Quality assessment of census results

Eric Schulte Nordholt

Statistics Netherlands

Contents (1)

- Quality in official statistics
- Data collection methods
- Costs
- Response burden
- Product quality
- Register-based statistics compared to statistical surveys
- Increasing role of administrative data in the statistical process
- Quality, administrative data, and the statistical process
- Data confidentiality

Contents (2)

- Combining research
- Data considerations in the Dutch Census 2011
- Results Source hyper dimension
- Results Metadata hyper dimension
- Results Data hyper dimension
- Conclusions from combined research
- Educational Attaiment File
- Imputing the Educational Attainment File
- Wrapping up

Quality in official statistics

- Definition of quality in statistics according to European "Code of practice"
- Product quality
  - Relevance
  - Accuracy
  - Timeliness and punctuality
  - Comparability and coherence
  - Accessibility and clarity
- Process quality
  - Best methods
  - Cost efficiency
  - Low response burden

Data collection methods

Options
- Census/full coverage statistical survey
- Sample survey
- Administrative registers

Administrative data are collected for administrative purposes
- Register-based statistics is secondary use of existing data

Decision on data collection method is a compromise between
- Cost efficiency
- Response burden
- Product quality

Costs

Current situation in many countries
- The NSIs have experienced budget cuts / restrictions
- Users demand new and more detailed statistics
- Must increase efficiency in production of statistics

Administrative data
- Almost no costs for data collection (for the NSI)
- Use resources on improving existing data instead of collecting data for statistical purposes
  - Supplement and correct existing data
  - Most resources used in establishing register-based statistical systems
  - But: systems must be maintained

Register-based statistics is not free of charge but normally less expensive than sample surveys and especially than traditional censuses

Response burden

Use of administrative data means no additional response burden

- For companies
  - "Reporting to authorities takes too much time"
- For citizens
  - "The authorities should not ask for information that I have already given"
- For the NSI
  - Increasing non-response problems in sample surveys and censuses

Product quality (1)

Relevance

- Register data is based on administrative definitions that may differ from statistical definitions
  - Units, coverage, variables, time references etc.
- "We have the right answers, but can we answer the right questions?"
- "The authorities picture of the world?"
- Combining data from different registers to improve relevance
- In some cases: additional data collection is necessary

Product quality (2)

Accuracy

- Registers normally have good quality for administrative purposes
- Improving accuracy by combining data from several registers
  - Editing for statistical purposes

Product quality (3)

Timeliness and punctuality

- Production time sometimes longer than for statistical surveys
  - Administrative process may take time (example: tax data)
  - Delay in updating of registers
  - Data extraction: necessary to wait some weeks or months and sometimes even longer

Product quality (4)

Comparability and coherence

- Building a coherent register-based statistical system
- Harmonising with statistics based on other sources
  - Experiences in the Netherlands

Accessibility and clarity

- Almost independent of data sources used

Register-based statistics compared to statistical surveys (1)

- Costs (++)
- Response burden (++)
- Relevance (-)
  - Not all variables are included in registers
  - Less direct control over data content
- Accuracy (o)
- Timeliness (-)

Register-based statistics compared to statistical surveys (2)

Administrative registers offer

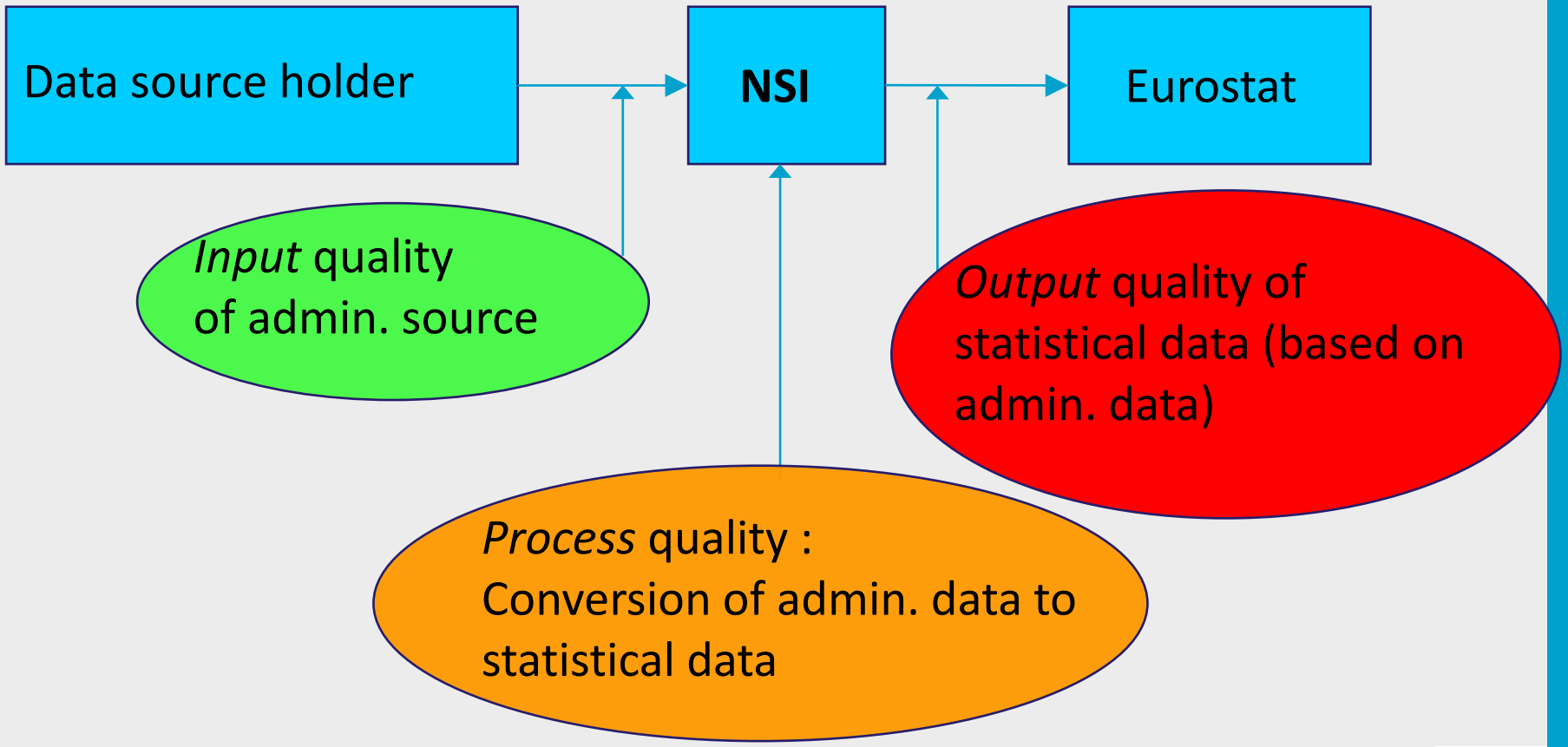
- Total coverage at a low cost
  - Statistics for small groups possible (compared to sample surveys)
- Annual (or more frequent) data for all variables
  - Annual "censuses"
- To produce statistics based on administrative data has proved to be efficient
- Register-based statistics have to be supplemented by information from sample surveys

Increasing role of administrative data in the statistical process

- More and more statistical institutes are using administrative sources for statistical purposes
  - Mainly to decrease costs and response burden

- However, as a result they:
  1. Become more *dependent* on data sources collected and maintained by *others*
     - Need to monitor the *quality* of those data sources when they *enter* the office
  2. Have to find new data sources that contain the information needed
     - Need to evaluate the usability of those data sources *prior* to use

Quality, administrative data, and the statistical process



Data confidentiality (1)

Legislation:

- Statistics Netherlands Act
- Netherlands Data Protection Act → since 2018 GDPR

These laws:

- authorize Statistics Netherlands to use personal data
- oblige Statistics Netherlands to take adequate measures aimed at privacy protection

Data confidentiality (2)

Measures:

- Linkage keys are anonymous, original personal identifiers are removed from the data
- Access rights to the microdata are restricted

Combining research

Development of a quality framework for administrative data

Data decisions on secondary sources in the Dutch Census of 2011

Data considerations in the Dutch Census 2011

Registers:
– Population Register (PR), 17 million records
– Jobs file, containing all employees
– Self-employed file, containing all self-employed
– Unemployment Benefit Register (UR)
– Social Security Register (SR)
– Education Register (ER)
– New Housing Register (HR)

Surveys:
– Labour Force Survey (LFS)

Results Source hyperdimension

Dimensions	*Data sources*				
	ER	UR	SR	HR	PR
1. Supplier	+	0	0	+	+
2. Relevance	0	+	+	0	+
3. Privacy and security	+	+	+	+	+
4. Delivery	-	+	+	+	+
5. Procedures	0	0	0	+	+

Low frequency of delivery

Suffers seriously from selective undercoverage

Purpose dataprovider unclear

Important variables are missing

Results Metadata hyperdimension

Dimensions	*Data sources*				
	ER	UR	SR	HR	PR
1. Clarity	+	+	+	+	+
2. Comparability	-	o	o	+	+
3. Unique keys	+	+	+	o	+
4. Data treatment	+	+	+	+	+

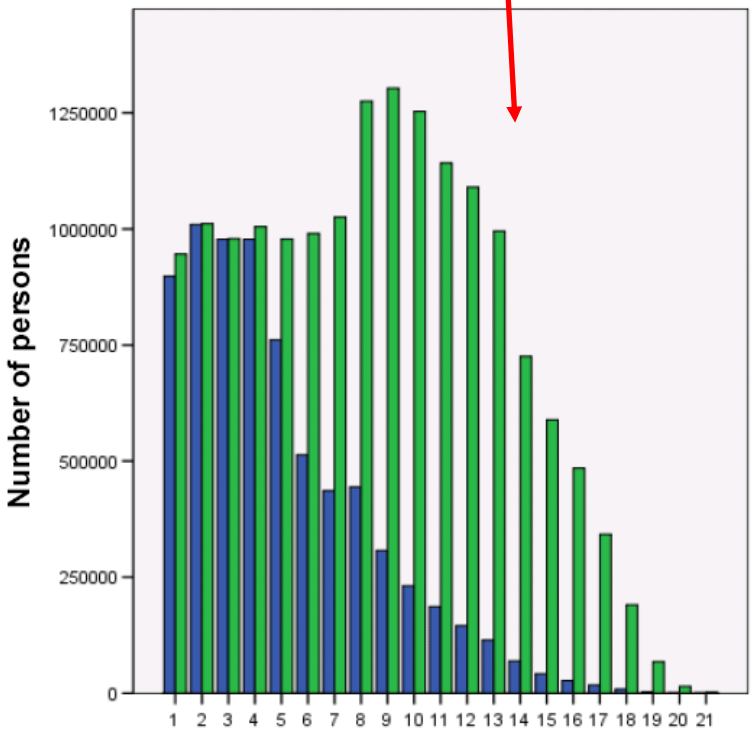
Time period in source can't be transferred easily to the time point needed

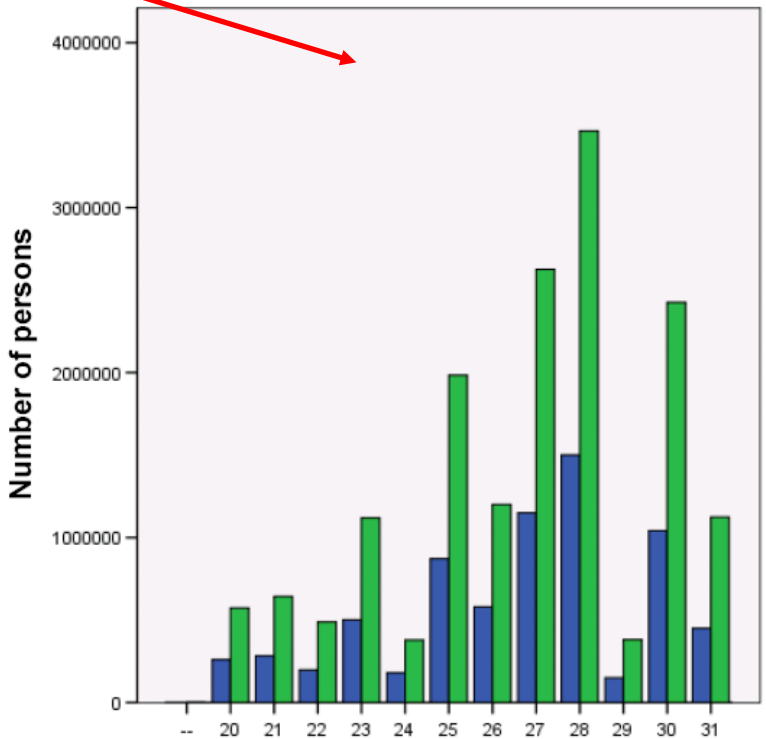
Time differences in reporting periods

Unique keys can't be easily used for linking

Results Data hyperdimension - completeness

Variable	*Number of missings*	*Percentage missing (%)*
Educational attainment	9.238.212	56,3
Current activity status	2.140.266	13,0





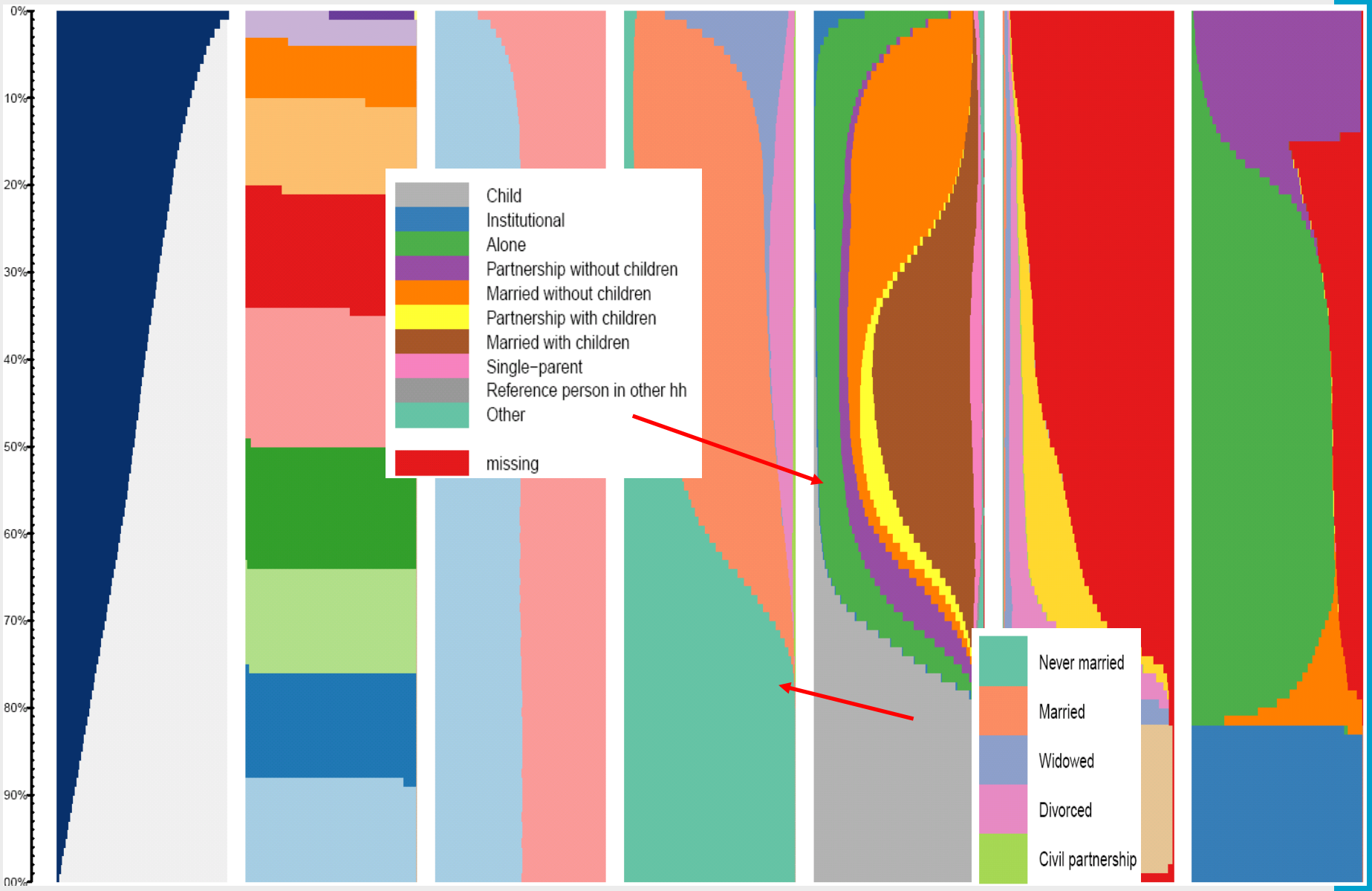
= Persons having a known education level

= All persons

Results Data hyperdimension – accuracy

Ageclass	Current activity status							
	Missing	0	1	2	3	4	5	6
1: [0, 5)	0	945861	0	0	0	0	0	0
2: [5, 10)	0	1011159	0	0	0	0	0	0
3: [10, 15)	0	978964	0	0	0	0	0	0
4: [15, 20)	34911	0	482180	33	0	487533	11	293
5: [20, 25)	113286	0	716411	106	0	147395	190	711
6: [25, 30)	142149	0	818167	107	0	28396	486	677
7: [30, 35)	163141	0	856030	129	0	4506	744	771
8: [35, 40)	216807	0	1053407	180	0	2418	1138	1056
9: [40, 45)	228634	0	1070204	228	0	1853	1076	1224
10: [45, 50)	236102	0	1013249	242	0	1134	1076	1434
11: [50, 55)	262473	0	875724	253	1	504	1261	1789
12: [55, 60)	330898	0	714959	263	39705	232	1776	2253
13: [60, 65)	390062	0	343089	122	256826	78	2348	2764
14: [65, 70)	8730	0	88209	1	628490	16	3	46
15: [70, 75)	5306	0	35690	1	548059	3	0	22
16: [75, 80)	3822	0	14705	0	466339	2	0	19
17: [80, 85)	2166	0	5897	0	333936	0	0	8
18: [85, 90)	1115	0	2360	0	186690	0	0	8
19: [90, 95)	405	0	662	0	66339	0	0	0
20: [95, 100)	162	0	136	0	14386	0	0	0
21: [100, ∞)	97	0	18	0	1450	0	0	0

[4] Current activity status: (0). Persons below minimum age for economic activity, (1) Employed, (2) Unemployed, (3) Pension or capital income recipients, (4) Students not economically active, (5) Homemakers, (6) Others


0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
00%
Child
Institutional
Alone
Partnership without children
Married without children
Partnership with children
Married with children
Single-parent
Reference person in other hh
Other
missing
Never married
Married
Widowed
Divorced
Civil partnership

Conclusions from combined research

– Quality of official statistics is an important aspect, especially when use is made of integrated data

– The virtual census has proved to be a successful concept in the Netherlands

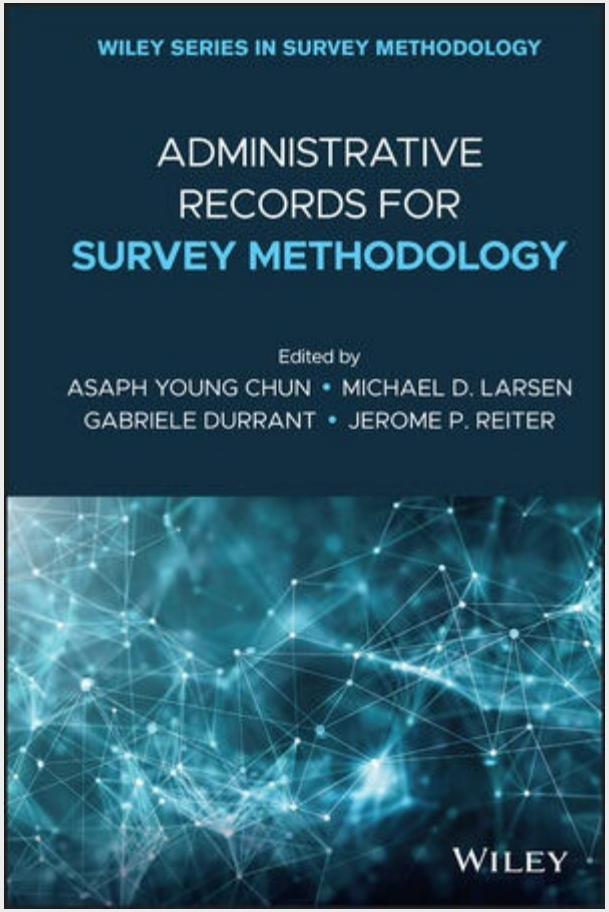
– The quality framework is a useful tool for making data decisions in the virtual census

– Follow-up: the Education Register has been improved (including newer data and foreign and private education) and is being used in the Census 2021

Conclusions from combined research

More information:
Daas, P., E. Schulte Nordholt, M. Tennekes and S. Ossen, 2021. Evaluation of the Quality of Administrative Data Used in the Dutch Virtual Census. In: Administrative Records for Survey Methodology, Wiley Online Library, 2021, pp. 63-83.

ISBN: 978-1-119-27204-5
April 2021
384 Pages



Educational Attainment File

- Complex integration process of microdata from LFS and examination registers
- New version containing also information on private education institutions available since 2016
- About 60 % of the records have information on highest level of education attained
- Weighting to known marginals of the population for statistics on level of education
- Better quality and more detailed education tables than before when only LFS information was used

Imputing the Educational Attainment File

In this project:

- Find a good imputation model for the Educational Attainment File (logistic regression model)
- Produce a set of hypercubes of the Census 2011 again, now by using the imputed Educational Attainment File
- Develop a set of quality indicators (basis for decision how detailed the future census publication will be)
- Plan for highest level of education attained in the Dutch Census 2021 (also of interest for other countries)

Wrapping up

- Administrative data quality evaluation and data confidentiality was the theme
- Many things can be explained, some things have to be experienced in your own (national) context
- Quality in an administrative context is very different from quality in a survey context
- Use of administrative data in future censuses will increase
- A lot of interesting work for the future!

Thank you very much for your attention!