Combining register and survey information in the Dutch Census 2011

Eric Schulte Nordholt
Senior researcher and project leader of the Census
Statistics Netherlands
Division Socio-economic and spatial statistics

e.schultenordholt@cbs.nl

4th Baltic Nordic Conference on Survey Statistics in Helsinki
24-28 August 2015

Contents (1)

- History of the Dutch Census
- The Dutch Census of 2011
- Data sources
- Combining sources: micro linkage
- Combining sources: micro integration
- Conditions facilitating use of registers
- Census tables
- Micro macro method
- Result on 2011 economic activity

Contents (2)

- Comparison with other countries
- Comparison with other years
- Promoting the 2011 Census
- Conclusions

History of the Dutch Census (1)

TRADITIONAL CENSUS

Ministry of Home Affairs:

1829, 1839, 1849, 1859, 1869, 1879 and 1889

Statistics Netherlands:

1899, 1909, 1920, 1930, 1947, 1960 and 1971

Unwillingness (nonresponse) and reduction expenses ⇒ no more traditional censuses

History of the Dutch Census (2)

ALTERNATIVE: VIRTUAL CENSUS

1981 and 1991: limited virtual censuses based on Population Register and surveys

development 1990's: more registers → integrated set of registers and surveys, SSD

2001 and 2011: complete virtual censuses based on the SSD with information at the municipality level

The Dutch Census of 2011

is based on the System of social statistical datasets (SSD) which

- **is a set of integrated microdata files with coherent and detailed demographic and socio-economic data on persons, households, jobs and benefits**
- **has no remaining internal conflicting information**

is part of the European Census

- **Eurostat: coordinator of EU, accession and EFTA countries in the European Census Rounds**
- **Census Table Programme, every 10 years**

Social statistics in the Netherlands develop in the direction of a permanent Virtual Census to be able to produce:

- **More crosstables over different domains**
- **More longitudinal information**
- **More flexible policy relevant output**

Data sources

Registers:

- **Population Register (PR) →**
  - illegal people excluded
  - homeless counted at last known address
- **Jobs file, containing all employees**
- **Self-employed file, containing all self-employed**
- **Fiscal administration**
- **Social Security administrations**
- **Pensions and life insurance benefits**
- **Housing registers**

Survey:

- **Labour Force Survey microdata around Census Day**

Combining sources: micro linkage

- **Linkage key:**

  Registers

  Citizen Service Number, unique

  Surveys

  Sex, date of birth,

  address (postal code and house number)

- **Linkage key replaced by RIN-person**

- **Linkage strategy**

  Optimizing number of matches

  Minimizing number of mismatches and missed matches

Combining sources: micro integration

- **Collecting data from several sources ⇒ more comprehensive and coherent information on aspects of a person's life**
- **Compare sources**
  - coverage
  - conflicting information (reliability of sources)
- **Integration rules**
  - checks
  - adjustments
  - imputations
- **Optimal use of information ⇒ quality improves**
- **Example: job period vs. benefit period**

Conditions facilitating use of registers

- Legal base (Statistics Act)
- Public approval (‘Big Brother is watching you’)
- Cooperation among authorities (mainly government organisations)
- Comprehensive and reliable register system (administrative versus statistical quality)
- Unified identification system (preferably unique ID-numbers)

Census tables (1)

Preliminary work before tabulating

Census Programme definitions:
not always clear and unambiguous, e.g. economic activity

Priority rules

- (characteristics of) main job (highest wage)
- employee or employer
- job or (partially) unemployed
- job or attending education
- job or retired
- engaged in family duties or retired
- age restrictions

Tabulating register variables:

Simply straightforward counting from SSD register data

Census tables (2)

Tabulating survey (and register) variables

Mass imputation?

- Pro's: reproducible results
- Con's: danger of oddities in estimates (e.g. highly educated baby)

Traditional Weighting?

- Pro's: simple, reproducible results (if same microdata and weights)
- Con's: no overall numerical consistency between survey and register estimates

Demand for overall numerical consistency

- one figure for one phenomenon idea
- all tables based on different sources (e.g. surveys) should be mutually consistent

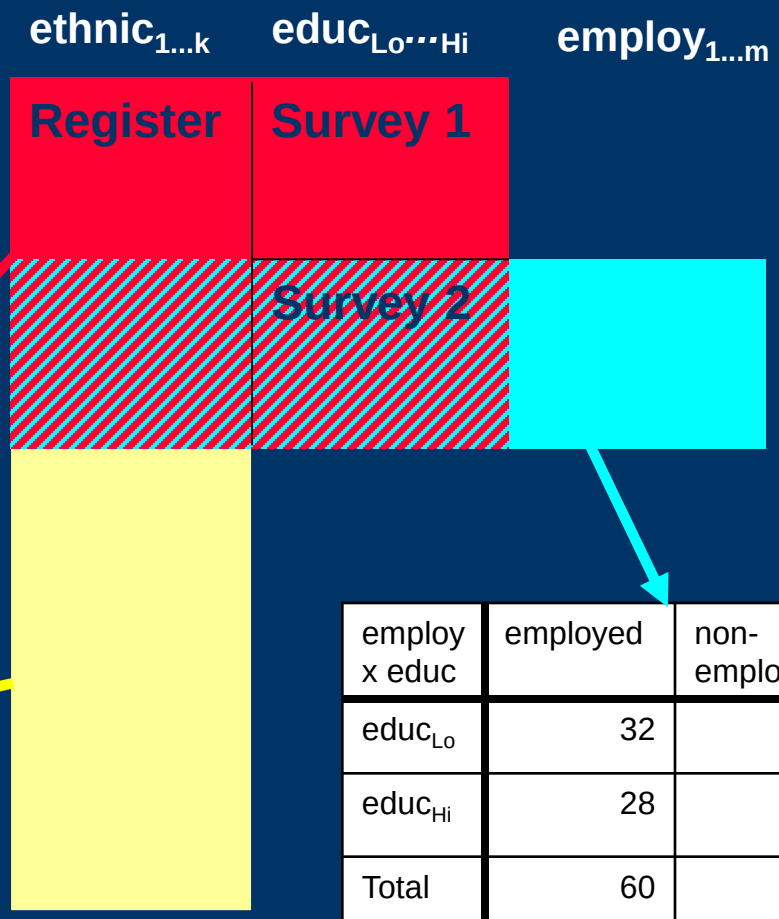
Census tables (3)

Ethnicity: register

Education: survey 1 and survey 2

Employment status: survey 2

Estimate: T1: educ x ethnic and T2: educ x employ



educ x ethnic	not-NL	NL	Total
$educ_{Lo}$	20	29	49
$educ_{Hi}$	9	42	51
Total	29	71	100

ethnic	not-NL	NL
Total	30	70

employ x educ	employed	non-employed	Total
$educ_{Lo}$	32	20	52
$educ_{Hi}$	28	20	48
Total	60	40	100

Census tables (4)

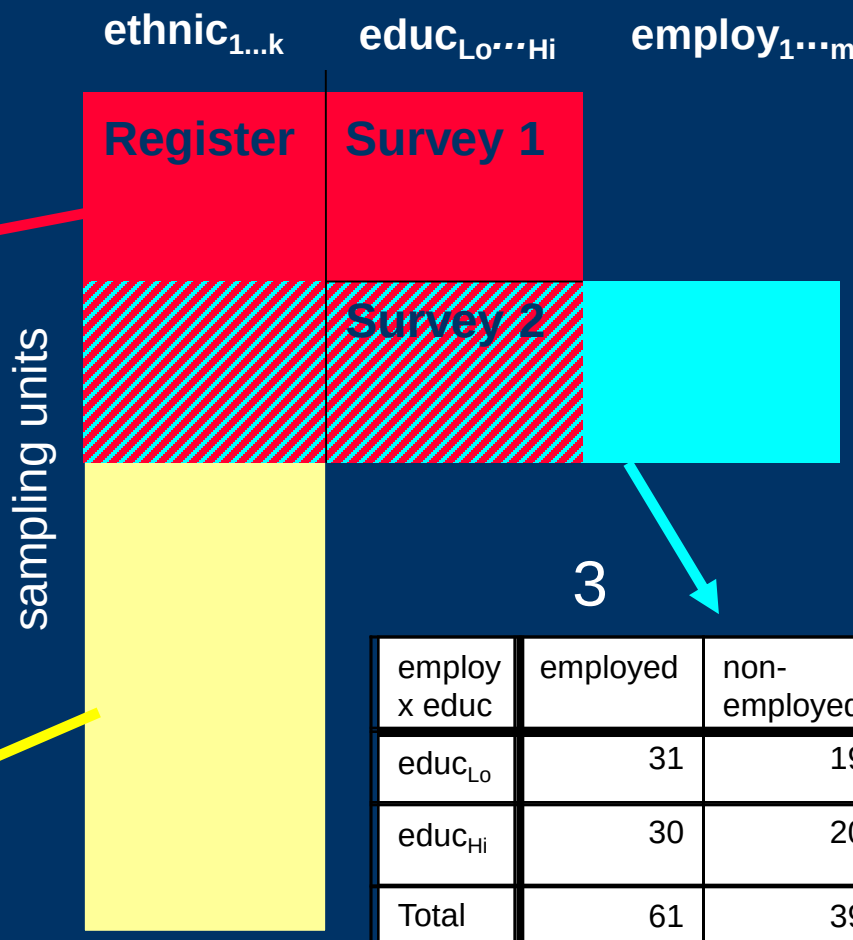
Repeated Weighting (RW) : tool to achieve numerical consistency (VRD-software)

Basic principles of RW:

- estimate table on most reliable source (mostly source with most records, e.g. register)
- estimate tables by calibrating on common margins of the current table and tables already estimated (auxiliary information)
- repeatedly use of regression estimator:
  - initial weights (e.g. survey weights) calibrated as minimal as possible
  - lower variances
  - no excessive increase of (non-response) bias (as long as cell size>>0)
- each table has its own set of weights

Census tables (5)

Calibrate on ethnic, then on educ x ethnic



2

educ x ethnic	not-NL	NL	Total
$educ_{Lo}$	20	30	50
$educ_{Hi}$	10	40	50
Total	30	70	100

1

ethnic	not-NL	NL
Total	30	70

3

employ x educ	employed	non-employed	Total
$educ_{Lo}$	31	19	50
$educ_{Hi}$	30	20	50
Total	61	39	100

Micro macro method (1)

Repeated Weighting works nicely, but in the 2011 Census a new requirement was introduced: hypercubes (= high dimensional tables)

Problem:
Very detailed tables contain many sample zeros that RW cannot handle

Solution 1: estimate subhypercubes
Solution 2: micro macro method (an IPF method) was introduced to estimate the interior of subhypercubes containing LFS variables

Micro macro method (2)

Results of the micro macro method are published if two conditions are fullfilled:

1. table margins estimated with RW are small enough

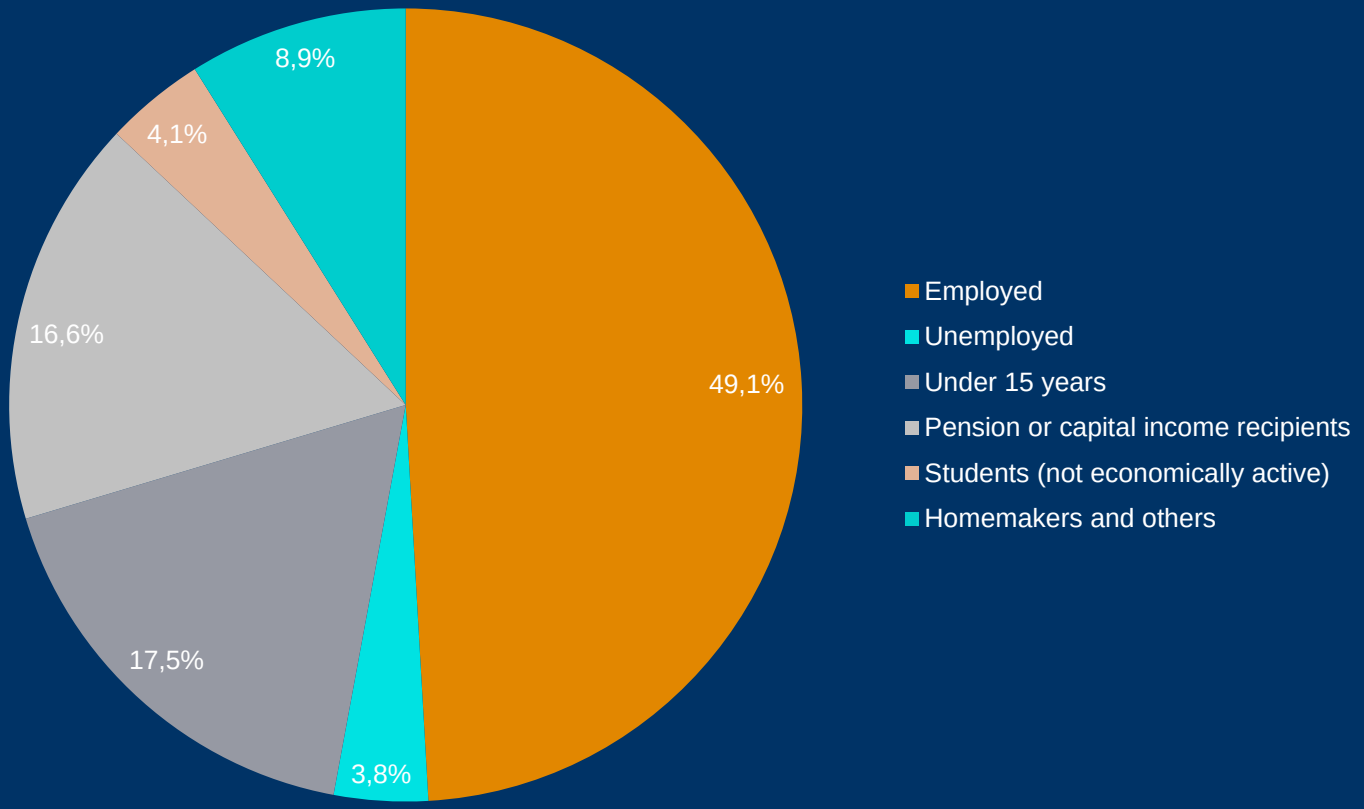
2. number of records in estmated cells are large enough

Criteria:

1. estimated relative inaccuracy of at most 20 percent (i.e. the estimated margins amount to 40 percent at most) which corresponds to a threshold of 25 persons

2. only table cells based on 5 or more persons are published

Result on 2011 economic activity



Comparison with other countries

Traditional Census (complete enumeration):

- Most countries in the world (including the UK and the US)

Traditional Census (partial enumeration) and Registers:

- Some countries (e.g. Germany, Poland and Switzerland)

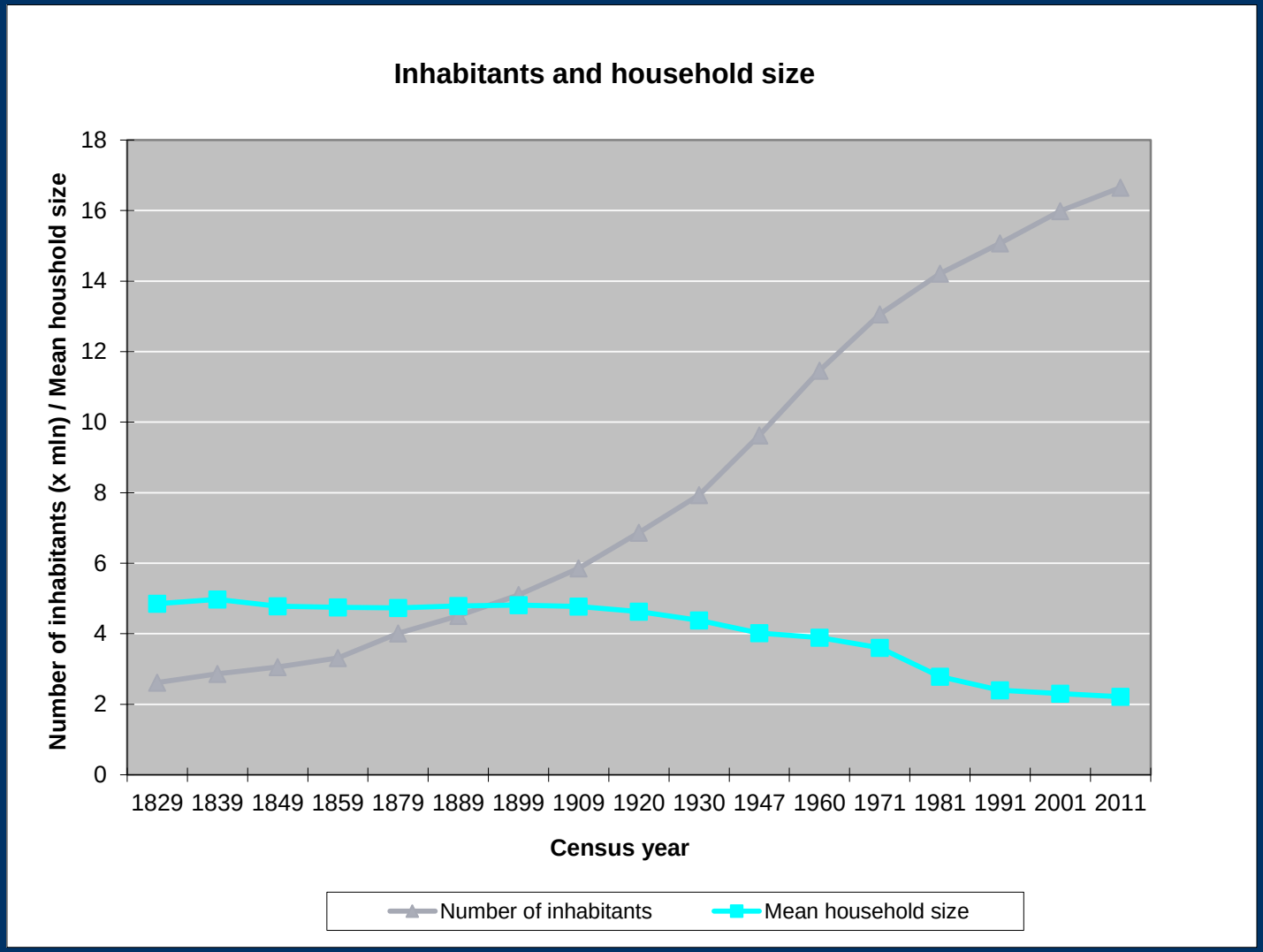
Rolling Census:

- France

Fully or largely register-based (Virtual) Census:

- Five Nordic countries (Iceland,Norway, Sweden, Finland and Denmark), the Netherlands, Belgium, Austria and Slovenia

Comparison with other years



Promoting the 2011 Census

Dutch Census Book

http://www.cbs.nl/en-GB/menu/themas/dossiers/historische-reeksen/publicaties/publicaties/archief/2014/2014-dutch-census-2011-pub.htm?Languageswitch=on

Census Hub:

https://ec.europa.eu/CensusHub2/

Conclusions (1)

Advantages:

- Relatively cheap (small cost per inhabitant)
- Quick (short production time)

Disadvantages:

- Dependent on register holders (statistics is not their priority), timeliness of registers, concepts and population of registers may differ from what is needed (keep good relations with the register holders!)
- Publication of small subpopulations sometimes difficult or even impossible because of limited information

Conclusions (2)

Other aspects:

- **Less attention for the results of a virtual census than for a traditional one**
- **Difficult to keep knowledge and software up-to-date (Census is running every ten years)**
- **Enormous international interest in virtual censuses**
- **A lot of interesting census work in the coming years!**

Thank you for your attention!

Time for questions and discussion