Check of raw data in IDS

Göran Broström

December 22, 2016

Data from scratch

First we read data, from IDS and from the created data file.

```
load("ef_fixed.rda")
ef <- ef_fixed
rm(ef_fixed)
source("R/read_IDS.R")
individual <- read_IDS()
```

Then we check the amount of information for each individual in IDS, present in *ef*.

```
id <- unique(ef$Id_I)
indiv <- individual[individual$Id_I %in% id, ]
xx <- with(indiv, tapply(Id_I, Id_I, length))
table(xx)
```

```
## xx
##   25   26   27   28   29   30   31   32   33   34   35   36   37   38   39
## 1627  397 2748 9433 5419 2072 1329 1338  826  665  639  611  549  537  474
##   40   41   42   43   44   45   46   47   48   49   50   51   52   53   54
##  438  416  428  389  341  345  335  318  268  252  226  219  175  201  195
##   55   56   57   58   59   60   61   62   63   64   65   66   67   68   69
##  164  150  146  133  128  115  106  100  117   97   79   70   73   67   72
##   70   71   72   73   74   75   76   77   78   79   80   81   82   83   84
##   44   55   60   51   46   44   29   31   31   25   21   26   25   18   22
##   85   86   87   88   89   90   91   92   93   94   95   96   97   98   99
##   13   14   16   23    9   12    9    7   13    8    9    4    5    7   11
##  100  101  102  103  104  105  106  107  108  109  111  112  113  114  116
##    7    5    5    3    6    2    3    1    1    2    2    1    1    1    1
##  122  123  124  127  129  132  138  139  140  179
##    1    1    1    1    2    1    1    1    1    1
```

OK, this is a little difficult to interpret, let's look at them in *skum*.

```
library(skum)
```

```
## Loading required package: eha
```

```
## Loading required package: survival
```

```
oj <- obs[obs$id %in% id, ]
length(unique(oj$id))
```

```
## [1] 0
```

```
dim(oj)
```

```
## [1]  0 33
```

Hmm, no connection between *id* in *observation* and *Id_I* in IDS-data. We have to go back to IDS. We start by noting that rows without any date can safely be discarded.

```r
oo <- rowSums(indiv[, 8:16])
indiv <- indiv[oo > 0.5, ]
xx <- with(indiv, tapply(Id_I, Id_I, length))
table(xx)
```

```
## xx
##    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17
##    4    9  117 1144  796  217  349  634 1543 5831 2201 1404 1854 2971 2054
##   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
##  848  903  841  608  605  564  618  599  509  498  434  448  425  384  375
##   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47
##  384  351  362  327  272  270  244  249  214  218  203  171  147  152  145
##   48   49   50   51   52   53   54   55   56   57   58   59   60   61   62
##  157  146  131  101  116  112   99   92   91   71   72   55   60   51   59
##   63   64   65   66   67   68   69   70   71   72   73   74   75   76   77
##   47   45   41   52   51   33   31   27   22   20   17   16   27   16   16
##   78   79   80   81   82   83   84   85   86   87   88   89   90   91   92
##   13   21   10   21   11    7   11    6   12    5    8    4    8    6    7
##   93   94   95   96   97   99  100  101  102  105  107  112  113  115  118
##   10    5    5    1    4    2    3    1    2    1    2    1    1    1    1
##  120  123  127  130  133  170
##    2    1    1    1    1    1
```

This tells us that four individuals had three notations in the IDS (modified), nine had four, etc. The guys with three (selected columns):

```r
idi <- names(xx[xx == 3])
indiv[indiv$Id_I %in% idi, c("Id_I", "Type", "Value", "Day", "Month", "Year")]
```

```
##           Id_I                 Type
## 17785 13884863           BIRTH_DATE
## 17786 13884863       BIRTH_LOCATION
## 17789 13884863 CHILDBIRTH_ASSISTANT
## 48927 13635521           BIRTH_DATE
## 48928 13635521       BIRTH_LOCATION
## 48931 13635521 CHILDBIRTH_ASSISTANT
## 65928 15666384           BIRTH_DATE
## 65929 15666384       BIRTH_LOCATION
## 65932 15666384 CHILDBIRTH_ASSISTANT
## 84346 18095053           BIRTH_DATE
## 84347 18095053       BIRTH_LOCATION
## 84350 18095053 CHILDBIRTH_ASSISTANT
##                                            Value Day Month Year
## 17785                                             16     7 1818
## 17786                        Missing information  16     7 1818
## 17789                                    Unknown  16     7 1818
## 48927                                              6     9 1872
## 48928                                SKELLEFTEÅ    6     9 1872
## 48931                                    Unknown   6     9 1872
## 65928                                             23     6 1894
## 65929 NORDAMERIKA (FÖRENTA STATERNA INKL CANADA)  23     6 1894
## 65932                                    Unknown  23     6 1894
## 84346                                             29     5 1871
## 84347                                SKELLEFTEÅ   29     5 1871
## 84350                           Midwife delivery  29     5 1871
```

Obviously only “birth” information. In contrast, the guy with 170 notations (watch up, several pages!):

```
idi <- names(xx[xx == 170])
indiv[indiv$Id_I %in% idi, c("Id_I", "Type", "Value", "Day", "Month", "Year")]
```

```
##             Id_I                 Type                Value Day Month Year
## 1588520 19690322           BIRTH_DATE                        9     8 1793
## 1588521 19690322       BIRTH_LOCATION          SKELLEFTEÅ    9     8 1793
## 1588522 19690322         BAPTISM_DATE                        0     0 1793
## 1588523 19690322     BAPTISM_LOCATION          SKELLEFTEÅ    0     0 1793
## 1588524 19690322 CHILDBIRTH_ASSISTANT              Unknown   9     8 1793
## 1588530 19690322           OCCUPATION                 PIG.   0     0    0
## 1588531 19690322           OCCUPATION                 PIG:   0     0    0
## 1588532 19690322           OCCUPATION                PIGAN   0     0    0
## 1588533 19690322           OCCUPATION                PIGAN   0     0    0
## 1588534 19690322           OCCUPATION                 PIGA   0     0    0
## 1588535 19690322           OCCUPATION                PIGAN   0     0    0
## 1588536 19690322           OCCUPATION                 PIGA   0     0    0
## 1588537 19690322           OCCUPATION                 PIGA   0     0    0
## 1588538 19690322           OCCUPATION                 PIGA   0     0    0
## 1588539 19690322           OCCUPATION                 PIGA   0     0    0
## 1588540 19690322           OCCUPATION                 PIGA   0     0    0
## 1588541 19690322           OCCUPATION                 PIG.   0     0    0
## 1588542 19690322           OCCUPATION                 PIGA   0     0    0
## 1588543 19690322           OCCUPATION                 PIG.   0     0    0
## 1588544 19690322           OCCUPATION                PIGAN   0     0    0
## 1588545 19690322           OCCUPATION                 PIG.   0     0    0
## 1588546 19690322           OCCUPATION                PIGAN   0     0    0
## 1588547 19690322           OCCUPATION                 PIGA   0     0    0
## 1588548 19690322           OCCUPATION                 PIG.   0     0    0
## 1588549 19690322           OCCUPATION                 PIGA   0     0    0
## 1588550 19690322           OCCUPATION                 PIGA   0     0    0
## 1588551 19690322           OCCUPATION                 PIG.   0     0    0
## 1588552 19690322           OCCUPATION                 PIG.   0     0    0
## 1588553 19690322           OCCUPATION                  PIG   0     0    0
## 1588554 19690322           OCCUPATION                 PIG.   0     0    0
## 1588555 19690322           OCCUPATION                  PIG   0     0    0
## 1588556 19690322           OCCUPATION INH PIG DISTRICTHJON   0     0    0
## 1588557 19690322           OCCUPATION     INHS. FATTIGHJON   0     0    0
## 1588558 19690322           OCCUPATION        DISTRIKTSHJON   0     0    0
## 1588559 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588560 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588561 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588562 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588563 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588564 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588565 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588566 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588567 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588568 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588569 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588570 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588571 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588572 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588573 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588574 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588575 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588576 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588577 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588578 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588579 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588580 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588581 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588582 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588583 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588584 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588585 19690322  OCCUPATION_STANDARD        DISTRIKTSHJON   0     0    0
## 1588586 19690322  OCCUPATION_STANDARD              INHYSES   0     0    0
## 1588587 19690322  OCCUPATION_STANDARD                 PIGA   0     0    0
## 1588588 19690322  OCCUPATION_STANDARD           FATTIGHJON   0     0    0
## 1588589 19690322  OCCUPATION_STANDARD              INHYSES   0     0    0
## 1588590 19690322  OCCUPATION_STANDARD        DISTRIKTSHJON   0     0    0
## 1588591 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588592 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588593 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588594 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588595 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588596 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588597 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588598 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588599 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588600 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588601 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588602 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588603 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588604 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588605 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588606 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588607 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588608 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588609 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588610 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588611 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588612 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588613 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588614 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588615 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588616 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588617 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588618 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588619 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588620 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588621 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588622 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588623 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588624 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588625 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588626 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588627 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588628 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588629 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588630 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588631 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588632 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588633 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588634 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588635 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588636 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588637 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588638 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588639 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588640 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588641 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588642 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588643 19690322     OCCUPATION_HISCO          -1            0     0    0
## 1588644 19690322     OCCUPATION_HISCO          54020         0     0    0
## 1588645 19690322     OCCUPATION_HISCO          62120         0     0    0
## 1588646 19690322     OCCUPATION_HISCO          99920         0     0    0
## 1588647 19690322     OCCUPATION_HISCO          -1            0     0    0
## 1588648 19690322     OCCUPATION_HISCO          99920         0     0    0
## 1588649 19690322     OCCUPATION_HISCO          -1            0     0    0
## 1588650 19690322          VACCINATION           Vaccinated   0     0    0
## 1588651 19690322          VACCINATION Infected by smallpox   0     0    0
## 1588653 19690322          OBSERVATION              Present   0     0    0
## 1588654 19690322          OBSERVATION              Present   0     0    0
## 1588655 19690322          OBSERVATION              Present   0     0    0
## 1588656 19690322          OBSERVATION              Present   0     0    0
## 1588657 19690322          OBSERVATION              Present   0     0    0
## 1588658 19690322          OBSERVATION              Present   0     0    0
## 1588659 19690322          OBSERVATION              Present   0     0    0
## 1588660 19690322          OBSERVATION              Present   0     0    0
## 1588661 19690322          OBSERVATION              Present   0     0    0
## 1588662 19690322          OBSERVATION              Present   0     0    0
## 1588663 19690322          OBSERVATION              Present   0     0    0
## 1588664 19690322          OBSERVATION              Present   0     0    0
## 1588665 19690322          OBSERVATION              Present   0     0    0
## 1588666 19690322          OBSERVATION              Present   0     0    0
## 1588667 19690322          OBSERVATION              Present   0     0    0
## 1588668 19690322          OBSERVATION              Present   0     0    0
## 1588669 19690322          OBSERVATION              Present   0     0    0
## 1588670 19690322          OBSERVATION              Present   0     0    0
## 1588671 19690322          OBSERVATION              Present   0     0    0
## 1588672 19690322          OBSERVATION              Present   0     0    0
## 1588673 19690322          OBSERVATION              Present   0     0    0
## 1588674 19690322          OBSERVATION              Present   0     0    0
## 1588675 19690322          OBSERVATION              Present   0     0    0
## 1588676 19690322          OBSERVATION              Present   0     0    0
## 1588677 19690322          OBSERVATION              Present   0     0    0
## 1588678 19690322          OBSERVATION              Present   0     0    0
## 1588679 19690322          OBSERVATION              Present   0     0    0
## 1588680 19690322          OBSERVATION              Present   0     0    0
## 1588681 19690322          OBSERVATION              Present   0     0    0
## 1588682 19690322          OBSERVATION              Present   0     0    0
## 1588683 19690322          OBSERVATION              Present   0     0    0
## 1588684 19690322          OBSERVATION              Present   0     0    0
## 1588685 19690322          OBSERVATION              Present   0     0    0
## 1588686 19690322          OBSERVATION              Present   0     0    0
## 1588687 19690322          OBSERVATION              Present   0     0    0
## 1588688 19690322          OBSERVATION              Present   0     0    0
## 1588689 19690322    START_OBSERVATION                Birth   9     8 1793
## 1588690 19690322      END_OBSERVATION           End source   0     0 1870
## 1588692 19690322         CIVIL_STATUS            Unmarried   0     0    0
## 1588695 19690322           DEATH_DATE                       30     4 1871
## 1588696 19690322       DEATH_LOCATION          SKELLEFTEÅ   30     4 1871
## 1588697 19690322         FUNERAL_DATE                        7     5 1871
## 1588698 19690322     FUNERAL_LOCATION  Missing information   7     5 1871
```

Almost 80 years of information . . .. What are the different possible values of *Type*?

```
sort(unique(indiv$Type))
```

```
##  [1] "ARRIVAL_FROM"          "BAPTISM_DATE"          "BAPTISM_LOCATION"
##  [4] "BIRTH_DATE"            "BIRTH_LOCATION"        "CHILDBIRTH_ASSISTANT"
##  [7] "CIVIL_STATUS"          "DEATH_DATE"            "DEATH_LOCATION"
## [10] "DEPARTURE_TO"          "END_OBSERVATION"       "FUNERAL_DATE"
## [13] "FUNERAL_LOCATION"      "MARRIAGE_DATE"         "MARRIAGE_LOCATION"
## [16] "OBSERVATION"           "OCCUPATION"            "OCCUPATION_HISCO"
## [19] "OCCUPATION_STANDARD"   "START_OBSERVATION"     "VACCINATION"
```

Dates

Let's fix dates in more readable form.

```
datum <- with(indiv, paste(Year, Month, Day, sep = "-"))
indiv$datum <- as.Date(datum, format = "%Y-%m-%d")
datum <- with(indiv, paste(Start_year, Start_month, Start_day, sep = "-"))
indiv$enter <- as.Date(datum, format = "%Y-%m-%d")
datum <- with(indiv, paste(End_year, End_month, End_day, sep = "-"))
indiv$exit <- as.Date(datum, format = "%Y-%m-%d")
indiv <- select(indiv, Id_I, Type, Value, datum, enter, exit)
indiv <- filter(indiv, !(Type %in% c("BAPTISM_LOCATION", "CHILDBIRTH_ASSISTANT",
                                     "FUNERAL_DATE", "FUNERAL_LOCATION",
                                     "MARRIAGE_LOCATION", "VACCINATION")))
indiv <- filter(indiv, !(Type %in% c("OCCUPATION", "OCCUPATION_HISCO")))
kable(head(indiv))
```

Id_I	Type	Value	datum	enter	exit
13884863	BIRTH_DATE		1818-07-16	NA	NA
13884863	BIRTH_LOCATION	Missing information	1818-07-16	NA	NA
10008393	BIRTH_DATE		1777-03-07	NA	NA
10008393	BIRTH_LOCATION	SKELLEFTEÅ	1777-03-07	NA	NA
10008393	BAPTISM_DATE		1777-03-08	NA	NA
10022921	BIRTH_DATE		1871-02-08	NA	NA

```
anydate <- indiv$datum
omiss <- is.na(anydate)
```

```
anydate[omiss] <- indiv$exit[omiss]
indiv$anydate <- anydate
rm(anydate)
length(unique(indiv$Id_I))
```

```
## [1] 35567
```

```
indiv <- filter(indiv, !is.na(anydate))
length(unique(indiv$Id_I))
```

```
## [1] 35567
```

We remove rows with *Type == “BAPTISM_DATE”, “BIRTH_LOCATION”, “DEATH_LOCATION*

```
indiv <- filter(indiv, !(Type %in% c("BAPTISM_DATE", "BIRTH_LOCATION", "DEATH_LOCATION")))
with(indiv, table(Type))
```

```
## Type
##        ARRIVAL_FROM          BIRTH_DATE        CIVIL_STATUS
##                3258               35567               31709
##          DEATH_DATE        DEPARTURE_TO     END_OBSERVATION
##               11580                6012               14213
##       MARRIAGE_DATE         OBSERVATION OCCUPATION_STANDARD
##               10502               16724               18160
##   START_OBSERVATION
##               34013
```

Now we check the min and max date for each person.

```
mind <- with(indiv, tapply(anydate, Id_I, min))
maxd <- with(indiv, tapply(anydate, Id_I, max))
wox <- data.frame(id = names(maxd),
                  first = as.Date(mind, origin = "1970-01-01"),
                  last = as.Date(maxd, origin = "1970-01-01"),
                  age = (maxd - mind) / 365.2425)
wox$birthYear <- wox$first %>%
    as.character() %>%
    substr(1, 4) %>%
    as.numeric()
id <- indiv$Id_I[indiv$Type == "DEATH_DATE"]
##wox$death <- FALSE
wox$death <- wox$id %in% id
```