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 2011-003
MAY 2011

An "R" package for the production of a Lexis database of fertility data

Dmitri Jdanov (jdanov@demogr.mpg.de)
Edward Nash (nash@demogr.mpg.de)

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.

An “R” package for the production of a Lexis database of fertility data

by Dmitri Jdanov and Edward Nash

Abstract

This technical report introduces software developed within the framework of the Human Fertility Database (HFD). The data on births provided by Statistical Offices are often classified only by calendar year and age of the mother or by calendar year and birth cohort of mother. For some countries and calendar years, birth data are available by five-year age intervals only. They may show broader or narrower ranges of available ages, they may include births with unknown age of the mother or unknown birth order, or they may show total births instead of live births. As part of the HFD project, a standardised methodology has been developed for the transformation of any set of raw data into data classified by single years of age and birth cohort, and (whenever possible) by birth orders Births with unknown age of the mother are distributed proportionally according to the birth data where age of the mother is specified. Within each age, births with unknown birth order are distributed proportionally across known birth orders. Aggregated age groups are additionally split into single-year ages by means of spline interpolation. Birth orders higher than five are combined into birth order 5+. Within each age, births are additionally split by year of birth of the mother (if such information is not present in input data). This Technical Report describes software in the form of packages for the free statistical computing environment “R” which implement the HFD methods to perform this manipulation.

Keywords: fertility, birth counts, splitting, spline interpolation, iterative proportional fitting, Human Fertility Database, R

Background

In the Human Fertility Database (HFD)[1], various output rates and indicators are produced using a standardised methodology from detailed uniform data on births presented by year, age, and year of birth of the mother and birth order, and female population exposure presented by year, age, and cohort (so called Lexis database). The Lexis database is produced from input data which may have a varying structure in terms of ages and birth orders, and is frequently less detailed than is required for the Lexis database. The data must therefore be manipulated to estimate the allocation of births to Lexis triangles in the age range 12- to 55+.

This technical report summarises the methodology which is used to perform this manipulation before introducing a package for the statistical programming environment “R” which implements these methods and presenting examples as to how the functions in the package may be used.

[1] A joint project of the Max Planck Institute for Demographic Research (MPIDR) and the Vienna Institute of Demography (VID), available at http://www.humanfertility.org

Notation

The HFD considers the reproductive span between age 12- ($x_{\min}$) to age 55+ ($x_{\max}$) and birth data for orders 1…5+. To enable the HFD methodology to be applied flexibly by other users who may wish to consider a greater or lesser parity range, the HFD Methods Protocol has been generalised in the description below: this generalised form is supported by the "R" packages, although the values for many parameters default to those used by the HFD. The notation i_b^+ is used here for the highest (open-interval) birth order.

Methods

Four major steps may be considered in the data processing, not all of which are required in all cases:

1. Production of estimates of female population exposure to risk.
2. Distribution of births at unknown age and/or order.
3. Splitting of open-ended and multiple age categories to single ages.
4. Splitting of single age categories to Lexis triangles.

Additional steps may be required where there are separately recorded late-registered births, particularly where these are recorded according to a different Lexis shape to the remaining data. In all cases, total births and data for each birth order are handled separately. In order to achieve balance of the order-specific data with the total births, an iterative proportional fitting (IPF) procedure is applied after each step. Note that in many cases, the data manipulations result in a non-integer estimate of the number of births in each category. The following subsections summarise the methods used, assuming the original data are in Lexis rectangles: the methods may be straightforwardly adjusted where the original data are in vertical parallelograms.

Production of female population exposure estimates

In the HFD, one of two different methods is used for estimation of female population exposure to risk by age, year, and birth cohort $E(x,t,c)$, depending upon availability of data. For cohorts where monthly birth data are available and mean date of birth $\bar{b}(c)$ (measured in years from 1st January of the year of birth) for the cohort of the mother can be calculated, the method (equations ((1) and ((2) below) is based on Calot & Sardon (2003). Where $\bar{b}(c)$ can not be determined, the method is based on Wilmoth et al (2007), using death counts in the Lexis triangle $D(x,t,c)$ to slightly improve the basic exposure estimate (equations ((3) and ((4)). In either case, the exposure estimate is fundamentally based on the population at age x on 1st January of years t, $P(x,t)$ and $t+1$, $P(x,t+1)$.

$$E(x,t,t-x-1) = P(x,t)\cdot \bar{b}(t-x-1) \qquad (1)$$

$$E(x,t,t-x) = P(x,t+1)\cdot \left(1-\bar{b}(t-x)\right) \qquad (2)$$

$$E(x,t,t-x) = \frac{P(x,t+1)}{2} + \frac{D(x,t-1,c)}{6} \qquad (3)$$

$$E(x,t,t-x-1)=\frac{P(x,t)}{2}-\frac{D(x,t,c)}{6} \tag{4}$$

Distribution of births at unknown age and/or order

Births with unknown age and/or order are in distributed proportionally between the defined categories. If $B_i^{UNK}(t)$ is the number of births with unknown age of mother and $B_i^{TOT}(t)$ is the total number of births across all ages at that birth order then the adjusted number of births in each age category x is

$$B_i^*(x)=B_i(x)\cdot\left(\frac{B_i^{TOT}(t)}{B_i^{TOT}(t)-B_i^{UNK}(t)}\right) \qquad 5$$

Births with either an unknown birth order or both unknown age and birth order may be similarly distributed: for the HFD, first births with both unknown age and birth order, then with unknown birth order and finally with unknown age are distributed. Thereafter, the IPF procedure is used to ensure that marginal totals (by age and by birth order) match.

Splitting of open-ended and multiple age categories to single ages

In many cases, births are not available by single years of age but by 5-year age bands, or there are open-interval categories at the lower and upper range, e.g. 15-, 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50+. There may also be a mixture of ranges where some births are recorded in 5 year groups and others in 1 year groups. A spline interpolation approach is used in these cases in order to split the whole range into single age categories.

It is first assumed that there are no births at age 11 ($x_{\min}-1$) or below and at age 55 ($x_{\max}$) or above – allowing closed-interval categories to be used at both upper and lower ends and giving n age groups $[x_i,x_{i+1}), i=0,...,n$, covering the age range $[x_{\min},x_{\max})$. The cumulative fertility rate $F(x)$ at the end of each age category is first calculated using equation (6 where $f(x_i;x_{i+1})$ is the mean unconditional fertility rate over the age interval $[x_i,x_{i+1})$, $B(x_i;x_{i+1})$is the number of births, and $E(x_i;x_{i+1})$ is the female population exposure within the age interval $[x_i,x_{i+1}), i=0,...,n$.

$$F(x)=\sum_{i=0}^{n}(x_{i+1}-x_i)f(x_i;x_{i+1})=\sum_{i=0}^{n}(x_{i+1}-x_i)\frac{B(x_i;x_{i+1})}{E(x_i;x_{i+1})} \tag{6}$$

The continuous approximation of this function is estimated by interpolation of the logit transformation using a Hermite cubic spline, i.e. the logit transform $Y(x)$ of $F(x)$ is produced according to equation ((7), the continuous approximation $\hat{Y}(x)$ is produced by Hermite cubic spline interpolation and then the continuous approximation $\hat{F}(x)$ through an inverse logit transformation of $\hat{Y}(x)$ according to equation ((8). Since the logit function is undefined at the extremes, -20 and +20 are substituted at $x_{\min}$ and $x_{\min}$ respectively, i.e. $Y(x_{\min})=-20$, $Y(x_{\max})=20$.

$$Y(x) = \text{logit}\left(\frac{F(x)}{F(x_{\max})}\right) = \log\left(\frac{F(x)}{F(x_{\max}) - F(x)}\right),\ Y(x_{\min}) = -20,\ Y(x_{\max}) = 20 \qquad (7)$$

$$\hat{F}(x) = \left[\frac{e^{\hat{Y}(x)}}{1 + e^{\hat{Y}(x)}}\right] F(x_{\max}) \qquad (8)$$

From the estimated continuous cumulative fertility function $\hat{F}(x)$, the fertility rate at each single age is estimated according to equation ((9), from which the number of births at each single age $\widehat{B}(x)$ may be estimated using equation ((10). Since the sum of the births estimated using this method in each original age interval is not intrinsically equal to the original birth count, the final estimate of the number of births at each age $B(x)$ is made by correcting $\widehat{B}(x)$ according to equation ((11).

$$f(x) = \int_{x}^{x+1} d\hat{F}(s) = \hat{F}(x+1) - \hat{F}(x) \qquad (9)$$

$$\widehat{B}(x) = f(x)E(x) \qquad (10)$$

$$B(x) = \widehat{B}(x)\frac{\sum_{z=x_i}^{x_{i+1}-1} \widehat{B}(z)}{B(x_i; x_{i+1})},\ x_i \le x < x_{i+1} \qquad (11)$$

Note that this procedure is applied independently for each calendar year and birth order (or total births), with the balance of marginal totals achieved by subsequently applying iterative proportional fitting.

Splitting of single age categories to Lexis triangles

In principle, the method used for splitting births recorded in single age categories to Lexis triangles (i.e. by birth cohorts) is the same as that used for splitting open- and multiple-age categories, and is again applied independently for each calendar year and birth order. The continuous cumulative fertility function $\hat{F}(x)$ is estimated based on the interpolation using Hermite cubic splines of a logit transformation of the discrete cumulative fertility function $F(x)$ (equations ((6)-((8)). The fertility rate in each of the upper and lower triangles is then calculated through integration of $\hat{F}(x)$, using the fact that the area of an elementary triangle is ½ (equations ((12) and ((13)), and this fertility rate may then be multiplied with the female population exposure to produce the initial estimate of the number of births in each Lexis triangle $\widehat{B}_U(x)$ and $\widehat{B}_L(x)$ (equations ((14) and ((15)). As the sum of the births estimated in the two triangles using this method is not necessarily the same as the original birth count, the final estimate of the births in each triangle is made by correcting $\widehat{B}_U(x)$ and $\widehat{B}_L(x)$ according to equations ((16) and ((17).

$$f_U(x) = 2\int_{x}^{x+1} (s - x)\, d\hat{F}(s) \qquad (12)$$

$$f_L(x) = 2\int_{x}^{x+1} [1 - (s - x)]\, d\hat{F}(s) \qquad (13)$$

$$\widehat{B}_U(x) = f_U(x)E_U(x) \tag{14}$$

$$\widehat{B}_L(x) = f_L(x)E_L(x) \tag{15}$$

$$B_U(x) = \widehat{B}_U(x)\frac{B(x)}{\widehat{B}_U(x) + \widehat{B}_L(x)} \tag{16}$$

$$B_L(x) = \widehat{B}_L(x)\frac{B(x)}{\widehat{B}_U(x) + \widehat{B}_L(x)} \tag{17}$$

Once the birth counts have been split independently for each birth order, the iterative proportional fitting procedure should be applied to ensure that marginal totals match.

Iterative proportional fitting

The iterative proportional fitting (IPF) procedure is an iterative method for estimating table cell values such that fixed marginal totals are obtained, details of which are presented e.g. in Fienberg (1970) and Bishop et al. (1975). In the case of the HFD, IPF is used to ensure that the sum of order-specific births within an age category is equal to the total births whilst the sum of births across age categories is equal to the known total. The marginal totals used are therefore the total births by age category (row totals) and the birth counts in each birth order within an aggregated age category (column totals).

The initial matrix cell values $B_i^0(x)$ are the result of splitting of aggregated age groups or redistribution of unknowns $B_i(x)$. At each iteration k for $k = 0,2,4,\ldots$ the difference between the cell sums and the marginal total is distributed proportionally between the cells, first by row (equation ((18)) and then by column (equation ((19)). This is repeated until the value of the distance function Λ between the cell values and the marginal totals (calculated according to equation ((20)) is less than 10^{-6}.

$$B_i^{k+1}(x) = \frac{B_i^k(x)}{\sum_i B_i^k(x)} \cdot B(x) \tag{18}$$

$$B_i^{k+2}(x) = \frac{B_i^{k+1}(x)}{\sum_x B_i^{k+1}(x)} \cdot B_i. \tag{19}$$

$$\Lambda^k = \sum_x \left(\sum_i B_i^k(x) - B(x) \right)^2 + \sum_i \left(\sum_x B_i^k(x) - B_I \right)^2 \tag{20}$$

The "R" packages `hfdIndbLoad` and `hfdIndb2ldb`

All calculations for the Human Fertility Database are programmed in R[2]; a number of the functions used which may be of more general interest are being made publically available as R packages accompanying Technical Reports.

[2] "R" (R Development Core Team, 2010) is a language and free software system for statistical computing and graphics.

R is usually operated in a command-line environment with commands entered by the user at the "R prompt". In the following sections, input at the R prompt is shown in **`> bold Roman type`**, with output from R shown in *`oblique type`*.

Content

As part of a modular system to encourage reuse of functions, the software described in this Technical Report is divided into two packages:

- `hfdIndbLoad` contains two utility functions for reading and writing births files in the HFD input database format, converting to/from an internal `data.frame` representation.
- `hfdIndb2ldb` implements the methods described in the previous section for redistribution of births with unknown parameters and splitting of aggregated age categories. This package depends on the external package `signal`, which is available via the CRAN repository network if not already installed[3].

The major functions in these packages are summarised in Table 1 and Table 2, with the parameters accepted by each function detailed in Table 3 and Table 4. The data formats used are described in more detail in the following section.

Table 1. Summary of functions in package `hfdIndbLoad`

Function name	*Purpose*	*Main input data and formats*	*Output data format*
`indbLoad`	Reading a HFD births input file and converting to internal `data.frame` format	Path to a file in HFD births input file format	`data.frame` containing a representation of the data from the file
`indb.print`	Writing a HFD births input file based on the contents of a `data.frame` in internal format	– `data.frame` representing HFD births input file in internal format – Path to a file to write	Data is written to file in standard HFD births input file format

Table 2. Major functions in package `hfdIndb2ldb`

Function name	*Purpose*	*Main in put data and formats*	*Output data format*
`indbExposure`	Calculation of female population exposure	– Population and death counts in HMD Lexis database format – Territorial adjustment factors in HMD	`data.frame` of female population exposure by Lexis triangles

[3] Use **`install.packages("signal", dependencies=TRUE)`** at the R prompt to install from a CRAN mirror

Function name	*Purpose*	*Main in put data and formats*	*Output data format*
		format – Monthly birth counts in HFD monthly input file format	
`indb.standard`	Standardisation of age/birth order range	`data.frame` representing HFD births input file in internal format	`data.frame` representing HFD births input file in internal format
`unk.unk`	Proportional distribution of births with unknown birth order and age	`data.frame` representing HFD births input file in internal format	`data.frame` representing HFD births input file in internal format
`unk.bo`	Proportional distribution of births with unknown birth order	`data.frame` representing HFD births input file in internal format	`data.frame` representing HFD births input file in internal format
`unk.age`	Proportional distribution of births with unknown age	`data.frame` representing HFD births input file in internal format	`data.frame` representing HFD births input file in internal format
`unk.IPF`	Ensuring marginal totals match after redistribution of births with unknown parameters	`data.frame` representing HFD births input file in internal format	`data.frame` representing HFD births input file in internal format
`split5.births`	Splitting of aggregated and open-ended age categories to single ages	`data.frame` representing HFD births input file in internal format	`data.frame` representing HFD births input file in internal format
`split1.births`	Splitting of single age birth counts to Lexis triangles	`data.frame` representing HFD births input file in internal format	`data.frame` representing HFD births input file in internal format
`indb2ldb.hfd`	Converting HFD input file representation to Lexis database representation	`data.frame` representing HFD births input file in internal format	`data.frame` representing HFD Lexis database births file, optionally also written to disk

Table 3. Description of arguments to functions in package hfdIndbLoad

Function	*Parameter*	*Description*
`indbLoad`	`fname`	Path to file containing HFD births input file
	`LDB`	Boolean indicating whether to filter out rows with LDB ≠ 1. Default: `TRUE`
`indb.print`	`indb`	`data.frame` representing a HFD births input file in internal format (as produced by `indbLoad`)
	`fname`	Path to file to write

Table 4. Description of arguments to major functions in package `hfdIndb2ldb`

Function	*Parameter*	*Description*
`indbExposure`	`pop`	Path to a HMD Lexis database file of female population and death counts
	`monthlyBirths`	Path to a HFD input file of monthly birth counts (`NULL` if not present)
	`tAdj`	Path to a HMD input file of territorial adjustment factors (`NULL` if not present)
	`outfile`	Path to a file to which to write calculated exposures (`NULL` for a default filename, `NA` for no output file)
	`countryCode`	Population code (usually three capital letters, e.g. `"USA"` or `"SWE"`). An attempt will be made to automatically determine this if not supplied
	`verbose`	Boolean indicating whether to report progress (default `TRUE`)
`indb.standard`	`indb`	`data.frame` representing a HFD births input file in internal format (as produced by `indbLoad`)
	`fname`	Path to a HFD births input file (to be read using `indbLoad`)
	`minA`	Minimum age to extend categories down to. Default: 10
	`maxA`	Maximum age to extend categories up to. Default: 60
	`maxO`	Maximum (open) birth order to retain. Default: 5
	`print`	Boolean indicating whether the resulting `data.frame` should be written to file (using `indb.print` with default options). Default: `FALSE`

Function	*Parameter*	*Description*
unk.unk unk.bo unk.age unk.IPF	dta	data.frame representing a HFD births input file in internal format. In most cases this should be have been preprocessed using indb.standard
split5.births	dta	data.frame representing a HFD births input file in internal format. In most cases this should be have been preprocessed using indb.standard and had births with unknown parameters redistributed
	exps	data.frame containing female population exposures by Lexis triangle in HFD Lexis database format as produced by indbExposure
	Lexis	The target shape for splitting to; either VV or RR. Default: RR
	verbose	Boolean indicating whether progress information should be reported. Default: TRUE
split1.births	dta	data.frame representing a HFD births input file in internal format. Births should be recorded only in elementary Lexis elements (1×1) or triangles with no unknowns
	exps	data.frame containing female population exposures by Lexis triangle in HFD Lexis database format
	verbose	Boolean indicating whether progress information should be reported. Default: TRUE
indb2ldb.hfd	indb	data.frame representing a HFD births input file in internal format. Births should be recorded only in Lexis triangles with no unknown age/order
	fname	Path to a file containing a HFD births input file containing births recorded only in Lexis triangles
	LDBfname	Path to file to write Lexis database to. Default: [PopName]birthsTRw.txt where [PopName] is the value of the PopName field
	printLDB	Boolean indicating whether Lexis database should be written to file. Default:

Function	*Parameter*	*Description*
		TRUE
	countryName	Long name of population, Default: value of PopName field.
	maxO	Maximum birth order to include. Default: maximum order included in the input (indb or fname).
	minAge	Minimum age to include. Default: 12
	maxAge	Maximum age to include. Default: 55

Data formats and structures

The main data format used is that of the HFD births input file. This is a comma-separated file whose structure is detailed fully in the HFD data formats documentation[4]. The variables contained in the file are PopName, Area, Year, YearReg, Age, AgeInt, Lexis, BirthOrder, OrderInt, PrevBirth, DurInt, Vital, Births, Access, Note1, Note2, Note3, RefCode and LDB. Missing values are represented with a single dot (`.`).

Files in this format may be read into a `data.frame` using the function `indbLoad`, and such a `data.frame` may be written to file using `indb.print`. These functions also perform conversion to an internal representation of some values which is more convenient for processing, in particular in allowing the field to be used as numeric values. The conversions which are performed are:

- `Missing values (.)` in the fields YearReg and AgeInt are replaced with `NA`.
- `Missing values (.)` in the fields Year are replaced with the value of the YearReg field.
- The values `+ ,-` in the field AgeInt are replaced with 100 and -100 respectively
- The values `UNK` and `TOT` in the field Age are replaced with -1 and 300 respectively.
- The values `TOT` and `UNK` in the field BirthOrder are replaced with the values 0 and -1 respectively.

The format of the HFD monthly births file is also detailed in the HFD data formats documentation.

The format of the territorial adjustments file used in `indbExposure` is found in the Human Mortality Database (HMD) input file documentation[5]: from this file, the factor type `Vx` (ratio of population size) is required where the territorial coverage is changed.

[4] See http://www.humanfertility.org/Docs/formats.pdf

[5] See http://www.mortality.org/Public/Docs/InputDBdoc.pdf

The HMD Lexis file is also a comma-separated file, containing the (unlabelled) columns Year, Age, Triangle, Cohort, Population and Deaths. Triangle contains either 1 or 2, such that Cohort + Age + Triangle – 1 = Year, i.e. 1 for the lower triangle and 2 for the upper triangle. The Population field contains for the lower triangle the population reaching exact age x during the year and for the upper triangle the population aged $\left[x, x+1\right)$ on 1st January: only the latter of these values is required in this case. Missing values are indicated with -1.

The HFD Lexis database file (and corresponding `data.frame`) for female population exposure contains 5 columns; Year, Age, Triangle, Cohort and Exposure, where Triangle is the same as the Triangle for the HMD Lexis database file. This file as produced by `indbExposure` is comma-separated. The file (and corresponding `data.frame`) for births contains the columns Year, Age, Cohort, Total, B1, … Bi_b^+p. The file as produced by `indb2ldb.hfd` is a tabulated text file, with missing values represented using a single point (`.`).

Installation and usage

All packages are written purely in R and may be obtained as a compressed CRAN-style repository archive, included with this Technical Report. Once the contents of the repository archive have been extracted (we assume here to the directory `C:\Temp\hfdPackages`), they may be installed and loaded as follows:

```
> install.packages(c("hfdIndb2ldb"), repos="file:C:/Temp/hfdPackages",
  type="source", dependencies = TRUE)
> library(hfdIndb2ldb)
```

The following examples assume that the files shown in Table 5 have been downloaded and are present in the current working directory.

Table 5. Files used in examples

Filename	*Contents*	*Source*
USAbirths.txt	HFD births input file	HFD country page[6]
USAmonthly.txt	HFD monthly births input file	
fUSA.txt	HMD Lexis file	HMD country ZIP archive[7]
USAtadj.txt	HMD territorial adjustment file	

We first load the births input file and inspect the contents using some graphical summaries (Figure 1 – Figure 3) of maximum birth order, presence of births at unknown age and/or order and presence of data cells requiring splitting:

```
> dta <- indbLoad("USAbirths.txt")
> tail(dta)
```

```
      PopName Area Year YearReg Age AgeInt Lexis BirthOrder OrderInt
63729     USA    1 2006    2006  46      1    RR          0        .
63730     USA    1 2006    2006  47      1    RR          0        .
63731     USA    1 2006    2006  48      1    RR          0        .
63732     USA    1 2006    2006  49      1    RR          0        .
63733     USA    1 2006    2006  50      5    RR          0        .
```

[6] See http://www.humanfertility.org/cgi-bin/countrypage.php?country=USA

[7] See http://www.mortality.org/cgi-bin/hmd/hmd_download.php

```
63734      USA    1 2006    2006 300     NA     .          0        .
      PrevBirth DurInt Vital  Births Access Note1 Note2 Note3 RefCode LDB
63729         .      .     1    1599      0     .     .     .      11   1
63730         .      .     1     859      0     .     .     .      11   1
63731         .      .     1     450      0     .     .     .      11   1
63732         .      .     1     280      0     .     .     .      11   1
63733         .      .     1     494      0     .     .     .      11   1
63734         .      .     1 4265555      0     .     .     .      11   1
```

```
> # determine and plot maximum birth order present in each year (Figure 1)
> plot(aggregate(dta$BirthOrder, list(Year = dta$Year), max), ylab =
  expression(i[max]), type = "h", ylim = c(0, max(dta$BirthOrder) + 1), yaxs
  = "i", main = "Maximum birth order by year", panel.first = grid(lty =
  "solid"))
```

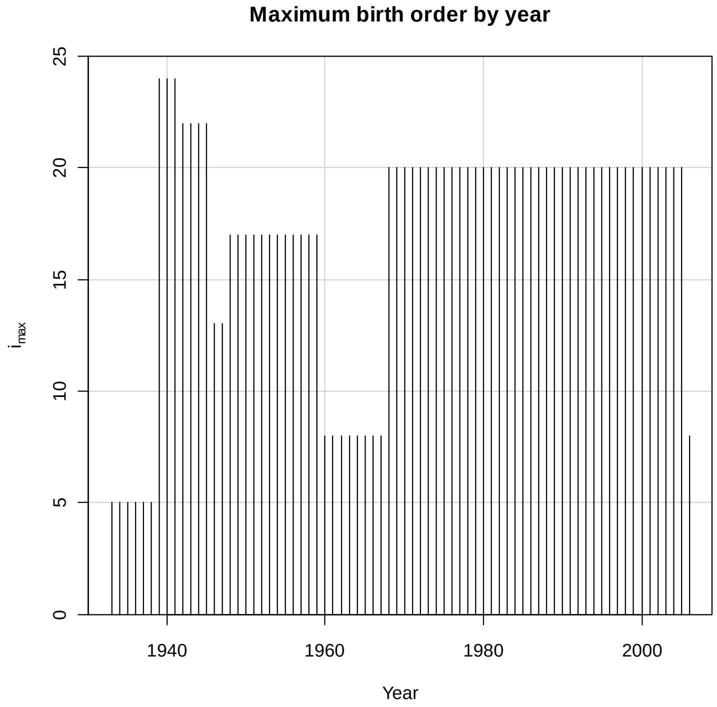

Figure 1. Maximum birth order by year in birth data for USA

```
> # now aggregate order 5+ and standardise age range
> dta <- indb.standard(dta)
> # determine and plot types of unknowns present (Figure 2)
> plot(rbind(cbind(unique(subset(dta, BirthOrder == -1, "Year")), Unknown =
  3), cbind(unique(subset(dta, Age == -1 & BirthOrder == -1, "Year")),
  Unknown = 2), cbind(unique(subset(dta, Age == -1, "Year")), Unknown = 1)),
  pch = 4, yaxt = "n", ylim = c(1, 3), main = "Types of unknown present by
  year", panel.first = {grid(ny = NA, lty = "solid"); abline(h = 1:3, col =
  "lightgrey"); axis(2, at=1:3, label=c("Age", "Age & Order", "Order"))})
```

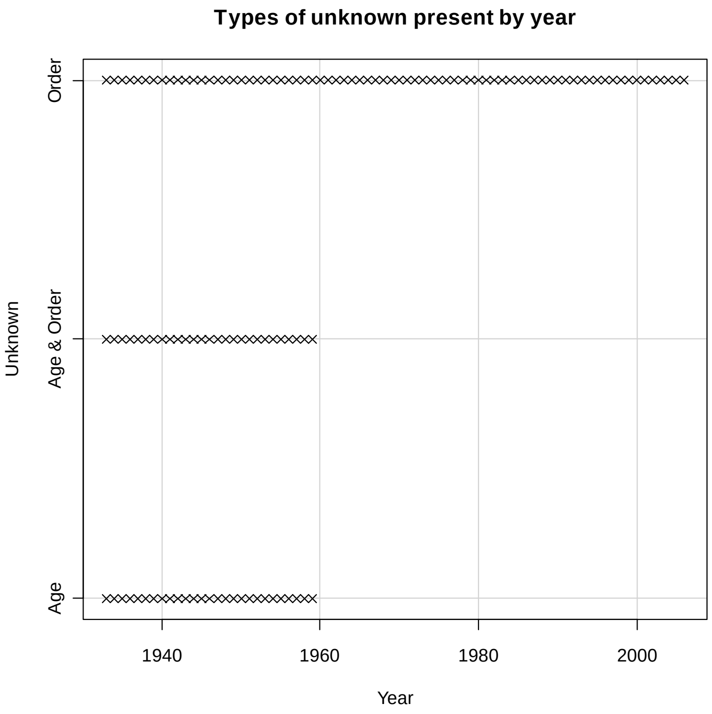

Figure 2. Unknown parameters by year in births data for USA

```
> # determine and plot splitting required in each year (Figure 3)
> plot(cbind(rep(unique(dta[!is.na(dta$AgeInt), "Year"]), each = 4),
  unlist(by(dta[!is.na(dta$AgeInt), ], dta$Year[!is.na(dta$AgeInt)],
  function(x){(1:4) * c(any(x$AgeInt == -100), !any(x$Lexis %in% c("TU",
  "TL")), any(x$AgeInt == 5), any(x$AgeInt == 100))}))), ylim = c(1, 4),
  xlab = "Year", ylab = "Splitting required", yaxt = "n", pch = 4,
  panel.first = {grid(ny = NA, lty = "solid"); abline(h = 1:4, col =
  "lightgrey"); axis(2, at=1:4, label=c("Open Lower", "1-Year", "5-Year",
  "Open Upper"))}, main = "Splitting required by year")
```

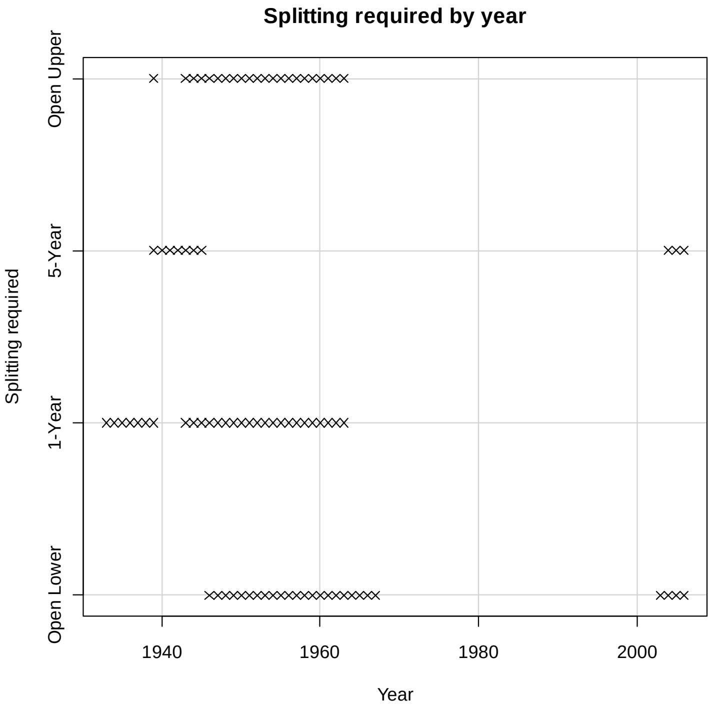

Figure 3. Splitting required by year in births data for USA

Having established which processing steps are necessary, we can begin processing the data by estimating the female population exposure by year and age:

```
> exps <- indbExposure("fUSA.txt", "USAmonthly.txt", "USAtadj.txt", NA)
```

We then perform distribution of the births with unknown parameters. The example of 1940 shows the need for the IPF procedure as the sum of order-specific births in each age category in this year after redistribution (column Sum) does not match the required total births (column Total), with some significant differences (column Diff):

```
> dta <- unk.unk(dta)
> dta <- unk.bo(dta)
> dta <- unk.age(dta)
> ageSumsCheck <- by(dta, dta$Year, function(dta){sums <- tapply(dta$Births,
  list(dta$Age, dta$BirthOrder), sum); sums <- sums[-nrow(sums), ]; res <-
  data.frame(Age = rownames(sums), Total = sums[, 1], Sum = rowSums(sums[,
  2:6])); res$Diff <- round(res$Sum - res$Total, 3); rownames(res) <- NULL;
  return(res)}, simplify = FALSE)
> head(ageSumsCheck$`1940`)
```

```
  Age      Total       Sum    Diff
1  10   3262.912   3262.09  -0.822
2  15 301292.937 301227.20 -65.732
3  20 739776.461 739683.91 -92.555
4  25 647039.424 647042.52   3.093
5  30 400951.517 401011.33  59.813
6  35 201485.087 201550.21  65.118
```

We therefore apply the IPF procedure to ensure that these marginal totals match and inspect the results to ensure that the differences are now at acceptable levels:

```
> dta <- unk.IPF(dta)
> ageSumsCheck <- by(dta, dta$Year, function(dta){sums <- tapply(dta$Births,
  list(dta$Age, dta$BirthOrder), sum); sums <- sums[-nrow(sums), ]; res <-
  data.frame(Age = rownames(sums), Total = sums[, 1], Sum = rowSums(sums[,
  2:6])); res$Diff <- round(res$Sum - res$Total, 3); rownames(res) <- NULL;
  return(res)}, simplify = FALSE)
> head(ageSumsCheck$`1940`) # note Diff is rounded to 3 decimal places
```

```
  Age      Total        Sum Diff
1  10   3262.912   3262.912    0
2  15 301292.937 301292.937    0
3  20 739776.461 739776.461    0
4  25 647039.424 647039.424    0
5  30 400951.517 400951.517    0
6  35 201485.087 201485.087    0
```

We can now perform splitting of 5-year and open-interval age categories into single age categories and then single age categories to Lexis triangles:

```
> dta <- split5.births(dta, exps)
> dta <- split1.births(dta, exps)
```

Finally, we convert the HFD input format data to the more convenient HFD Lexis and output database format with data arranged by Year, Age and Cohort:

```
> dta <- indb2ldb.hfd(dta, printLDB = FALSE)
> head(dta)
```

```
  Year Age Cohort      Total         B1       B2 B3 B4 B5p
1 1933  12   1921    6.51388    6.51388 0.000000  0  0   0
2 1933  12   1920   40.33961   40.33961 0.000000  0  0   0
3 1933  13   1920  198.40026  198.40026 0.000000  0  0   0
4 1933  13   1919  333.28975  333.28975 0.000000  0  0   0
5 1933  14   1919  850.05280  848.87639 1.176412  0  0   0
6 1933  14   1918 1270.58279 1262.63439 7.948402  0  0   0
```

Summary

This technical report has presented R packages which allow the processing of births data which may be gathered in a varying age and birth order structure to produce the HFD Lexis database, which contains standardised files with birth counts classified by birth order and mother's year of birth and age. Female population exposure by age and birth cohort may also be estimated, and such estimates are in any case required for splitting aggregated age categories.

The major functions and their parameters were described before their usage was illustrated by showing the end-to-end processing for a single country. Further functions are available for special cases: for details of these and for more detail on the functions described here, the online help in R may be used and the source code may be inspected.

References

Bishop Y., Fienberg S., Holland P. (1975). Discrete Multivariate Analysis: Theory and Practice. MIT University Press.

Calot, G., Sardon, J.-P. (2003). "Methodology for the calculation of Eurostat's demographic indicators. Detailed report by the European Demographic Observatory (EDO)". Population and social conditions 3/2003/F/no 26, 146 p.

Fienberg S. (1970). An iterative procedure for estimation in contingency tables, The Annals of Mathematical Statistics, Vol. 41, N. 3, 907-917.

Jasilioniene, A., Jdanov, D. A., Sobotka, T., Andreev, E. M., Zeman, K., Nash, E. J., and Shkolnikov, V. M. (with contributions of Goldstein, J., Philipov, D. and Rodriguez, G.) (2010). Methods Protocol for the Human Fertility Database. URL http://www.humanfertility.org/Docs/methods.pdf

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Wilmoth, J. R., Andreev, K., Jdanov, D., Glei, D.A. with the assistance of C. Boe, M. Bubenheim, D. Philipov, V. Shkolnikov, P. Vachon. (2005). Methods Protocol for the Human Mortality Database. URL http://www.mortality.org/Public/Docs/MethodsProtocol.pdf.