



# The integration of educational attainment data from administrative and survey data

part 2: Estimation of census tables with mass imputation

Jacco Daalmans

22 August 2023

# Dutch virtual census





# Backgrounds

Every 10 years

About 60 tables about people and houses



# Virtual census

- Data sources already available at Statistics Netherlands
- No single data source contains all census variables.
- All data can be linked at micro level
- All variables available from (integral) administrative registers except:
  - Educationale attainment: approx. 60% of population covered
  - Occupation: approx. 3% of population covered
- Missing data need to be estimated: weighting, (mass) imputation

# Virtual census

- Data sources already available at Statistics Netherlands
- No single data source contain all census variables.
- All data can be linked at micro level
- All variable available from (integral) administrative registers except
  - Educationale attainment: approx. 60% of population covered
  - Occupation: approx. 3% of population covered
- Missing data need to be estimated: weighting vs (mass) imputation

# Mass imputation

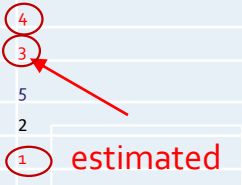
# (Mass) imputation

- Estimate all missing information for each individual person

Compile census tables from completely filled data

Easy and attractive: results can be made for detailed subpopulations

Age	Sex	Education
Y	M	1
Y	M	8
Y	M	2
Y	M	3
Y	F	4
Y	F	3
Y	F	5
Y	F	2
Y	F	1
Y	F	7
Y	F	2
Y	F	3
Y	F	4
Y	F	3
Y	F	2
Y	F	6
O	M	3
O	M	4
O	M	8
O	M	1
O	F	2
O	F	4
O	F	5
O	F	8





# Mass imputation



Erroneous conclusions after imputation:  
“Dog owners who never buy dog food”

# Mass imputation

- is applied for education, but not for occupation, because of higher data coverage (60% versus 3%)
- Risks of inappropriate conclusions are limited:
  - a) all census variables are used as auxiliary variables  
-> all relevant relations are taken into account
  - b) imputations are especially made for the census.  
These are deleted afterwards to avoid misuse for other applications.

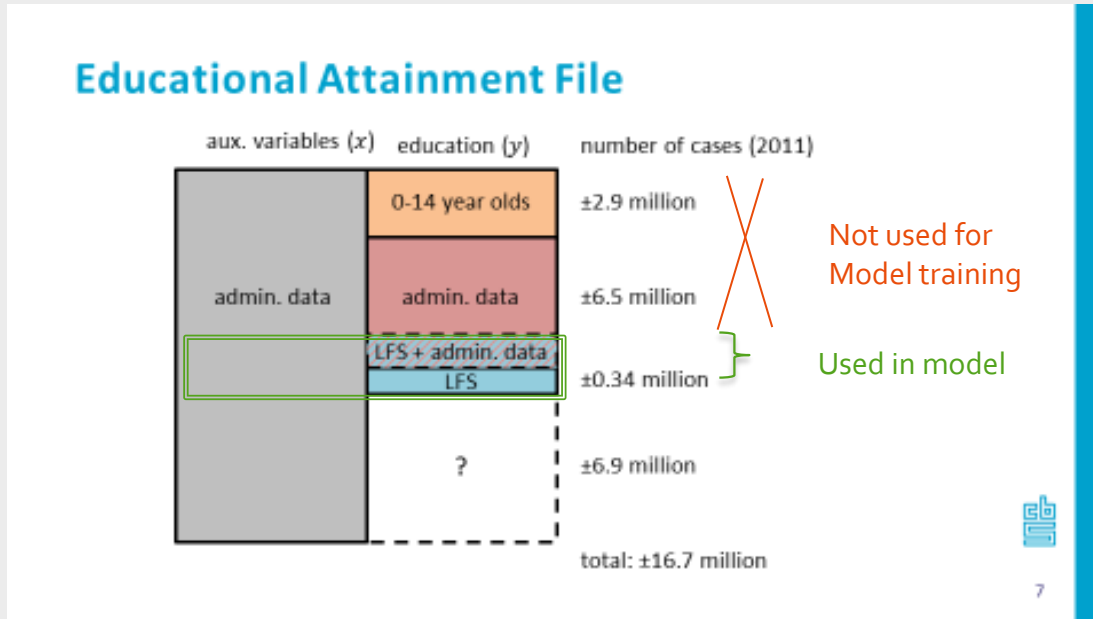
# Imputation: multinomial logistic regression

- Estimate model from people with available education.  
Apply to people with missing education
- Explanatory variables: Other census variables & income
- Estimate probabilities for each of the 8 education categories for each person given the characteristics of that person
- Use the probabilities to derive imputations (stochastically)



# Imputation multinomial regression

Imputation model estimated from Educational Attainment File (EAF)



The admin data within EAF are not used for model estimation due to selectivity<sup>12</sup> (overrepresentation of older and higher educated people)

# Research on imputation method

## Machine Learning versus regression

- Machine learning (Gradient boosting & neural networks) better estimates the distribution of educational attainment for the entire Dutch population
- However, for specific subpopulations (e.g. 53 year old men) regression works much better, due to implicit variable selection in ML methods
- The main of the census is to count all kinds of subpopulations
- Therefore, regression is the most appropriate method (at this moment)

# The end

– Thank you for listening

