.
2. Population counts.
3. Births

With data arranged this format, it becomes easier to derive additional variables, to compare between groups, and to automate the production of outputs for each country separately.

Note it is important that 'location' variables are compatible. In the case of demographic data, is is particularly important to ensure that the years and ages to be combined all refer to either Lexis squares, or Lexis triangles, or Lexis parallograms, and not to an inconsistent mixture of these. Combining data in this way would be incorrect, the analogy within demography of combining data that use different coordinate reference systems within spatial data analysis.

The functions presented here will all make use of variables aggregated to Lexis squares using calender time in years. However similar functions could be written to fetch other forms of compatible data.

Data preparation

A zipped file is available alongside this technical report. This contains a subset of the data available from the HFD and HMD bulk download options, as described above. The functions described here will work using this subset of files, but in order to make best use of the functions I recommend that the full datasets are downloaded to the appropriate locations as detailed below.

The zipped files associated with this technical report will work as-is. However, to get the full benefits of the batch processing functions presented here it is recommended that, after trying the functions out on the subsets of data contained in the zipped files, the full datasets from the HMD and HFD are loaded into the raw_data subdirectories as detailed below.

The directory structure assumed is:

- base_dir
    - data
        - raw_data
            - hfd
            - hmd
        - derived_data
    - scripts
    - figures
        - asfr
        - population
        - logmort

To avoid violating the HFD and HMD user agreements, the HFD and HMD datasets are not included in the zipped files. Instead, please download the zipped files from the respective websites, and unzip these files into the following directories:

- For the HFD: into the directory base_dir/data/raw_data/hfd/
- For the HMD: into the directory base_dir/data/raw_data/hmd/

For further details please see the appendix. The zipped file associated with this technical report contains a small number of files arranged in the structure above.

Loading the functions

The functions are contained in a script file, within the scripts directory above, called 'functions'. The location of the working directory within an R session can be found using

```
getwd()
```

If the present working directory is not the correct base directory, i.e. the directory containing the unzipped files associated with this technical report, then it can be set using the 'setwd()' function. For example, if the base directory is in the location "d:/hfd_hmd_filemanagement/", then the present working directory of the R session can be set to this using

```
setwd("d:/hfd_hmd_filemanagement/")
```

So long as the working directory has been set correctly, then both the functions can be loaded using the source function, as follows:

```
source("scripts/functions.R")
```

In order to check that the functions have been loaded correctly, we can check for all objects in the R workspace using 'ls()', or the functions only using 'ls.str()'. If the functions have been loaded correctly then either call should show the following functions to be within the R workspace

- merge_lexis_square_hfd
- merge_lexis_square_hmd
- generate_scp_pop
- generate_scp_log_mort
- generate_scp_asfr

The two functions beginning with merge_ will merge and combine the data from the HFD and and HMD respectively, and so will be the focus of part one of this technical report. The three functions beginning with 'generate' will be used to produce SCPs using the data as arranged by the merge_ functions; they will be the focus of part two of this technical report.

Part one: merging and combining data from many files

The 'merge_lexis_square_hfd' function

With the HFD data downloaded, unzipped, and in the location specificed above, we can use the merge_lexis_square_hfd function by specifying the location of the HFD directory as its first argument, named 'loc'.

```
hfd_tidy <- merge_lexis_square_hfd(loc="data/raw_data/hfd/")
```

The function takes extracts the following variables from separate HFD data: ASFR, pop....

We can look at the first few rows of the data using the head command.

```
head(hfd_tidy)
```

We can also change how the hfd_tidy data is presented to us within R by using the tbl_df() function from the 'dplyr' package. After doing this, we can see the first few observations, as well as other information about the object such as its size, just by typing the object name

```
hfd_tidy <- tbl_df(hfd_tidy)
hfd_tidy
```

The data can be saved into the directory `data/derived` as follows:

```
write.csv("data/derived/hfd_lexis_square.csv", row.names=FALSE)
```

The 'merge_lexis_square_hmd' function

Although it is a more complex function internally, the merge_lexis_square_hmd function can be used in exactly the same way as the merge_lexis_square_hfd function, by passing the location of the correct directory to the function as its first argument, called 'loc'

```
hmd_tidy <- merge_lexis_square_hmd(loc="data/raw_data/hmd/")
```

As with the HFD object earlier, the first few observation of the combined dataset can be shown in the R console as follows

```
head(hmd_tidy)
```

or

```
hmd_tidy <- tbl_df(hmd_tidy)
hmd_tidy
```

Zipped example files with this technical report

This technical report comes with a zipped file. This contains the functions described here, and the directory structure detailed above. . To begin with, please try out these the functions using this small subset of the data. However, to make best use of the functions and their batch processing functionality, please download the full HFD and HMD datasets into the appropriate directories as described above.

Part Two: Batch generating outputs

With appropriate and compatible data combined into a single file as described above, it becomes easier to perform the same operation on each of many subgroups within the dataset as a batch process. As an example of this, and with the hmd_tidy and hfd_tidy datasets generated before, the functions `generate_scp_asfr`, `generate_scp_population` and `generate_scp_logmort` are provided.

These functions generate shaded contour plots of fertility rates, population sizes, and mortality rates, for each country available in the hmd_tidy and hfd_tidy datasets. Shaded contour plots use both contour lines and shades to show how the value of a variable varies as a function of age, year and, in the case of HMD data, gender. [5–8]

The data from each country/code is used to generate a separate SCP image file, with a filename that reports the subset of the data used, and the range of years over which the data are available for that country. Therefore, it is possible to generate dozens of outputs, in this case high resolution images, using just one line, as shown below:

```
generate_scp_asfr(dta_hfd)
generate_scp_population(dta_hmd)
generate_scp_logmort(dta_hmd)
```

For example, the `generate_scp_asfr` function generates 31 separate image files, one for each country/code subgroup contained in the data, each contained in the figures/asfr subdirectory. Two examples of this, for the Netherlands and the Ukraine, are provided in the figures below:

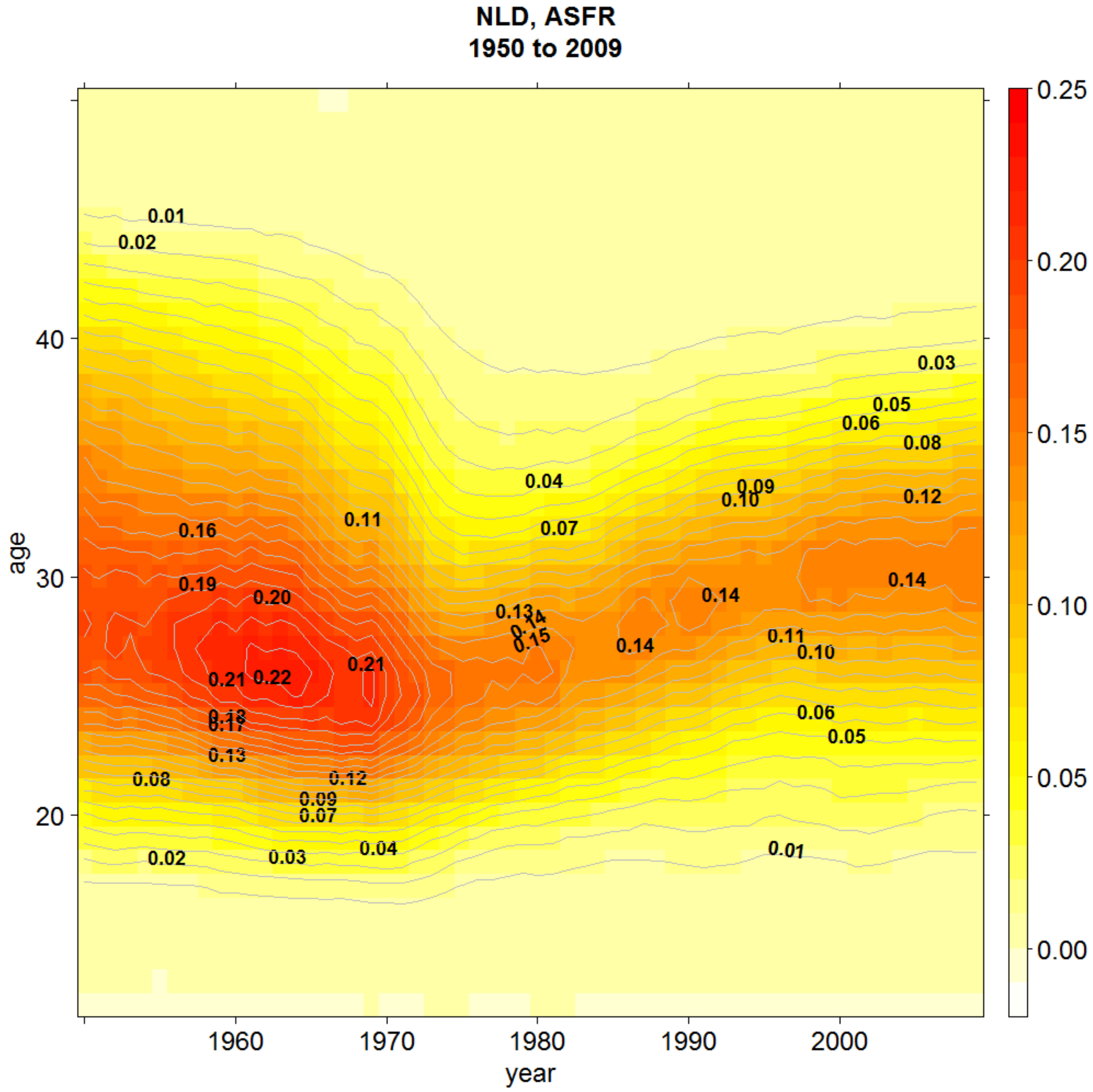

Figure 1 SCP of ASFRs in the Netherlands, 1950 to 2009

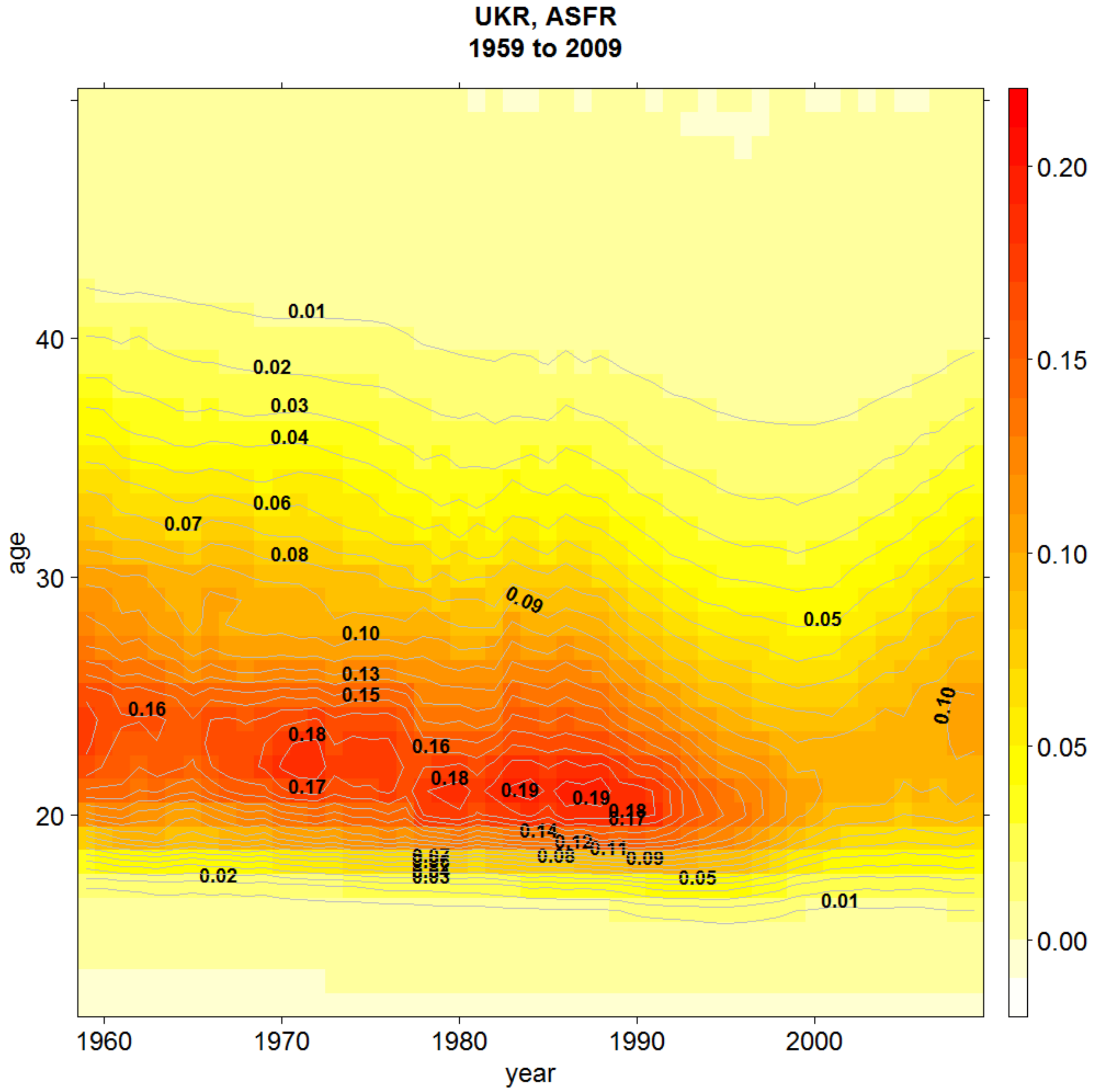

Figure 2 SCP of ASFR for Ukraine, 1959 to 2009

Discussion

This technical report has described a small number of functions that help to automate certain batch processing operations that may be involved when working with data from the HFD and HMD. Two of these functions work to combine appropriate data from a range of different files into a format that is easy to work with; three other functions take these derived datasets and produce a large number of images on the basis of them.

The functions described here are intended to be useful to users of HMD and HFD data in their own right, but also to be introductions to a much broader paradigm of data management and batch processing in R that, once adopted, can make the development of additional functions much easier and additional batch processes much more streamlined. Users are encouraged to explore the contents of the functions, paying attention to how they make use of the plyr and dplyr packages, and how their contents could be used as the blueprint for performing a much wider range of batch merging and batch processing operations with the data. Please also visit the website, and feel free to contact me with any queries or suggestions:

https://github.com/JonMinton/human_fertility_database/

References

1 Wickham H. plyr. 2015.http://cran.r-project.org/web/packages/plyr/plyr.pdf

2 Wickham H. The Split-Apply-Combine Strategy for Data Analysis. *J Stat Softw* 2011;**40**.

3 Wickham H, Francois R, Rstudio. dplyr. 2015.http://cran.r-project.org/web/packages/dplyr/index.html

4 Wickham H. Tidy Data. *J Stat Softw* 2014;**59**.http://www.jstatsoft.org/v59/i10

5 Vaupel JW, Gambill BA, Yashin AI. *Thousands of Data at a Glance: Shaded Contour Maps of Demographic Surfaces*. Laxenburg, Austria: 1987. http://user.demogr.mpg.de/jwv/pdf/Vaupel-IIASA-RR-87-016.pdf

6 Vaupel JW, Wang Z, Andreev K, *et al. Population Data at a Glance: Shaded Contour Maps of Demographic Surfaces over Age and Time*. University Press of Southern Denmark 1997. http://www.abebooks.co.uk/servlet/BookDetailsPL?bi=2944819605

7 Minton J, Vanderbloemen L, Dorling D. Visualizing Europe's demographic scars with coplots and contour plots. *Int J Epidemiol* 2013;**42**:1164–76. doi:10.1093/ije/dyt115

8 Minton J. Real geographies and virtual landscapes: Exploring the influence on place and space on mortality Lexis surfaces using shaded contour maps. *Spat Spatiotemporal Epidemiol* 2014;**10**:49–66. doi:10.1016/j.sste.2014.04.003

Appendix: Accessing the full datasets

Accessing all available data from the HFD

- Go to humanfertility.org and select login from the Registration section of the column on the left of the page.
- Once logged in, select 'Zipped Data Files' under the 'DATA' section of the column on the left of the page.
- Within the table 'Data by type', look for the link to the data type 'All types of HFD' data. Click on this link to begin the download.

The size of the HFD file is around 25Mb. Once unzipped, this increases to around 144Mb.

Accessing all available data from the HMD

- Go to mortality.org and log in.
- Select 'Zipped Data Files'; scroll to the bottom of the page and click on the link in the table 'All countries for the HMD'.
- Once logged in, select 'Zipped Data Files' under the 'DATA' section of the column on the left of the page.
- Within the table 'Data by type', look for the link to the data type 'All types of HFD' data. Click on this link to begin the download.

The size of the HMD file is around 311 Mb. Once unzipped, this increases to around 1.24 Gb.

Structure of HFD directory

The HFD has a relatively straightforward directory structure: within 'hfd' is a directory called 'Files', and within this directory is a directory called 'zip_w'. A total of 56 files are in this directory, and there are no additional directories.

Structure of the HMD directory

The HMD directory has a more complex, branched structure. Once unzipped, the directory opens to 46 separate folders, each labelled with the country code of the country whose data they contain. Each of these country folders then has the same internal directory structure. As an example here is the directory structure of the first country by code, AUS:

- AUS
  - CHECKS
  - DOCS
  - InputDB
  - LexisDB
  - STATS

The STATS directory then contains 55 data files in comma-separated value (CSV) format. Within this technical document, the aim will be to extract data from the files `Deaths_1x1.txt` and `Populations.txt` from each of these subdirectories, but the methods described can be generalised to other operations.