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 2011-001
JANUARY 2011

A collection of "R" packages for the production of period fertility tables and some summary fertility indicators

Edward Nash (nash@demogr.mpg.de)
Aiva Jasilioniene (jasilioniene@demogr.mpg.de)
Evgeny Andreev (andreev@demogr.mpg.de)
Kryštof Zeman

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

A collection of “R” packages for the production of period fertility tables and some summary fertility indicators

by Edward Nash, Aiva Jasilioniene, Evgeny Andreev and Kryštof Zeman

Abstract

As part of the Human Fertility Database project, a standardised methodology has been developed for the production of period and cohort fertility tables and other fertility indicators. Packages have been implemented for the free statistical computing language and environment “R” in order to allow other researchers easy access to these methods. A previous Technical Report (TR-2010-007, Nash et al. 2010) introduced the package for the production of cohort fertility tables. This Technical Report introduces four further packages which, when used together, enable the calculation of five basic fertility indicators (age-specific and cumulative fertility rates, total fertility rate, mean age at birth and cohort parity progression ratios) and period fertility tables from the basic inputs of birth counts, estimated female exposure and, where available, census or register data of female population by parity.

Keywords: period fertility table, period fertility indicators, age-specific fertility rate, cumulative fertility rate, total fertility rate, mean age at birth, parity progression ratios, period fertility index of total fertility, Human Fertility Database, R

Background

In the Human Fertility Database (HFD)[1], a joint project of the Max Planck Institute for Demographic Research (MPIDR) and the Vienna Institute of Demography (VID), based at the MPIDR, fertility tables are produced by age and parity for both birth cohorts and periods. Technical Report TR-2010-007 by Nash et al. (2010) considered the production of cohort fertility tables. This report is devoted to summary fertility indicators and period fertility tables. The functions required to produce these have been structured in a collection of distinct but interrelated packages in “R” to improve maintainability and to facilitate flexible use.

This report first reviews the methodology detailed in the HFD methods protocol before describing the implementation of these methods in “R”. Finally, the use of the functions provided is illustrated with some examples.

Notation

The HFD considers the reproductive span between age 12- ($x_{\min}$) to age 55+ ($x_{\max}$) and birth data for orders 1…5+, allowing the consideration of parities 0…4+. To enable the HFD methodology to be applied flexibly by other users who may wish to consider a greater or lesser age or parity range, the HFD Methods Protocol has been generalised in terms of age range and birth order 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

[1] http://www.humanfertility.org

highest (open-interval) birth order and i_l^+ for the highest (open-interval) parity, where $i_l^+ = i_b^+ - 1$.

The notation used in the HFD Methods Protocol is also simplified here in the handling of cohort and period data: no distinction is made where the formula is in principle identical and so e.g. $B_i(x)$ in this Technical Report is used for all of $B_i(x,t,c)$, $B_i(x,t)$, $B_i(t,c)$ and $B_i(x,c)$ for birth counts by birth order and age by Lexis triangles (age in completed years, year, cohort), rectangles (age in completed years, year), vertical parallelograms (year, cohort or age reached during year) and horizontal parallelograms (age in completed years, cohort) respectively. Furthermore, no distinction is made between total births and rates and order-specific births and rates: where appropriate (e.g. for age-specific fertility rates) the total values may be substituted for the order-specific ones.

Summary fertility indicators based on unconditional rates

Five summary fertility indicators based on unconditional fertility rates are currently featured in the Human Fertility Database; age-specific fertility rate (ASFR)[2], cumulative fertility rate (CFR), total fertility rate (TFR), mean age at birth (MAB) and cohort parity progression ratios (PPR). Table 1 summarises the Lexis shapes for which each of these indicators is calculated. Each of these, except parity progression ratios, may be calculated for each birth order individually or for all birth orders combined.

Table 1. Summary indicators based on unconditional rates featured in the HFD

			Lexis shape			
			Triangle	**Rectangle**	**Horizontal parallelogram**	**Vertical parallelogram**
Indicator	**Abbr.**	**Notation**	**TR**	**RR**	**VH**	**VV**
Age-specific fertility rate	ASFR	$f_i(x)$	✓[3]	✓	✓	✓
Cumulative fertility rate	CFR	$CFR_i(x)$		✓	✓	
Total fertility rate	TFR	TFR_i		✓	✓	
Mean age at birth	MAB	MAB_i		✓	✓	
Parity progression ratios	PPR	$PPR_{i-1,i}$			✓	

The formulae used for calculating these basic summary indicators are:

$$f_i(x) = \frac{B_i(x)}{E(x)} \quad (1)$$

[2] Unconditional rates, also called incidence rates, rates of the second kind, frequencies, densities or reduced rates (Bongaarts and Feeney, 2006; Kohler and Ortega, 2002; Wunsch, 2006).
[3] Not currently featured in the Human Fertility Database.

$$CFR_i(x) = \sum_{z=x_{\min}}^{x-1} f_i(z) \qquad (2)$$

$$TFR_i = \sum_{x=x_{\min}}^{x_{\max}} f_i(x) \qquad (3)$$

$$MAB_i = \frac{\sum_{x=x_{\min}}^{x_{\max}} \bar{x} \cdot f_i(x)}{\sum_{x=x_{\min}}^{x_{\max}} f_i(x)} \qquad (4)$$

where $B_i(x)$ is births of order i at age x, $E(x)$ the female population exposure at age x and $\bar{x} = x + a(x)$ with a being the average share of the age interval $[x, x+1)$ lived before giving birth, for the HFD assumed to be 0.5 for all ages.

$$PPR_{0,1}(c) = TFR_1(c) \qquad (5)$$

$$PPR_{i-1,i}(c) = \frac{TFR_i(c)}{TFR_{i-1}(c)}, \text{ for } 1 < i < i_l^+ \qquad (6)$$

where c is the cohort.

Period fertility tables

The key input in period fertility tables is the age- and parity-specific distribution of the female population of reproductive age. These distributions may be obtained from cohort fertility tables, “golden” censuses that provide the initial parity distribution in one base year or directly from population censuses or registers. In the latter case, the fertility tables are census or register-based. Thus, period fertility tables can be calculated only if either a sufficiently long series of order-specific data is available to calculate a cohort fertility table or population census or register data of the female population by parity is available. The packages described in this report allow the use of any of these methods for generating the distribution, depending on the available data.

The production of period fertility tables according to the HFD methodology may be separated into two steps:

1. The mid-year estimates of the age- and parity-specific distribution of women are produced, and thereafter the age- and parity-specific female population exposure for each period and age to be included in the table.
2. The period fertility table is produced based on the estimated age- and parity-specific female population exposure and the observed age- and order-specific period birth counts.

Once the period fertility table has been produced, summary indicators such as the table mean age at birth and the period fertility index (PATFR)[4] may be produced based on this. Each of these stages will now be considered separately.

[4] Called the summary index of fertility controlling for age and parity by Rallu and Toulemon (1994)

Estimation of mid-year age/parity distributions

The mid-year estimate of the age/parity distribution $w_i(x,t)$ consists of relative weights which are used to produce parity-specific exposure estimates through multiplying with the estimated total female population exposure. As such, for each period t and age x:

$$\sum_{i=0}^{i_l^+} w_i(x,t) = 1 \qquad (7)$$

These weights are calculated from cohort table populations at each parity, $l_i(x,c)$, which in turn may be obtained from a cohort fertility table and/or based upon census or register data. The weights are calculated according to the following formulae (note that $l_0(x_{\min}, t - x_{\max})$ in equations 10 and 11 is the radix of the cohort fertility table):

$$w_i(x_{\min}, t) = 1, \text{ for } i = 0 \qquad (8)$$

$$w_i(x_{\min}, t) = 0, \text{ for } i = 1 \ldots i_l^+ \qquad (9)$$

$$w_i(x,t) = \frac{l_i(x, t-x) + l_i(x+1, t-x-1)}{2 \cdot l_0(x_{\min}, t - x_{\min})}, \text{ for } x_{\min} < x < x_{\max} \qquad (10)$$

$$w_i(x_{\max}, t) = \frac{l_i(x, t-x)}{l_0(x_{\min}, t - x_{\min})} \qquad (11)$$

The potential methods for producing $l_i(x,c)$ from census/register data are now considered.

Producing $l_i(x,c)$ from census or register data

The population parity weights on the census date may be calculated as:

$$w_i(x, t_{cens}^T) = \frac{P_i^{cens}(x)}{P_{TOT}^{cens}(x)} \qquad (12)$$

where P_i^{cens} is the female population at parity i on the date of the census.

If the census date T is not 1st January or 31st December (in which case the weights are taken as representing the situation on 1st January of the following year), they must first be adjusted to 1st January using interpolation of the age-specific fertility rates as described in section 5.1 of the HFD Methods Protocol. These weights on 1st January are denoted here as $w_i(x, t_{cens}^{Jan1})$. Where the weights on 1st January of the subsequent year are not known, they may also be estimated using age-specific fertility rate data as described in section 5.2 of the HFD Methods Protocol, giving $w_i(x, [t_{cens}+1]^{Jan1})$.

Cohort fertility table population by parity values $l_i(x,c)$ may then be calculated for the cohorts enumerated in the census or register:

$$l_i(x, t_{cens} - x) = radix \cdot w^C{}_i(x, t_{cens} - x) \qquad (13)$$

Where the radix is the standardised cohort fertility table population size (10000 for the HFD) and the cohort parity weightings $w^C{}_i(x, t_{cens} - x)$ is calculated as:

$$w^C{}_i(x, t_{cens} - x) = \frac{w_i(x-1, t_{cens}^{Jan1}) + w_i(x, [t_{cens}+1]^{Jan1})}{2} \quad (14)$$

These values of $l_i(x,c)$ may then be used in the equations (8) to (11) to produce the mid-year estimate of the age-parity distribution $w_i(x,t)$. Additionally, the series of $l_i(x,c)$ values may be continued by using the cohort fertility table births, allowing the production of $w_i(x,t)$ for years subsequent to the census, known as the "golden census" method. In this way, period fertility tables may be produced for periods for which parity weightings cannot be estimated using cumulation of cohort fertility alone.

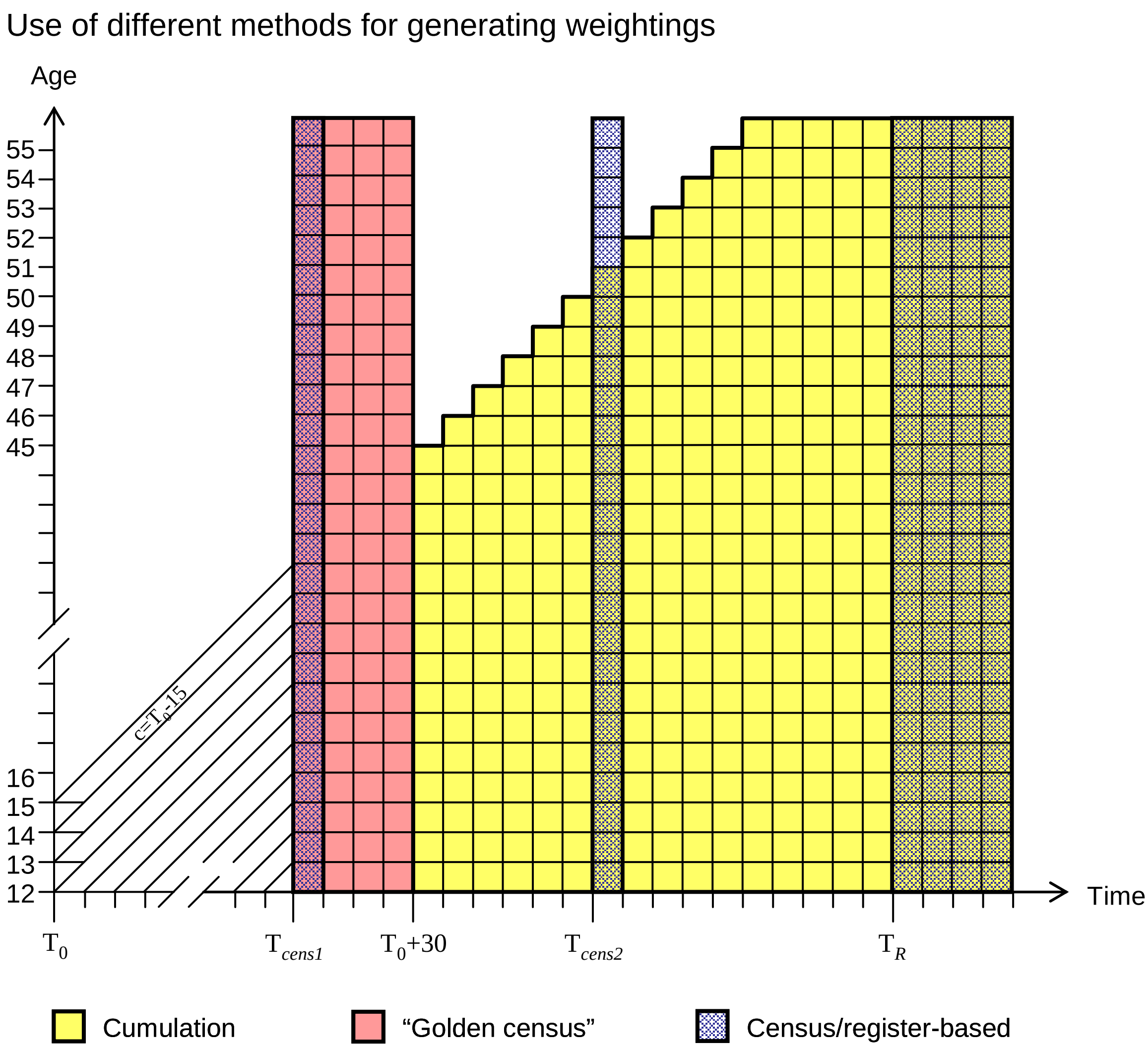

Figure 1. Lexis regions for which weights may be generated using different methods

Figure 1 illustrates Lexis regions for which each method of generating population by parity weightings may typically be used. T_0 is the year at which records of births by biological birth order commence. T_{cens1} and T_{cens2} are years in which a population census was recorded, with T_R marking the start of a population register. Requiring a cohort to have been observed from age 15 to age 45, the cumulation of fertility rates may be used to produce weightings from year $T_0 + 30$, although these weightings do not initially cover the full age range. Using the census from T_{cens1} as a "golden census” allows the production of a set of weightings for a longer period of time. Since the method of cumulation of fertility rates is however generally preferred to the “golden census” method, the weightings switch to the cumulative ones in year $T_0 + 30$ (note that the R packages allow for this preference to be reversed and the entire table to be produced cumulating forward from the “golden census”). Census/register based period fertility tables may be generated for the years T_{cens1}, T_{cens2} and from T_R onwards.

Calculation of period fertility table

The period fertility table as specified for the Human Fertility Database contains 7 indicators for each birth order/parity, summarised in Table 2. Typically, these columns are arranged by order/parity and then indicator ($w_0(x)$, $m_1(x)$, …, $L_{i_l^+}(x)$, $Sb_{i_b^+}(x)$), with two initial columns (Year, x).

Table 2. Summary of indicators included in the HFD period fertility tables

Notation	*Description*
$w_{i-1}(x)$	Relative distribution of female population exposure by parity
$m_i(x)$	Age- and parity- specific fertility rates[5] in age interval $[x, x+1)$
$q_i(x)$	Conditional probability of giving i^{th} birth in age interval $[x, x+1)$
$l_{i-1}(x)$	Table population by parity $i-1$ at age x
$b_i(x)$	Table number of births of order i in age interval $[x, x+1)$
$L_{i-1}(x)$	Table population exposure at parity i-1 in age interval $[x, x+1)$
$Sb_i(x)$	Cumulative table births of order i by age x

The initial basis for the production of the indicators is the female population exposure $E(x)$ by calendar year and age and the relative distribution of female population distribution by parity $w_{i-1}(x)$, whose production was described in the previous section. From these, the parity-specific population exposure may be calculated:

$$E_{i-1}(x) = E(x) \cdot w_{i-1}(x), \; i = 1 \ldots i_b^+ \qquad (15)$$

Based on this estimated parity-specific population exposures and the observed period births by order, the age- and parity specific fertility rates may be trivially calculated:

$$m_i(x) = \frac{B_i(x)}{E_{i-1}(x)}, \; i = 1 \ldots i_b^+ \qquad (16)$$

The probability of an i^{th} birth, which forms the basis for the remainder of the indicators in the period fertility table, may be calculated from these conditional rates together with the average share of the age interval $[x, x+1)$ lived before giving birth, $a(x)$, assumed in the HFD to equal 0.5 for all ages and periods:

[5] Also called conditional rates, occurrence-exposure rates, rates of the first kind, intensities, hazard rates or risks (Bongaarts and Feeney, 2006; Kohler and Ortega, 2002; Wunsch, 2006).

$$q_i(x) = \frac{m_i(x)}{1+[1-a(x)]\cdot m_i(x)},\ i = 1\ldots i_b^+ \quad (17)$$

The calculation of $l_{i-1}(x)$, $b_i(x)$ and $L_{i-1}(x)$ is interlinked, with the values of each indicator at age x required in order to calculate the indicators at age $x+1$:

$$l_0(x_{\min}) = radix \text{ (10000 for the HFD)} \quad (18)$$

$$l_i(x_{\min}) = 0 \text{ for } i = 1\ldots i_l^+ \quad (19)$$

$$l_0(x) = l_0(x-1)\cdot[1-q_1(x-1)] \quad (20)$$

$$l_i(x) = l_i(x-1) - b_{i+1}(x-1) + L_{i-1}(x-1)\cdot m_i(x-1) \text{ for } i = 1\ldots[i_l^+ - 1] \quad (21)$$

$$l_{i_l^+}(x) = l_{i_l^+}(x-1) + L_{i_l^+ -1}(x-1)\cdot m_{i_b^+}(x-1) \quad (22)$$

$$b_i(x) = L_{i-1}(x)\cdot m_i(x) \text{ for } i = 1\ldots i_b^+ \quad (23)$$

$$L_0(x) = l_0(x) - l_0(x)\cdot q_1(x)\cdot[1-a(x)] \quad (24)$$

$$L_i(x) = l_i(x) + l_{i-1}(x)\cdot q_i(x)\cdot[1-a(x)] - l_i(x)\cdot q_{i+1}(x)\cdot[1-a(x)] \text{ for } i = 1\ldots[i_l^+ - 1] \quad (25)$$

$$L_{i_l^+}(x) = l_{i_l^+}(x) + l_{[i_l^+ -1]}(x)\cdot q_{i_b^+}(x)\cdot[1-a(x)] \quad (26)$$

The calculation of cumulative births may be trivially calculated from the table births:

$$Sb_i(x) = \sum_{z=x_{\min}}^{x-1} b_i(z) \quad (27)$$

Summary indicators based on the period fertility table

Two summary indicators based on the period fertility table are featured in the Human Fertility Database; the period fertility index of total fertility (PATFR) and the table mean age at birth (TMAB). These are based on the table births, and may be calculated for each individual birth order or for all birth orders combined using the following formulae:

$$PATFR_i = \frac{\sum_{x=x_{\min}}^{x_{\max}} b_i(x)}{10{,}000} \quad (28)$$

$$TMAB_i = \frac{\sum_{x=x_{\min}}^{x_{\max}} \bar{x}\cdot b_i(x)}{\sum_{x=x_{\min}}^{x_{\max}} b_i(x)} \quad (29)$$

where $\bar{x} = x + a(x)$ with a being the average share of the age interval $[x, x+1)$ lived before giving birth, for the HFD assumed to be 0.5 for all ages and periods.

The “R” packages `hfdBasicIndicators`, `hfdParity`, `hfdPeriodFertilityTable` and `hfdLexisManipulation`

All calculations for the Human Fertility Database are programmed in R[6]; a number of the functions used which may be of more general interest are being made publically available.

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`*.

Contents

To provide a clear structure and enable flexible re-use, the functions implemented to perform the calculations described here are split into a number of R packages. Due to the package mechanisms provided by R, the user must not install and load each package individually: the required packages may be automatically loaded by the R system using the commands as demonstrated in the section “Installation and basic usage” so that only the package `hfdPeriodFertilityTable` must be explicitly installed and loaded.

- `hfdBasicIndicators` provides functions for the production of the five indicators described in the section “Summary fertility indicators”.
- `hfdParity` provides functions for the manipulation of parity datasets to produce the relative distribution of female population exposure by parity $w_{i-1}(x)$ by cumulation of cohort fertility and/or from register/census data.
- `hfdPeriodFertilityTable` provides functions for production of period fertility tables, individual indicators and summary indicators based on the PFT.
- `hfdLexisManipulation` provides a number of functions which are used for manipulating data held in Lexis shapes and is a dependency for the packages `hfdBasicIndicators` and `hfdParity`. Although these functions are primarily utility functions used internally, they may also be called directly if required. The main function of end-user interest is `produceLexisShapes` which aggregates a `data.frame` of Lexis triangle data to Lexis squares or parallelograms.

Additionally, the package `hfdCohortFertilityTable`, described in a previous technical report (Nash et al., 2010), is required by the packages `hfdParity` and `hfdPeriodFertilityTable` as well as the external package `abind` which may be obtained from a CRAN mirror if not already installed in R.

The major functions provided by each of these packages are summarised in Tables 3 to 6 in Appendix A. Furthermore, the packages `hfdLexisManipulation`, `hfdParity`, and `hfdPeriodFertilityTable` contain some sample data which is used in the package examples and may be used as a basis for experimenting with the available functionality. These datasets are described in Table 9 in Appendix B. Additionally, `hfdPeriodFertilityTable` provides a utility function

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

(`array2dataframe.pft`) for reformatting the 3D `array` (Year × Age × Order) outputs from individual functions to a `data.frame` analogous to the usual HFD ouptut formats with data arranged in "long format" with factors Year and Age followed by data columns. More information on the data formats is provided in the next section.

The arguments to all main functions are described in detail in Tables 7 and 8. Further information and detailed descriptions of other functions included in the packages is available in the online documentation which may be retrieved by typing e.g. **`?calculate.pft`** at the R prompt.

Most of the default values for function parameters correspond to those used for the Human Fertility Database, but it is possible for users to specify the majority of parameters relevant to the processing, such as the handling of records with unknown parity and the range of ages for which cohorts must be observed. Additionally, there is no restriction on the range of ages or birth orders which can be processed: the ranges from the input data are used, and so fertility tables and other indicators may be produced for higher birth orders than 5, or for a lower range (e.g. to only 4+), or for a wider or narrower range of ages than that featured in the HFD. It is however generally the responsibility of the package user to ensure that the range of ages and birth orders/parities in the input data is consistent for each function.

Data formats and structures

For convenience, all packages work with standard R data types, particularly `data.frame` and `array`, avoiding the need for conversions to/from specialist data types. NA is generally used to represent all missing or implausible values.

In general, where a `data.frame` is accepted as a parameter, the expected structure matches that of the corresponding Human Fertility Database output file (e.g. for births, exposures or rates). This can also be seen in the example data provided with the packages, e.g.:

```
> head(someBirths.TR)

  Year Age Cohort Total   B1 B2 B3 B4 B5p
1 1970  12   1958   130   81 29 14  6   0
2 1970  12   1957    90   76  9  4  1   0
3 1970  13   1957   234  223  8  2  1   0
4 1970  13   1956   391  379  9  2  1   0
5 1970  14   1956  1179 1139 36  3  1   0
6 1970  14   1955  1952 1882 64  5  1   0
```

```
> tail(someBirths.RR)

    Year Age Total B1 B2 B3 B4 B5p
435 1999  50    27  9  7  5  2   4
436 1999  51    17  6  3  3  1   4
437 1999  52    12  3  4  3  2   0
438 1999  53     3  1  0  0  1   1
439 1999  54     5  1  1  1  1   1
440 1999 55+     1  0  0  1  0   0
```

Data files from the HFD "output database" (counts, rates, indicators and fertility tables sections of each country page) may be easily read into a `data.frame` in R which may then be used directly as input to many functions using the command read.table with the options header = TRUE, skip = 2, na.strings = ".", e.g.:[7]

```
> usaBirths <- read.table("USAbirthsTRbo.txt", header = TRUE, skip = 2,
  na.strings = ".")
```

Additionally, the package hfdLexisManipulation contains a convenience function read.table.hfd which may be used to read such tables and has options to automatically remove suffixes + and – for open interval age categories and to trim incomplete rows (where one or more factors has a NA value). This function may be called e.g.

```
> usaBirths <- read.table.hfd("USAbirthsRR.txt", dropSuffixes = TRUE)
```

Furthermore, many functions can read input directly from file: by default, read.table is called on this file with the arguments for HFD data, but users may supply further arguments appropriate to other file formats. Alternatively, existing data.frames may be reformatted to match the usual HFD format

Additionally, 3D arrays can be used as input for many functions. These arrays must have named dimensions, the order of which matches the order in which columns appear in HFD output datasets: Year, Age, Order for Lexis squares, Cohort, Age, Order for horizontal parallelograms and Cohort, Year, Order for vertical parallelograms. Such an array may be generated in R by e.g.

```
> randomBirths.RR <- array(runif(1860), dim = c(20, 31, 3), dimnames =
  list(Year = 1990:2009, Age = c(15:45), Order = c(1:3)))
```

Note that in general, the subscripts + and - used to indicate open-intervals are ignored for the purpose of calculation, although in most cases they will be propagated to the output if present in the input.

For census and register data, the basic input format expected is that of the HFD "Female population by age and parity", described fully in the HFD data formats documentation[8]. This is a comma-separated file which may be read into an R data.frame using the command read.csv, e.g.:

```
> read.csv("CZEparity.txt", header = TRUE, na.strings = ".")
```

Either the path to the file or the data.frame representation may be used as input. Note that strictly only the variables Day, Month, Year, Age, AgeInt, Cohort, CohInt, Parity, ParityInt, Population and LDB are required, and that generally either Age and AgeInt or Cohort and CohInt will be empty (indicated with . in the file and NA in a data.frame).

Installation and basic 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("hfdPeriodFertilityTable"),
  repos="file:C:/Temp/hfdPackages",type="source", dependencies = TRUE)
```

[7] This and other examples assume that the relevant files have been downloaded from the HFD to the R working directory, which may be located by entering **getwd()** at the R prompt.
[8] See http://www.humanfertility.org/Docs/formats.pdf

```
> library(hfdPeriodFertilityTable)
```

Note that due to the defined dependencies between the packages, these two commands will actually install and load all five packages described here. Note that it may be necessary to first install the external package `abind` from a CRAN mirror, if it is not already installed in your R system, using the command:

```
> install.packages("abind")
```

For simplicity in this section, the example data from the packages is used as input: the next section shows an example using the published HFD output files. Once packages are installed, the sample data can be loaded using the `data` function, e.g.:

```
> data(someBirths.TR, someExpos.TR)
```

The Lexis triangle data may be aggregated to another shape using `produceLexisShapes`, e.g.

```
> someBirths.VV <- produceLexisShapes(someBirths.TR, "VV", "births")
> someExpos.VV <- produceLexisShapes(someBirths.TR, "VV", "births")
```

Rates and indicators may then be calculated using these aggregated datasets, e.g.:

```
> someAsfr.VV <- calculate.asfr(someBirths.VV, someExpos.VV, "VV")
```

Inputs may also be read from file for many functions, e.g.

```
> data(sweRegisterPath)
> sweRegisterPath # path depends on where the package was installed to

[1] "U:/R/R/win-library/2.11/hfdParity/sampleFiles/SWEparity-extract.txt"

> sweWeights <- parity.census.weights(sweRegisterPath)
> sweWeights <- parity.1Jan.weights(sweWeights, "SWEasfrVVbo.txt")
```

The different methods of calculating the relative distribution of female population exposure by parity produce weightings, and thus period fertility tables, for different ranges of year:

```
> cum <- calculateAgeParityDistribution.cumulative(someBirths.TR,
  someExpos.TR)
> gol <- calculateAgeParityDistribution.golden(someBirths.TR, someExpos.TR,
  sweWeights, 1981)
> cen <- calculateAgeParityDistribution.census(someBirths.TR, someExpos.TR,
  sweWeights)
> reg <- calculateAgeParityDistribution.register(someBirths.TR,
  someExpos.TR, sweWeights)
> dimnames(cum)$Year

[1] "2001" "2002" "2003" "2004" "2005" "2006"

> dimnames(gol)$Year

 [1] "1981" "1982" "1983" "1984" "1985" "1986" "1987" "1988" "1989" "1990"
[11] "1991" "1992" "1993" "1994" "1995" "1996" "1997" "1998" "1999" "2000"
[21] "2001" "2002" "2003" "2004" "2005" "2006"

> dimnames(cen)$Year

[1] "1981" "1982" "1983" "1984" "1985"

> dimnames(reg)$Year

[1] "1981" "1982" "1983" "1984" "1985"
```

Despite in this case giving the same range of years, the ‘census’ and ‘register’ methods produce slightly different weightings: the ‘census’ method estimates the parity distribution on 1st January of the subsequent year based on fertility rates whereas the ‘register’ method uses the following year’s register data:

```
> head(round(reg["1981",,]-cen["1981",,], 5))
```

```
    Parity
Age       w0x      w1x      w2x      w3x     w4x
  12 0.00000  0.00000  0.00000  0.00000  0e+00
  13 0.00039 -0.00038 -0.00001  0.00000  0e+00
  14 0.00157 -0.00151 -0.00006  0.00000  0e+00
  15 0.00435 -0.00405 -0.00029 -0.00002  0e+00
  16 0.00819 -0.00717 -0.00094 -0.00008 -1e-05
  17 0.01215 -0.00982 -0.00204 -0.00026 -3e-05
```

Once the parity distribution has been calculated using a suitable method, the female population exposure by parity or a whole period fertility table and subsequently summary indicators may be produced:

```
> E <- calculate.E.pft(gol, "SWEexposRRbo.txt")
> pft <- calculate.pft("SWEbirthsRRbo.txt", "SWEexposRR.txt", gol,
  returnType = "list")
> patfr <- calculate.patfr(pft$b) # only b is needed, not the whole PFT!
```

For functions which produce output as 3D arrays, these may be converted to data frames, which may be more convenient for many operations:

```
> class(E)
```

```
[1] "array"
```

```
> head(E)
```

```
[1] 54127.36 52509.64 54707.98 55147.26 54659.28 54257.87
```

```
> E <- array2dataframe.pft(E)
> head(E)
```

```
  Year  x      E0x        E1x        E2x       E3x E4x
1 1981 12 54127.36   0.000000  0.0000000 0.0000000   0
2 1981 13 57222.18   2.624344  0.0000000 0.0000000   0
3 1981 14 60193.89  13.470179  0.2492243 0.0000000   0
4 1981 15 59624.73  54.569440  1.4744534 0.0000000   0
5 1981 16 60727.46 234.272042  3.9772044 0.0000000   0
6 1981 17 57251.45 706.508349 27.8051673 0.8055999   0
```

In the next section, more examples illustrate in depth how various functions may be applied to perform further analysis and plotting of results, and how the published files from the HFD website may be used as input.

Further examples

Comparing TFR and PATFR using published files from the HFD

The period total fertility rate and the period fertility index of total fertility are both indicators with a similar interpretation, despite different methods of calculation. The following script produces both indicators and plots the graph shown in Figure 2 allowing their values to be compared. In this graph both the indicators are shown for both total and birth order 1: it can be seen that particularly for order one, the TFR values vary rather more than the corresponding PATFR values.

For this example, the births and exposures by Lexis triangles (`SWEbirthsTRbo.txt` and `SWEexposTR.txt`) and parity input (`SWEparity.txt`) files from the HFD for Sweden are used. Note that the HFD Lexis triangle output files for births contain aggregated ages 12- and 55+: in this example we therefore trim the datasets to ages 13-54 (closed interval). This also demonstrates how the functions may be used with data with a different age range to the HFD. In the case of Sweden there are in any case no births recorded at ages 12- or 55+, as is shown by the check in the following script.

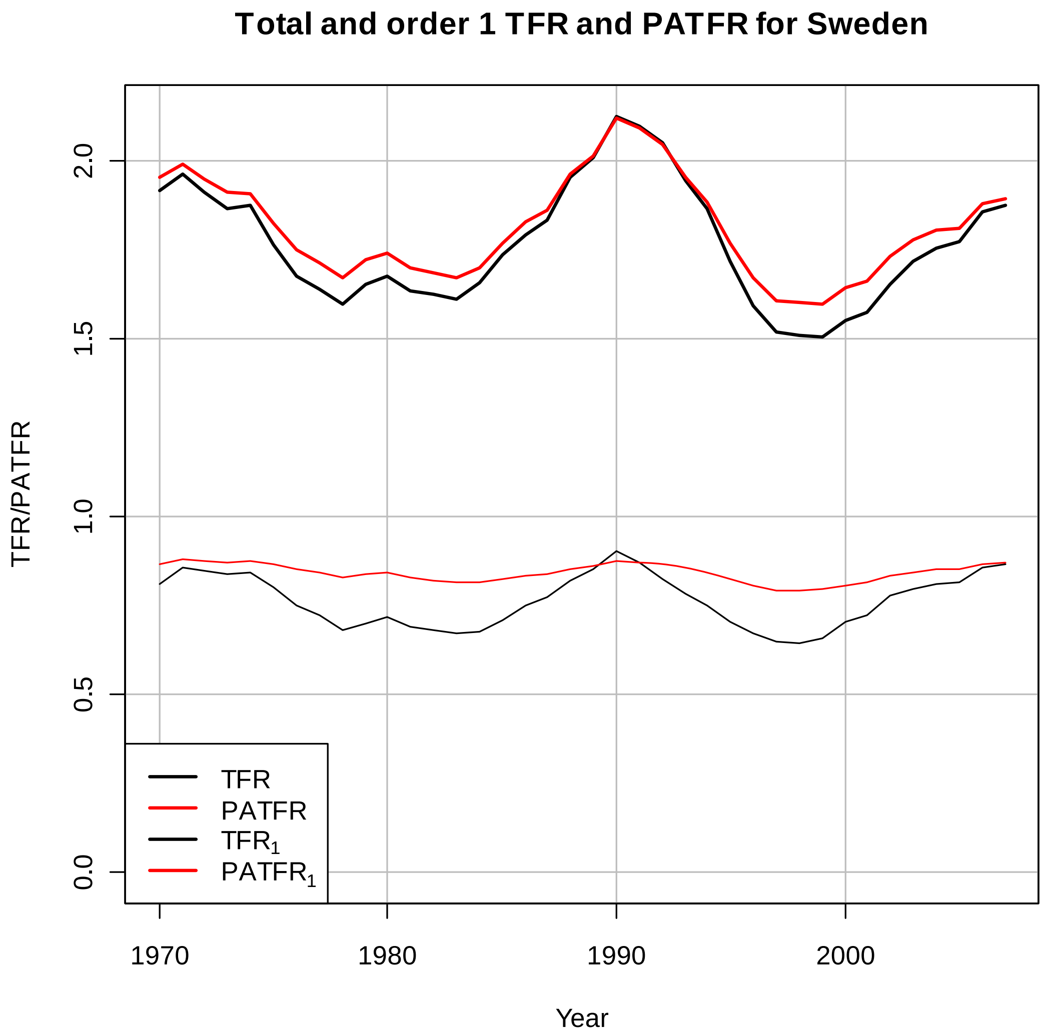

Figure 2. Total and order 1 TFR and PATFR for Sweden

```
> # Step 1: Read in Lexis triangle data
> birthsTR <- read.table.hfd("SWEbirthsTRbo.txt")
> exposTR <- read.table.hfd("SWEexposTR.txt")
> # Step 2: Check no births in open-interval age groups, then drop them
> all(birthsTR[grepl("[+-]", birthsTR$Age), "Total"] == 0)
[1] TRUE
> birthsTR <- birthsTR[!grepl("[+-]", birthsTR$Age), ]
> birthsTR$Age <- as.numeric(levels(birthsTR$Age)[birthsTR$Age])
> exposTR <- exposTR[exposTR$Year %in% birthsTR$Year & exposTR$Age %in%
  birthsTR$Age, ]
> # Step 3: Produce births by Lexis squares and then period TFR
> birthsRR <- produceLexisShapes(birthsTR, "RR", "births")
> exposRR <- produceLexisShapes(exposTR, "RR", "expos")
> asfrRR <- calculate.asfr(birthsRR, exposRR, "RR")
> tfr <- calculate.tfr(asfrRR, "RR")
> # Step 4: Calculate relative distribution of population exposure
> wc <- parity.census.weights("SWEparity.txt", 13:55) # ages to max+1
> w1J <- parity.1Jan.weights(wc)
```

```
> w <- calculateAgeParityDistribution.golden(birthsTR, exposTR, w1J, 1970)
> # Step 5: Calculate PFT and thereafter PATFR
> pft <- calculate.pft(birthsRR, exposRR, w)
> patfr <- calculate.patfr(pft)
> # Step 6: Perform plotting
> plot(tfr[, 1:2], ylim = c(0, max(tfr[, 2], patfr[, 2], na.rm = TRUE)),
  main = "Total and order 1 TFR and PATFR for Sweden", type = "n", ylab =
  "TFR/PATFR")
> grid(col = "gray", lty = "solid")
> lines(tfr[, 1:2], lwd = 2)
> lines(patfr[, 1:2], lwd = 2, col = "red")
> lines(tfr[, c(1, 3)])
> lines(patfr[, c(1, 3)], col = "red")
> legend("bottomleft", c("TFR", "PATFR", expression(TFR[1]),
  expression(PATFR[1])), col = rep(c("black", "red"), 2), lwd = c(2, 2, 1,
  1), bg = "white")
> box()
```

Lexis “heatmap” of first-order conditional fertility rates

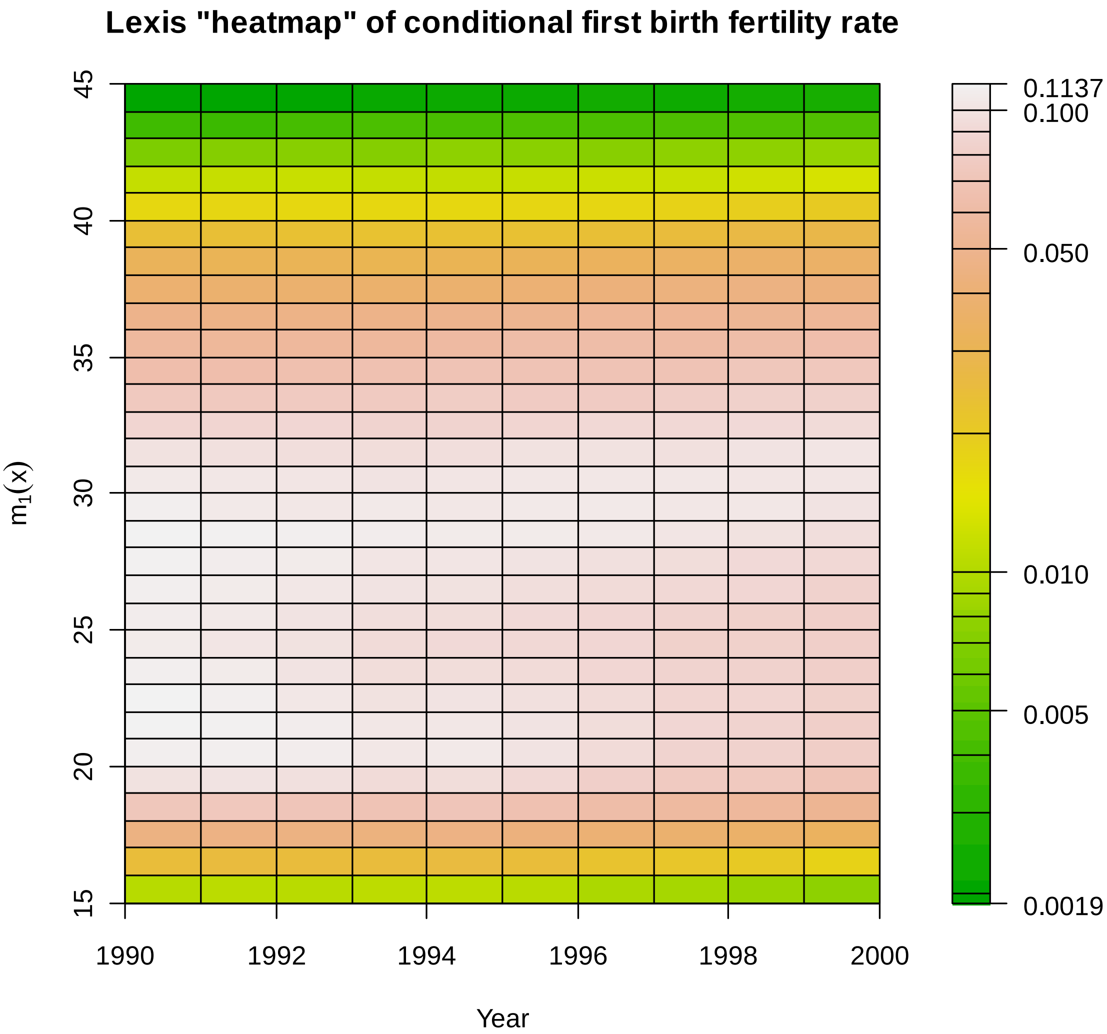

Figure 3. Lexis “heatmap” of conditional first birth fertility rate from sample data

This example shows how, if the distribution of female population by parity is available, an individual indicator from the period fertility table may be straightforwardly calculated and the values graphically displayed to enable easy analysis. The colour scale is applied to the log-transformed rates in order to give better contrast in the region with the highest values.

```
> # Step 1: calculate m
> data(someBirths.RR, someExpos.RR, someWeightings)
> E <- calculate.E.pft(someWeightings, someExpos.RR)
> m <- calculate.m.pft(someBirths.RR, E)
> # Step 2: trim age range so that colour scale can show more detail
> ages <- 15:44
> yrs <- as.numeric(dimnames(m)$Year)
> m <- m[, as.character(ages), ]
> # Step 3:  setup plotting area
> layout(matrix(c(1,2), ncol = 2), widths = c(6, 1))
> # Step 4: plot heatmap with log scale for the colours
> image(yrs + 0.5, ages + 0.5, log(m[, , "m1x"]), col = terrain.colors(256),
  main = "Lexis \"heatmap\" of conditional first birth fertility rate", xlab
  = "Year", ylab = expression(m[1](x)))
> abline(h = ages, v = yrs)
> box()
> # Step 5: add legend
> par.orig <- par(mar = c(5, 0.1, 4, 3.5) + 0.1)
> plot.new()
> mr <- range(m[, , 1], na.rm = TRUE)
> lvls <- seq(from = mr[1], to = mr[2], length.out = 256) # colour intervals
> plot.window(0:1, mr, xaxs = "i", yaxs = "i", log = "y")
> image(0.5, lvls, log(array(lvls, dim = c(1, 256))), col =
  terrain.colors(256), add = TRUE)
> axis(4, mr, round(mr, 4), las = 1, lwd = 0, lwd.ticks = 1) # max/min vals
> axis(4, c(10^-(1:2), 0.5 * 10^-(1:2)), las = 1, lwd = 0, lwd.ticks = 1)
> abline(h = c(seq(0.001, 0.009, 0.001), seq(0.01, 0.1, 0.01)))
> box()
> # Finally: reset graphics parameters
> par(par.orig)
```

Summary

This Technical Report has introduced a collection of “R” packages which may be used to produce period fertility tables and other period fertility indicators based on the methodology used in the Human Fertility Database. These packages allow the HFD methods to be straightforwardly applied, allowing users to experiment with their own data or with different parameters. The methods were summarised and the major functions included in the packages were described, with examples illustrating their usage. Further details on the methodology may be found in the HFD Methods Protocol (Jasilioniene et al., 2010), and further details on the software may be found in the online documentation included in the packages, and by examining the source code.

References

Bongaarts, J. and Feeney, G. (2006). The Quantum and Tempo of Life-Cycle Events. *Vienna Yearbook of Population Research*, 2006: 115-151.

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.

Kohler, H.-P. and Ortega, J.A. (2002). Tempo-adjusted period parity progression measures, fertility postponement and completed cohort fertility. *Demographic Research* 6(6), pp. 91–144.

Nash, E., Jasilioniene, A. and Andreev, E. (2010). An “R” package for the production of cohort fertility tables, Rostock/Germany, Max Planck Institute for Demographic Research, TR-2010-007. URL http://www.demogr.mpg.de/papers/technicalreports/tr-2010-007.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.

Rallu, L. and Toulemon, L.. (1994). Period fertility measures: The construction of different indices and their application to France, 1946-89. *Population, An English Selection* 6, pp 59-94.

Wunsch, G. (2006). Rates, Frequencies, and Probabilities; in Caselli, G., Vallin, J., and Wusch, G. (Eds.) *Demography and Synthesis*, Vol. 1. Amsterdam et al.: Elsevier, pp. 79-87.

Appendix A: Major functions

Table 3. Major functions included in package `hfdBasicIndicators`

Function name	*Purpose*	*Main input data and formats*	*Output data format*
`calculate.asfr`	Calculation of age-specific fertility rates	Births and female population exposures as – File or file connection (e.g. URL) in HFD output file structure – `data.frame` with columns as in HFD output files – 3D `array` with named dimensions Year/Cohort, Cohort/Age and Order	Data frame with columns as in HFD output files
`calculate.cfr`	Calculation of cumulated fertility rates	Age specific fertility rates as – File or file connection (e.g. URL) in HFD output file structure – `data.frame` with columns as in HFD output files – 3D `array` with named dimensions Year/Cohort, Cohort/Age and Order	Data frame with columns as in HFD output files
`calculate.tfr`	Calculation of total fertility rates	Age specific fertility rates in RR or VH as – File or file connection (e.g. URL) in HFD output file structure – `data.frame` with columns as in HFD output files – 3D `array` with named dimensions Year/Cohort, Cohort/Age and Order	Data frame with columns as in HFD output files
`calculate.mab`	Calculation of mean age at birth	Age specific fertility rates in RR or VH as – File or file connection (e.g. URL) in HFD output file structure – `data.frame` with columns as in HFD output files – 3D `array` with named dimensions Year/Cohort, Cohort/Age and Order	Data frame with columns as in HFD output files

Function name	*Purpose*	*Main input data and formats*	*Output data format*
`calculate.ppr`	Calculation of cohort parity progression ratios	Age specific fertility rates in VH as – File or file connection (e.g. URL) in HFD output file structure – `data.frame` with columns as in HFD output files – 3D `array` with named dimensions Year/Cohort, Cohort/Age and Order	Data frame with columns as in HFD output files

Table 4. Functions included in package `hfdParity` for producing relative distribution of female population exposure

Function name	*Purpose*	*Main input data and formats*	*Output data format*
`calculateAgeParityDistribution.cumulative`	Calculation of relative distribution of female population exposure – by cumulating cohort fertility	– Births and female population exposures by Lexis triangles as a `data.frame` with columns as in the HFD output files.	3D `array` of relative distribution of female population exposure with named dimensions Year, Age, Parity.
`calculateAgeParityDistribution.golden`	– by cumulating cohort fertility forward from initial values obtained from a census or register	– Births and female population exposures by Lexis triangles as a `data.frame` with columns as in the HFD output files. – Female population distribution on 1st January as a `data.frame` with the structure produced by `parity.1Jan.weights` or 3D `array` with named dimensions Year, Age and Parity	As for `.cumulative`
`calculateAgeParityDistribution.census`	– using only data from censuses	As for `.golden`	As for `.cumulative`
`calculateAgeParityDistribution.register`	– using only data from a population register	As for `.golden`	As for `.cumulative`

Table 5. Major functions included in package `hfdParity` for processing census and register parity data

Function name	*Purpose*	*Main input data and formats*	*Output data format*
`parity.census.weights`	Calculation of relative population weights by parity on the day of a census	Female population by parity counts as – File in the standard HFD input format for `XXXparity.txt` – `data.frame` with columns as in the HFD input file	A `data.frame` of the parity distribution of females on the census day for each year and age. Columns are Year, Month, Day, Age, Format, w0, w1, …, wi_l^++, where format is one of ACY or ARDY
`parity.1Jan.weights`	Calculation of relative population weights by parity on 1st January of a census year	Relative parity distributions of the female population as – A `data.frame` as produced by `parity.census.weights` – A file or file connection representing this in comma-separated format Age specific fertility rates in VH as – File or file connection (e.g. URL) in HFD output file structure – `data.frame` with columns as in HFD output files	A `data.frame` of the parity distribution of females on 1st January for each year and age. Columns are Year, Age, w0, w1, …, wi_l^+p.

Table 6. Main calculation functions included in package `hfdPeriodFertilityTable`

Function name	*Purpose*	*Main input data and formats*	*Output data format*
`calculate.pft`	Calculation of whole period fertility table	Births by birth order and female population exposure by RR as – File or file connection (e.g. URL) in HFD standard format for RR output files – `data.frame` with columns as for HFD RR output files – 3D `array` with named dimensions Year, Age and Order Relative distribution of female population exposure as – 3D `array` with named dimensions Year, Age and Parity – `data.frame` with columns Year, Age, w0, …, wi_l^+p – File or file connection to data in format as for `data.frame`	– `data.frame` in HFD standard format for `XXXpft.txt` – 3D `array` of indicators with named dimensions Year, Age and Indicator – `list` containing each indicator in a separate `array` by Year, Age and Order
`calculate.E.pft`	Calculation of female population exposure by parity	Births by birth order and relative distribution of female population exposure in same respective formats as for `calculate.pft`	3D `array` of female population exposure by parity with named dimensions Year, Age and Parity

Function name	*Purpose*	*Main input data and formats*	*Output data format*
`calculate.m.pft`	Calculation of conditional fertility rates by year, age and birth order	Births by birth order and female population exposure by parity by RR as – File or file connection (e.g. URL) in HFD standard format for `XXXbirthsRRbo.txt` and `XXXexposRRpa.txt` respectively. – `data.frame` with columns as for HFD output files – 3D `array` with named dimensions Year, Age and Order/Parity	3D `array` of conditional fertility rates with named dimensions Year, Age and Order
`calculate.q.pft`	Calculation of conditional probabilities of giving birth by year, age and birth order	– File or file connection (e.g. URL) in HFD standard format for `XXXmi.txt` – `data.frame` with columns as for `XXXmi.txt` – 3D `array` of conditional fertility rates with named dimensions Year, Age and Order	3D `array` of conditional probabilities of giving birth with named dimensions Year, Age and Order
`calculate.b.l.L.pft`	Calculation of table births, populations and exposures by year, age and birth order	3D arrays of conditional rates and birth probabilities with named dimensions Year, Age and Order	`list` with entries `b`, `l` and `L`, each containing a 3D `array` with named dimensions Year, Age and Order/Parity
`calculate.patfr`	Calculation of period fertility index of total fertility by year and birth order	Period fertility table (or table births) as – File or file connection (e.g. URL) in HFD standard format for `XXXmi.txt` – `data.frame` with columns as for `XXXpft.txt` – 3D `array` of PFT indicators with named dimensions Year, Age and Order/Indicator	`data.frame` of period fertility index of total fertility by birth order with columns as for `XXXpatfr.txt`

Function name	*Purpose*	*Main input data and formats*	*Output data format*
`calculate.pmab`	Calculation of table mean age at birth by year and birth order	Period fertility table (or table births) as – File or file connection (e.g. URL) in HFD standard format for `XXXmi.txt` – `data.frame` with columns as for `XXXpft.txt` – 3D `array` of PFT indicators with named dimensions Year, Age and Order/Indicator	`data.frame` of table mean age at birth by birth order with columns as for `XXXpmab.txt`

Table 7. Parameters accepted by major functions in the packages `hfdBasicIndicators`, `hfdParity` and `hfdPeriodFertilityTable`

Function name	*Parameters*
`calculate.asfr`	`births, expos, shape, …`
`calculate.cfr`	`asfr, shape, observeFrom, …`
`calculate.tfr`	`asfr, shape, observeFrom, observeTo, …`
`calculate.mab`	`asfr, shape, observeFrom, observeTo, a, …`
`calculate.ppr`	`ctfr, shape, …`
`calculateAgeParityDistribution…`	
`… .cumulative`	`bTr, eTr, observeFrom, observeTo, …`
`… .golden`	`bTr, eTr, wiyT, goldenYear, observeFrom, observeTo, switchToCumulative, …`
`… .census`	`bTr, eTr, wiyT, …`
`… .register`	`bTr, eTr, wiyT, …`
`parity.census.weights`	`parity, ages, reqRange, maxParity, unkMethod, outfile, country, ldb, …`
`parity.1Jan.weights`	`census.weights, asfr.vv.bo, outfile`
`calculate.pft`	`birthsRR, exposRR, w, a, returnType, …`
`calculate.E.pft`	`w, exposRR, …`
`calculate.m.pft`	`birthsRR, E, filterPlausible, adjustE, …`
`calculate.q.pft`	`m, a, filterPlausible, …`
`calculate.b.l.L.pft`	`m, qix, a, …`
`calculate.patfr`	`pft, radix, …`
`calculate.pmab`	`pft, a, …`

Table 8. Description of individual parameters to functions

Parameter name	*Default value*	*Description*
`births`	–	Observed births as a `data.frame`, 3D `array` or path to file
`expos`	–	Female population exposure as a `data.frame`, 3D `array` or path to file
`shape`	`"VH"` for `calculate.ppr`	String indicating the Lexis shape in which data is held; `"TR"`, `"RR"`, `"VH"` or `"VV"`
`asfr`	–	Age-specific fertility rates as a `data.frame`, 3D `array` or path to file
`observeFrom`	`15`	Age from which cohorts must have been observed to calculate cohort indicators
`observeTo`	`50` for `hfdBasicIndicators`, `45` for `hfdParity`	Minimum age to which cohorts must be observed
`ctfr`	–	Cumulative fertility rates as `data.frame`, 3D `array` or path to file
`bTr`	–	Observed births by Lexis triangles as a `data.frame`
`eTr`	–	Female population exposure by Lexis triangles as a `data.frame`
`wiyT`	–	Relative female population distribution by parity on 1st January of census/register year(s) as a `data.frame` or 3D `array`
`goldenYear`	–	Census/register year to use as “golden census”
`switchToCumulative`	`TRUE`	Whether to switch relative parity weight calculation to the cumulative method once cohorts have been observed for a sufficiently long period
`parity`	–	HFD `XXXparity.txt` input file, or `data.frame` representing this
`ages`	`12:56`	Ages for which to calculate relative parity distribution
`reqRange`	`c(15,49)`	Range of ages which must be all included in parity data in order to calculate a distribution

Parameter name	*Default value*	*Description*
`maxParity`	`4`	i_l^+; the maximum (open-interval) parity for which the distribution should be calculated
`unkMethod`	`c(default="proportional")`	Method used for distribution females of unknown parity as a character vector with labels for a default method and for each year with a non-default method and entries either `"proportional"` or `"allToZero"`.
`outfile`	`NULL`	Path to a file to which output should be written.
`country`	`NULL`	Country name.
`ldb`	`TRUE`	Whether only rows in the parity file with LDB=1 should be used.
`census.weights`	–	Relative distribution of female population by parity on 1st January as a `data.frame` or path to file
`asfr.vv.bo`	–	Age-specific fertility rates in Lexis vertical parallelograms as a `data.frame` or path to file.
`birthsRR`	–	Observed births by birth order in Lexis squares as a `data.frame`, 3D `array` or path to file
`exposRR`	–	Female population exposure in Lexis squares as a `data.frame`, 3D `array` or path to file
`w`	–	Relative distribution of female population exposure by parity as a `data.frame`, 3D `array` or path to file
`a`	`0.5`	Average share of the age interval $\left[x, x+1\right)$ lived before giving birth
`returnType`	`"data.frame"`	How the PFT should be returned; one of `"data.frame"`, `"array"` (all indicators in one 3D `array`) or `"list"` (containing each indicator in a separate `array`)
`E`	–	Female population exposure by parity as a `data.frame`, 3D `array` or path to file

Parameter name	*Default value*	*Description*
`filterPlausible`	FALSE	Whether implausible values should be replaced with `NA` in the result
`adjustE`	`TRUE`	Whether parity-specific exposures less than 5 should be replaced with 0 to avoid irregularities caused by very small populations
`m`	–	Conditional fertility rates as a 3D `array` , or (for `calculate.q`) `data.frame`, or path to file
`qix`	–	Conditional probabilities of giving birth as a 3D `array`
`pft`	–	Period fertility table (or period births) as a `data.frame`, 3D `array` or path to file
`radix`	10000	Population size in period fertility table
…	–	Arguments passed to `read.table` to configure reading data from file.

Appendix B: Sample datasets

Table 9. Sample datasets provided

Package	*Dataset*	*Description*
`hfdLexisManipulation`	`someBirths.TR`	Births by birth order and total in Lexis triangles as a `data.frame` in the structure of the HFD output files.
	`someExpos.TR`	Female population exposure in Lexis triangles as a `data.frame` in the structure of the HFD output files.
`hfdParity`	`czeCensus`	Female population by parity data from censuses in the Czech Republic as a `data.frame` in HFD input format. This is the same data as the file CZEparity.txt available from the HFD page for the Czech Republic.
	`sweRegisterPath`	Character string holding the path to a file containing an extract of the female population by parity data from the Swedish population register in HFD input format. This file is installed as part of the package and is an extract from the file SWEparity.txt available from the HFD page for Sweden.
`hfdPeriodFertilityTable`	`someBirths.RR`	Births by birth order and total in Lexis squares as a `data.frame` in the structure of the HFD output files.
	`someExpos.RR`	Female population exposure in Lexis squares as a `data.frame` in the structure of the HFD output files.
	`someWeightings`	Relative distribution of female population exposure as a 3D `array` with named dimensions Year. Age and Parity.