30 January 2007

Structure of the Input Database

The Input Database (InputDB) contains three types of data: deaths, population, and births. For each population, we have 8 different files:

- XXXdeath.txt (for death data),
- XXXpop.txt (for population data),
- XXXbirth.txt (for birth data),
- XXXref.txt (references),
- XXXnote.txt (for specific notes),
- XXXcom.pdf (for background and documentation),
- XXXtadj.txt (for territorial adjustment factors if relevant), and
- XXXreadme.txt (with a descriptive identifier for this population).

where XXX is a population code (3-7-digits, all character values must be uppercase). The first three digits represent a 3-letter country code (e.g., FRA, USA, for details see http://www.mortality.org/Public/CountryList.html). An optional fourth digit identifies the national population, where "T" denotes the total population (i.e., including both military and civilian) and "C" denotes the civilian population. In cases where this information is known or unclear (e.g., not consistent across deaths and population estimates), the fourth digit may be omitted (if the data represent the national population) or indicated by "_" (if the data represent a subpopulation identified by the 5th-7th digits). The fifth, sixth, and seventh digits may be used to identify a subpopulation (e.g., region/province, racial/ethnic group) as needed. This code may be any combination of up to three numbers of uppercase characters. For example,

NOR	identifies the national population of Norway (unknown/uncertain whether military are included)
FRAT	identifies the total national population of France
FRAC	identifies the civilian national population of France
NZL_NP	identifies the national population of New Zealand
NZL_MA	identifies the Maori population of New Zealand

The Background and Documentation file (XXXcom.pdf) is in portable document format (.pdf). All other files are in ASCII format. Each data file (XXXdeath.txt, XXXpop.txt, XXXbirth.txt, and XXXtadj.txt) contains standard headings (first line), which represent field identifiers. In these files each line (record) is defined independently from other records (i.e., each record contains all necessary information, and the sequence of the records has no significance). Each record is unique (i.e., there are no duplicate records). We use CRLF ("\r\n") combination of characters as a record delimiter and a comma (",") as the field delimiter. Missing values are coded as a single dot ("."). Optionally, the data files may contain a number of spaces to improve text file readability. The spaces have no other function.

Description of formats:

1. Deaths

File name: XXXdeath.txt

Heading:

PopCode, Area, Year, YearReg, YearInterval, Sex, Age, AgeInterval, Lexis, RefCode, Access, Deaths, NoteCode1, *NoteCode2*, *NoteCode3, LDB*

For each death count, we have one record. Each death count (field "*Deaths*") is determined by population code (field "*PopCode*"), geographical coverage (field "*Area*"), calendar year (field "*Year*"), year of registration ("*YearReg*"), the length of year interval of the Lexis element (field "*YearInterval*"), sex (field "*Sex*"), age (field "*Age*"), the length of age interval (field "*AgeInterval*"), the shape of Lexis element (field "*Lexis*"), reference code (field "*RefCode*"), and type of access (field "*Access*"). Each record also contains three fields that link to specific comments ("*NoteCode1*", "*NoteCode2*", and "*NoteCode3*"). The final field ("*LDB*") indicates whether the data are used to create the Lexis database. Note: If this file includes death counts for a particular calendar year, but an observation is omitted for a particular sex, age, and lexis combination, then for the purposes of calculating mortality estimates, we assume there were no deaths for this group.

The format of the fields is as follows:

1.1) Population code (3-7-digits). Same code ("XXX") as in the file name.

1.2) Area (2-digit). For example: 01, 10, 20, …. This field serves to reflect territorial (or population) coverage.

1.3) Calendar year (4-digit). Year in which the deaths occurred.

1.4) Year of registration (4-digit). In recent years, some countries (e.g., Finland) have reported deaths registered in current year, but which actually occurred earlier in time. In such cases, *Year*, *Age,* etc. are coded to refer to the date of occurrence, while *YearReg* is set to the registration year. If no distinction is made between year of occurrence and year of registration[1], then *Year = YearReg.*

1.5) The length of the year interval of the Lexis element (1- or 2-digit). For example: 1, 2, 3, 5, 10.

1.6) Sex (1-char). The character 'm' denotes males, 'f' denotes females, and 'b' indicates both.

1.7) Age (1-, 2- or 3-digit or 3-char). For age groups, the value is always equal to the lower age limit, and TOT[2] and UNK stand for total and unknown ages, respectively.

1.8) The length of the age interval (1- or 2-digit or '+' for open age interval). For example: 1, 2, 3, 5, 10, +.

1.9) Lexis element (2-char). This field denotes the shape of the Lexis element, where: TL=lower triangle, TU=upper triangle, RR=rectangle, VV=parallelogram with vertical left and right sides (i.e., period-cohort), VH=parallelogram with horizontal upper and lower sides (i.e., age-cohort); RV=Same as VV except also includes TL for the first age in the interval (e.g., cohorts aged 0-4 on Dec 31st—includes those born in the current calendar year).

1.10) Reference code (numeric, no fixed-length). A numeric code that identifies a data source provided in the file XXXref.txt (see description for XXXref.txt).

1.11) Access (1-char). This field indicates the confidentiality/accessibility of the data ('C' - confidential, 'O' – publicly accessible, 'U' – unknown).

1.12) Deaths (numerical field, no fixed length). Number of deaths.

[1] Recording deaths by year of occurrence deaths is a new trend in statistical publications. In earlier years, deaths were recorded only by year of registration.

[2] The total is always taken from the original data. It may be different from the total yielded by summing up age specific counts.

1.13) Note codes (numeric, no fixed length). These three fields ("*NoteCode1*", "*NoteCode2*", "*NoteCode3*") link to specific notes contained in the file "XXXnote.txt". These fields may be empty (denoted by a single dot '.').

1.14) LDB (numeric, 0/1). This field ("*LDB*") is coded "1" if the data are used to create the Lexis database and coded "0" if not.[3]

[3] We are currently in the process of adding this field. Therefore, it is not yet included for all countries.

Example: Total National Population of France

```
PopName,Area,  Year,  YearReg,  YearInterval, Sex,  Age,   AgeInterval,  Lexis, RefCode,  Access,  Deaths,  NoteCode1,  NoteCode2,  NoteCode3,  LDB
 FRATNP,    30, 1907,      1907,              1,   m,      0,              1,    TL,        10,        C,  43118,          1,          .,          .,
 FRATNP,    30, 1907,      1907,              1,   m,      0,              1,    TU,        11,        C,  16110,          2,          3,          .,
 FRATNP,    30, 1907,      1907,              1,   m,      1,              1,    TL,        10,        C,   5822,          .,          .,          .,
 FRATNP,    30, 1907,      1907,              1,   m,      1,              1,    TU,        10,        C,   4806,          1,          .,          .,
    ……..  ……   ……       ……            ……   ….   …..            ….   …..      …..      …..   …….  …………..   ………..   …………
 FRATNP,    30, 1907,      1907,              1,   m,    123,              +,    TL,         5,        C,      0,          2,          5,         10,
 FRATNP,    30, 1907,      1907,              1,   m,  UNK,              .,     .,        10,        C,      0,          .,          .,          .,
 FRATNP,    30, 1907,      1907,              1,   m,   TOT,              .,     .,        10,        O, 415491,          .,          .,          .,
```

2. Population

File name: XXXpop.txt

Heading:
PopCode, Area, Sex, Age, AgeInterval, Type, Day, Month, Year, RefCode, Access, Population, NoteCode1, *NoteCode2*, *NoteCode3, LDB*

Each population size value (field "*Population*") is determined by population code (field "*PopCode*"), geographic coverage (field "*Area*"), sex (field "*Sex*"), age group (field "*Age*"), the length of age interval (field "*AgeInterval*"), type code (field "*Type*"), the day, month, and calendar year corresponding to the date of the census count or population estimate (fields "*Day*", "*Month*", "*Year*", respectively), source of data (field "*RefCode*"), and type of access (field "*Access*"). Each record also contains three fields that link to specific comments ("*NoteCode1*", "*NoteCode2*", and "*NoteCode3*"). The final field ("*LDB*") indicates whether the data are used to create the Lexis database.

The format of the fields is as follows:

2.1) Population code (3-7-digits, the same as for deaths).

2.2) Area (2-digit, the same as for deaths).

2.3) Sex (1-char, the same as for deaths).

2.4) Age (1-, 2- or 3-digit or 3-char, same as for deaths). Note: If Type='B', then this field represents age at the end of the calendar year in which the census occurred (i.e., year of census minus year of birth).

2.5) The length of age interval (1- or 2-digit or '+', same as for deaths).

2.6) Type (1-char). The character 'C' denotes census count, 'O' denotes official estimates by statistical offices, 'R' denotes register population counts, 'E' indicates other estimates, and 'B' denotes census count classified by year of birth (not age).

2.7) Day (2-digit). Day corresponding to the date of the census or population estimate. Possible values: 01, 02, …, 31

2.8) Month (2-digit). Month corresponding to date of the census or population estimate. Possible values: 01, 02, …, 12

2.9) Year (4-digit). Year corresponding to the date of the census or population estimate.

2.10) Reference code (numeric, the same as for deaths).

2.11) Access. (1-char, the same as for deaths).

2.12) Population (numerical field, no fixed length). Population size.

2.13) Note codes (numeric, the same as for deaths)

2.14) LDB (numeric, 0/1). This field ("*LDB*") is coded "1" if the data are used to create the Lexis database and coded "0" if not.[4]

[4] We are currently in the process of adding this field. Therefore, it is not yet included for all countries.

Example: Total National Population of France

PopName,	*Area,*	*Sex,*	*Age,*	*AgeInterval,*	*Type,*	*Day,*	*Month,*	*Year,*	*RefCode,*	*Access,*	*Population,*	*NoteCode1,*	*NoteCode2,*	*NoteCode3,*	*LDB*
FRATNP,	30,	m,	0,	1,	E,	01,	01,	1899,	10,	O,	363100,	1,	.,	.,	1
FRATNP,	30,	m,	1,	1,	E,	01,	01,	1899,	10,	O,	350100,	2,	3,	.,	1
FRATNP,	30,	m,	2,	1,	E,	01,	01,	1899,	10,	O,	345800,	2,	5,	10,	1
……..	…..	….	….	…..	…..	…..	…..	……	……	…..	………	…………..	…………	…………	1
FRATNP,	30,	m,	100,	+,	E,	01,	01,	1899,	10,	C,	10,	1,	.,	.,	1

3. Births

File name: XXXbirth.txt

Heading: *PopCode, Area, Sex, Year, YearReg, RefCode, Access, Births, NoteCode1*, *NoteCode2*, *NoteCode3, LexisDB*

Each birth count (field "*Births*") is specified by population code (field "*PopCode*"), geographical coverage (field "*Area*"), sex (field "*Sex*"), calendar year (field "*Year*"), year of registration (field "*YearReg*"), reference code (field "*RefCode*") and type of access (field "*Access*"). Each record also contains three fields that link to specific comments ("*NoteCode1*", "*NoteCode2*", and "*NoteCode3*"). The final field ("*LDB*") indicates whether the data are used to create the Lexis database.

The format of the fields is as follows:

3.1) Population code (3-7-digit, the same as for deaths).

3.2) Area (2-digit, the same as for deaths).

3.3) Sex (1-char, the same as for deaths).

3.4) Calendar year (4-digit). Year in which the births occurred.

3.5) Year of registration (4-digit, the same as for deaths). In the recent years, some countries have reported births that occurred in an earlier year with the statistics for the current year. In such cases *Year*, *Age,* etc. are coded to refer to the date of occurrence, while *YearReg* is set to the registration year. If no distinction is made between year of occurrence and year of registration, then *Year* = *YearReg*.

3.6) Reference code (numeric, the same as for deaths).

3.7) Access. (1-char, the same as for deaths).

3.8) Births (numerical field, no fixed length). The number of live births.

3.9) Note codes (numeric, the same as for deaths)

3.10) LDB (numeric, 0/1). This field ("*LDB*") is coded "1" if the data are used to create the Lexis database and coded "0" if not.[5]

[5] We are currently in the process of adding this field. Therefore, it is not yet included for all countries.

Example: Total National Population of France

PopName,	*Area,*	*Sex,*	*Year,*	*YearReg,*	*RefCode,*	*Access,*	*Births*	*NoteCode1,*	*NoteCode2,*	*NoteCode3,*	*LDB*
FRATNP,	30,	m,	1899,	1899,	11,	O,	435485,	1,	.,	.,	1
FRATNP,	30,	m,	1900,	1900,	10,	O,	425139,	2,	3,	.,	1
FRATNP,	30,	m,	1901,	1901,	11,	O,	440012,	2,	5,	10,	1
FRATNP,	30,	m,	1902,	1902,	10,	O,	434410,	1,	.,	.,	1
……..	……	…..	…….	……..	…….	……	……..	……..	……..	……..	1

4. Notes

File name: XXXnote.txt

This file contains specific notes pertaining to individual data points. These notes are identified by a note code that links to specific data points in the data files. The first line of each record contains a numeric code (the same as *NoteCode1, NoteCode2*, and *NoteCode3* in the data files). Lines following the note code contain specific notes (which may be written in a free form, but shouldn't include the blank lines). Blank lines are used for separating the records. Each subpopulation within a country must have a XXXnote.txt file, but it is left to the discretion of the country specialist whether those files are unique (containing only information for that particular subpopulation) or whether they are simply copies of one master file (containing information for all subpopulations within that country) with different filenames (e.g., NZL_NMnote.txt, NZL_MAnote.txt).

Example:

1
This number is probably not correct, but we don't have other data

2
For this year and age we need additional information

………………………………………………………………………………

5. References

File name: XXXref.txt

This file contains the description of sources. Each data source has a corresponding record in this file. Each record in this file is separated by a blank line. Each subpopulation within a country must have a XXXref.txt file, but it is left to the discretion of the country specialist whether those files are unique (containing only information for that particular subpopulation) or whether they are simply copies of one master file (containing information for all subpopulations within that country) with different filenames (e.g., NZL_NMref.txt, NZL_MAref.txt). Each record contains the following fields:

- *RefCode:* Reference code (the same as in the data files)
- *Source:* Full (or as full as possible) reference to the source

- *Comments:* First, list the type of data (e.g., deaths, births, census counts, population estimates) using the following format: "(Deaths, Births)". This field may also contain comments from the country specialist about this submission.
- *Date:* Date of incorporation into Database (format: dd.mm.yyyy)
- *Reference Person:* Person who added this entry to the reference file

The name of each field is included in the record on a separate line.

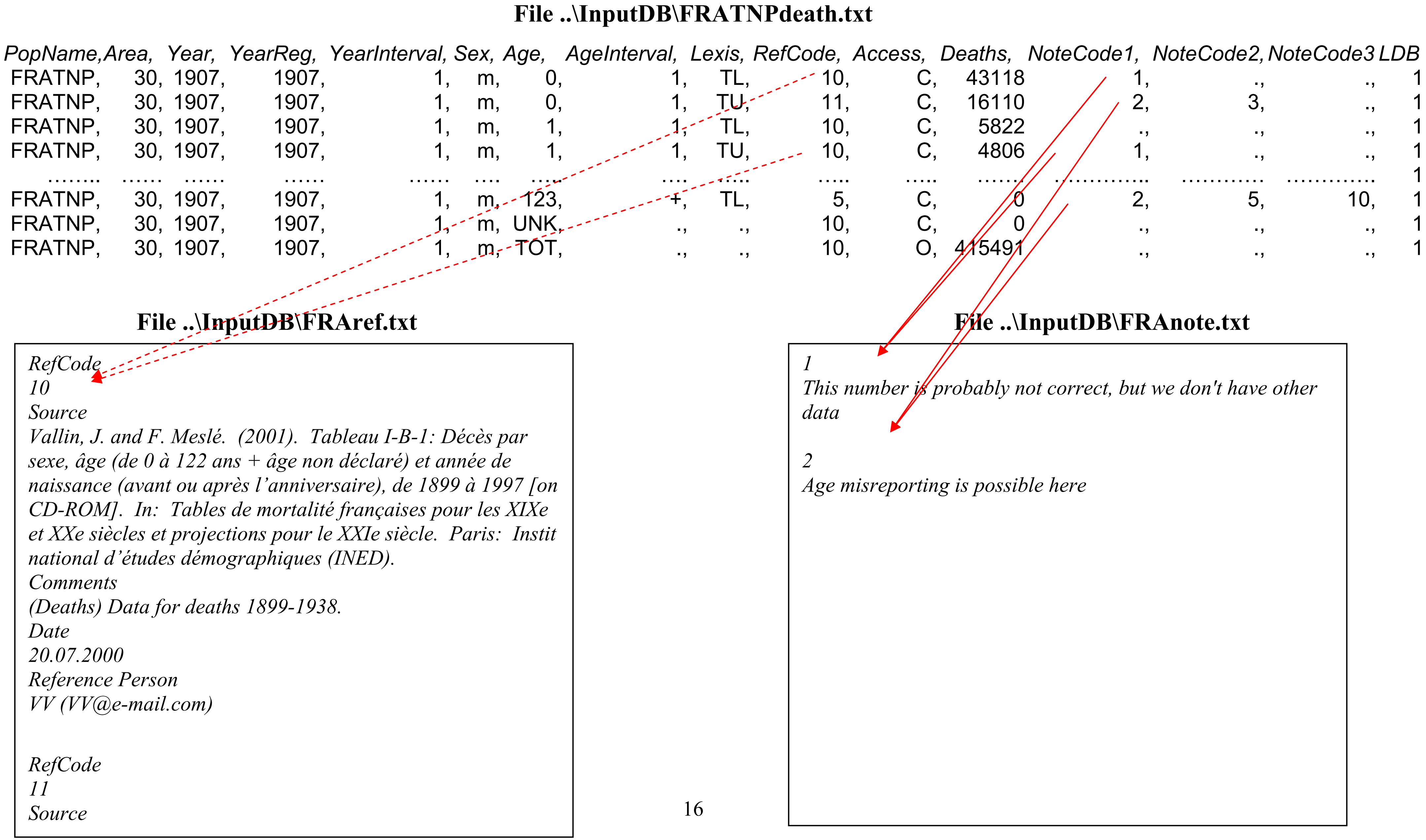
Example:

RefCode
10
Source
Vallin, J. and F. Meslé. (2001). Tableau I-C-1: Population par sexe et âge (de 0 à 100 ans), au 1 janvier, de 1899 à 1998, avec deux estimations selon le territoire pour les années de changement de territoire [revised post-publication]. In: Tables de mortalité françaises pour les XIXe et XXe siècles et projections pour le XXIe siècle. Paris: Instit national d'études démographiques.
Comments
(Population Estimates) Data for population estimates 1899-1998. Data received on CD-ROM accompanying publication.
Date
20.07.2000
Reference Person
VV (VV@e-mail.com)

Refcode
1
Source
Instituto Nazionale di Statistica (ISTAT). (1995). Tavola 2.1 -
Popolazione residente per sesso, stato civile e singolo di eta.
Pp. 73-74 in: Capitolo 2 - L'Italia Oggi, Immagini del Paese,
Popolazione e Abitazioni, Fascicolo Nazionale Italia, 13th
Censimento Generale Della Popalazione e Delle Abitazioni,
20 ottobre 1991. Roma: ISTAT.
Comments
(Census counts) Data from 1991 census. Received data in electronic file from Graziella Caselli (originally sent to Kirill Andreev). Hard copies of
published tables from University of California, Berkeley Library.
Date
10.11.2000
Reference Person
RR (RR@e-mail.com)

The connection between notes, references, and data records is presented in Figure 1.

Figure 1. Connection between files in the InputDB



File ..\InputDB\FRATNPdeath.txt

```
PopName,Area, Year, YearReg, YearInterval, Sex, Age,  AgeInterval, Lexis, RefCode, Access, Deaths, NoteCode1, NoteCode2, NoteCode3 LDB
FRATNP,    30, 1907,     1907,          1,   m,     0,          1,    TL,      10,     C,  43118         1,         .,         .,   1
FRATNP,    30, 1907,     1907,          1,   m,     0,          1,    TU,      11,     C,  16110         2,         3,         .,   1
FRATNP,    30, 1907,     1907,          1,   m,     1,          1,    TL,      10,     C,   5822         .,         .,         .,   1
FRATNP,    30, 1907,     1907,          1,   m,     1,          1,    TU,      10,     C,   4806         1,         .,         .,   1
  ……..   ……   ……      ……          ……    ….   ….         ….   …..     …..    …..   …….  …………..   …………   …………   1
FRATNP,    30, 1907,     1907,          1,   m,   123,          +,    TL,       5,     C,      0         2,         5,        10,   1
FRATNP,    30, 1907,     1907,          1,   m,   UNK,          .,     .,      10,     C,      0         .,         .,         .,   1
FRATNP,    30, 1907,     1907,          1,   m,   TOT,          .,     .,      10,     O, 415491         .,         .,         .,   1
```

File ..\InputDB\FRAref.txt

RefCode
10
Source
Vallin, J. and F. Meslé. (2001). Tableau I-B-1: Décès par sexe, âge (de 0 à 122 ans + âge non déclaré) et année de naissance (avant ou après l'anniversaire), de 1899 à 1997 [on CD-ROM]. In: Tables de mortalité françaises pour les XIXe et XXe siècles et projections pour le XXIe siècle. Paris: Instit national d'études démographiques (INED).
Comments
(Deaths) Data for deaths 1899-1938.
Date
20.07.2000
Reference Person
VV (VV@e-mail.com)

RefCode
11
Source

File ..\InputDB\FRAnote.txt

1
This number is probably not correct, but we don't have other data

2
Age misreporting is possible here

6. Background and Documentation

File name: XXXcom.pdf

This file is constructed by the country specialist and made available to users. For a given population, the Background and Documentation for the primary population (e.g., Total National Population) should contain the information that pertains to the population as a whole (e.g., completeness of coverage, data quality, historical information, border changes, etc.). The Background and Documentation for *subpopulations* should refer the user to the "main" Background and Documentation, but provide any additional information that is specific to that subpopulation. Information related to specific data points should be included in the Notes file rather than the Background and Documentation file. For example, comments regarding the source of data from country C for calendar year Y are given in "*Comments*" field of the reference file. However, comments regarding comparative characteristics of different data sources for year Y of country C should be stored in XXXcom.pdf. This file contains all (important) information that is known to country specialist. Nevertheless, it cannot be used as a full description of statistical system or history of the country.

7. Territorial adjustment factors

File name: XXXtadj.txt

Heading: *PopCode, Year, Age, Area1, Area2, Sex, RefCode, Access, Type, Value, NoteCode1*, *NoteCode2*, *NoteCode3*

Each territorial adjustment factor (field "*Value*") is specified by population code (field "*PopCode*"), calendar year (field "*Year*"), age (field "*Age*"), geographical coverage before (field "*Area1*") and after (field "*Area2*") territorial changes, sex (field "*Sex*"), reference code (field "*RefCode*"), type of access (field "*Access*"), and type of factor (field "*Type*"). Each record also contains three fields that link to specific comments ("*NoteCode1*", "*NoteCode2*", and "*NoteCode3*"). Changes in population coverage (e.g., a change from covering the *de facto* population to the *de jure* population) may also be treated as a territorial change for purposes of making calculations.

The format of the fields is as follows:

7.1) Population code (3-7-digit, the same as for deaths).

7.2) Calendar year (4-digit). Year for which this adjustment factor was calculated.

7.3) Age (1-, 2- or 3-digit). Age for which this adjustment factor was calculated. If the adjustment factor does not depend on age (e.g., ratio of births), this field should be coded as a single dot (".").[6]

7.4) Area1 (2-digit). This field indicates the code of the territorial (or population) coverage ***before*** territorial changes.

7.5) Area2 (2-digit). This field indicates the code of the territorial (or population) coverage ***after*** territorial changes.

7.6) Sex (1-char, the same as for deaths). The character 'm' denotes males, 'f' denotes females, and 'b' indicates both sexes.

7.7) Reference code (numeric, the same as for deaths).

7.8) Access (1-char). This field indicates the confidentiality/accessibility of data ('C' - confidential, 'O' – publicly accessible, 'U' – unknown).

7.9) Type (2-char). Type of factor: Rb = ratio of births, Vx = ratio of the population size at a specific age, Rd=ratio of the death counts at a specific age. See Appendix D of the Methods Protocol for details.

7.10) Value (real number). The value of the adjustment factor. The ratio of births is defined as $R_b(t) = \frac{B^+(t)}{B^-(t)}$, where $B^+(t)$ is number of births in the territory denoted by the *Area2* code (after the change), and $B^-(t)$ is number of births in the territory denoted by the *Area1* code (before the change). Similarly, the ratio of population size is defined as $V(x,t) = \frac{P^+(x,t)}{P^-(x,t)}$, where $P^+(x,t)$ is the population in the territory denoted by the *Area2* code and $P^-(x,t)$ is the population in the territory denoted by the *Area1* code. Thus, if a territory was added in *Year=t*, both $R_b(t)$ and $V(x,t)$ should be greater than one, whereas if

[6] For Vx factors, there should be a record for each single year of age (0 to 130). Nonetheless, the values may be the same for a range of ages. For example, at older ages (e.g., 90+) it is usually better to calculate the Vx factor based on aggregating across the open age interval in order to smooth out random variation. In this case, the *value* would be the same for ages 90, 91,…130.

a territory was lost, these factors should be less than one. [Note: The data (births/population) for both the numerator and denominator come from the same year (preferably the year of the territorial change); it is just the definition of the territory that differs.] The ratio of death counts, $R_d(x,t)$, is used in special cases where a change in population coverage (for purposes of counting deaths) occurs in the middle of a year (e.g., see Appendix 2 of the Background and Documentation file for New Zealand for a discussion of the change in the definition of ethnicity on September 1, 1995). This adjustment factor is needed to account for an increase (or decrease) in deaths resulting from the mid-year change in definition. It is calculated according to the following formula:

$$R_d(x,t) = \frac{\overline{M}(x,t-1,t+1)}{M(x,t)}$$

where $M(x,t)$ is the actual age-specific mortality rate for year in which the change occurs (where deaths are based partly on the old and partly on the new definition, but population estimates represent the old definition); $\overline{M}(x,t-1,t+1)$ is the average of the age-specific mortality rates for the prior year (t-1) and the subsequent year (t+1). By applying this ratio to the original death count in year t, we get the approximate number of deaths based on the old definition of population coverage

7.11) Note codes (numeric, the same as for deaths).

Example: Total National Population of France

PopName,	*Year,*	*Age,*	*Area1,*	*Area2,*	*Sex,*	*RefCode,*	*Access,*	*Type,*	*Value,*	*NoteCode1,*	*NoteCode2,*	*NoteCode3*
FRATNP,	1899,	.,	20,	30,	m,	1,	O,	Rb,	1.007,	1,	.,	.
FRATNP,	1914,	.,	30,	40,	m,	2,	O,	Rb,	0.811,	1,	.,	.
FRATNP,	1914,	25,	30,	40,	m,	11,	O,	Vx,	0.824,	2,	5,	10
FRATNP,	1914,	26,	30,	40	m,	11,	O,	Vx,	0.824,	1,	.,	.
……..	……	…..	…….	……..		…….	……	……..		……..	……..	……..

8. Descriptive Identifer

File name: XXXreadme.txt

The first line of this file contains a descriptive identifer for the population (i.e., name of country or area, type of population, and population subgroup) that will be used in the header for all data files displayed on the HMD website. For populations where the population classifer (e.g., total or civilian) is indeterminate, it is particularly important that the population is described to the best of your knowledge. At the discretion of the country specialist, other unspecified information (e.g. history of updates) can be included on subsequent lines.

For example:

FRATNPreadme.txt reads:

France, National Total Population

FRACNPreadme.txt reads:

France, National Civilian Population

ITA_NPreadme.txt reads:

Italy, National Population (excludes military deaths, but population includes military)