Simulation as a way to combine information from multiple sources

Maarten L. Buis

Institut für Soziologie
Eberhard Karls Universität Tübingen

An example application

- We want to look at social reproduction of eduction
  - that is: We want to take a sample of women and look at the education of their children
- We have a sample of children and the education of their mother, and
- We know quite a bit about when women get children.
- Can't we combine these two bits of information to get the estimate we want?

Inequality of Educational Opportunity versus Social Reproduction

- Inequality of Educational Opportunity takes the offspring and “looks back” at the parents, while reproduction takes the parents and “looks forward” towards the offspring.
- Inequality of Educational Opportunity $\rightarrow$ How close is a country to the meritocratic model?
- Reproduction of Educational categories $\rightarrow$ What are the long term processes that shape the structure of society?

Aim of the talk

- Aim of this talk is partly methodological.
  - How to combine information from different sources using simulation.
  - Build intuition on the “empirical content” of the results of such a simulation.
- and is partly substantive.
  - People who were born more recently are more likely to attain higher levels of education
  - People who have higher educated parents are themselves more likely to attain higher levels of education
  - Does educational expansion lead to further educational expansion in the next cohort?

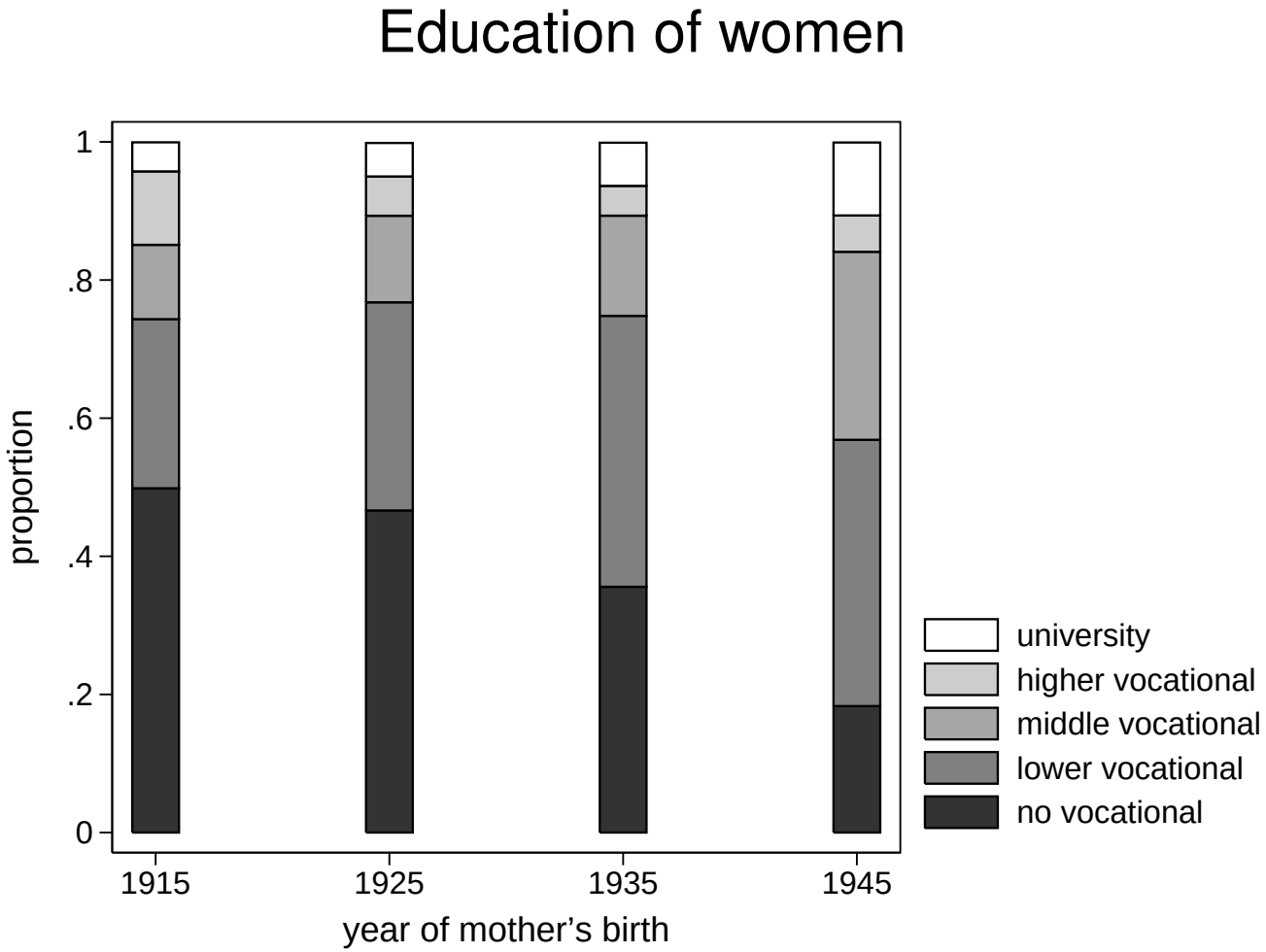
Empirical content

- The argument made by empirical research can be stylized as follows:
  1. You have a question (and theory and hypotheses)
  2. You look around (collect data)
  3. You summarize what you have seen (estimate statistical model)
  4. You answer question
- The information that makes the conclusion in step 4 credible is the information collected in step 2.
- That is often not the only information that is used: often it is necessary to add additional “information” in step 3 in the form of assumptions.
- We want to have an intuition how strong the information from step 2 is relative to the “information” from step 3.

Possible effects of educational expansion in the mother's generation

- People who were born more recently are more likely to attain higher levels of education

Possible effects of educational expansion in the mother's generation

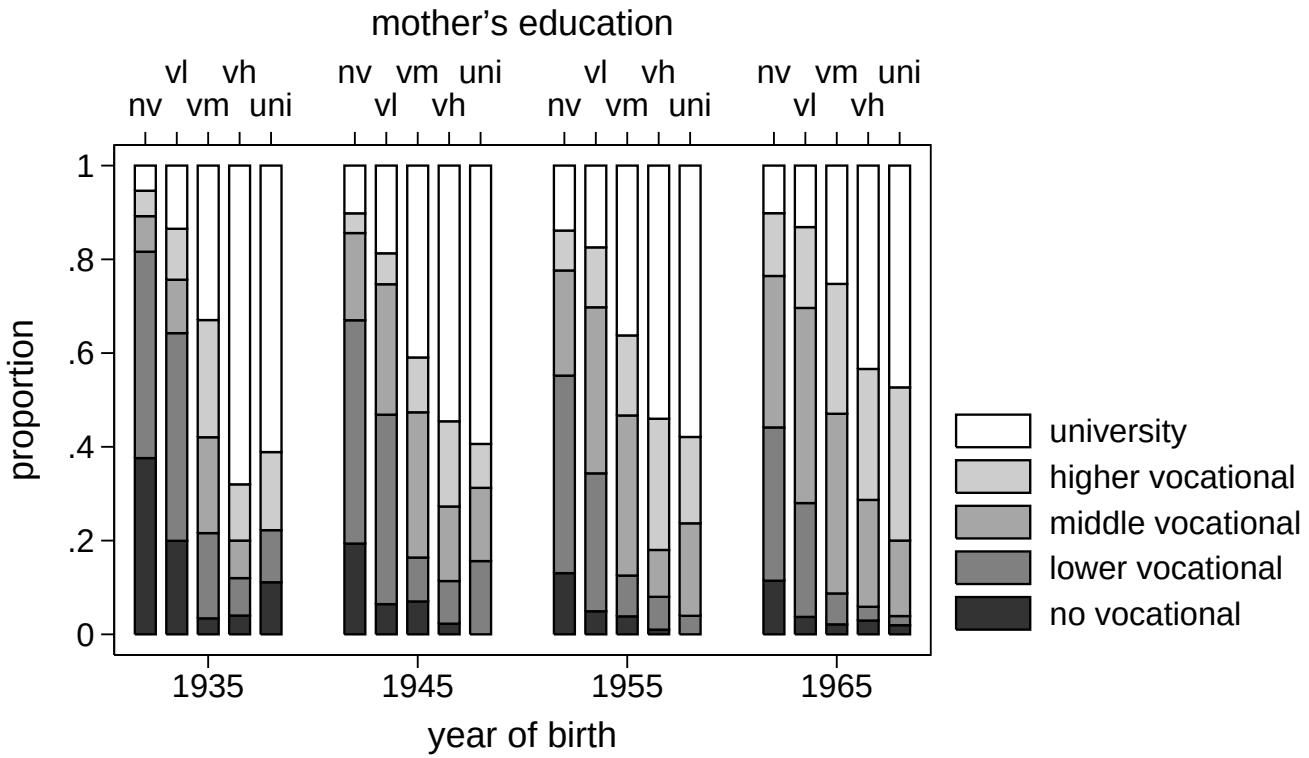


Possible effects of educational expansion in the mother's generation

- People who were born more recently are more likely to attain higher levels of education
- People who have higher educated parents are themselves more likely to attain higher levels of education

Possible effects of educational expansion in the mother's generation

Education of offspring conditional on education of mother



nv = no vocational,
vl = lower vocational, vm = middle vocational, vh = higher vocational,
uni = university

Possible effects of educational expansion in the mother's generation

- People who were born more recently are more likely to attain higher levels of education
- People who have higher educated parents are themselves more likely to attain higher levels of education
- Does educational expansion lead to further educational expansion in the next cohort?

Challenges

- Estimation requires a sample of mothers who have children who finished their education.
  - This is not a common design.
  - The most recent cohort that can be studied was born approx. 60 years ago.
- Alternative is to reconstruct such a dataset using simulation:
  - It is possible to use a wider range of datasources.
  - Exploration of effects in terms of counterfactual scenarios is also natural within this framework

Scenarios

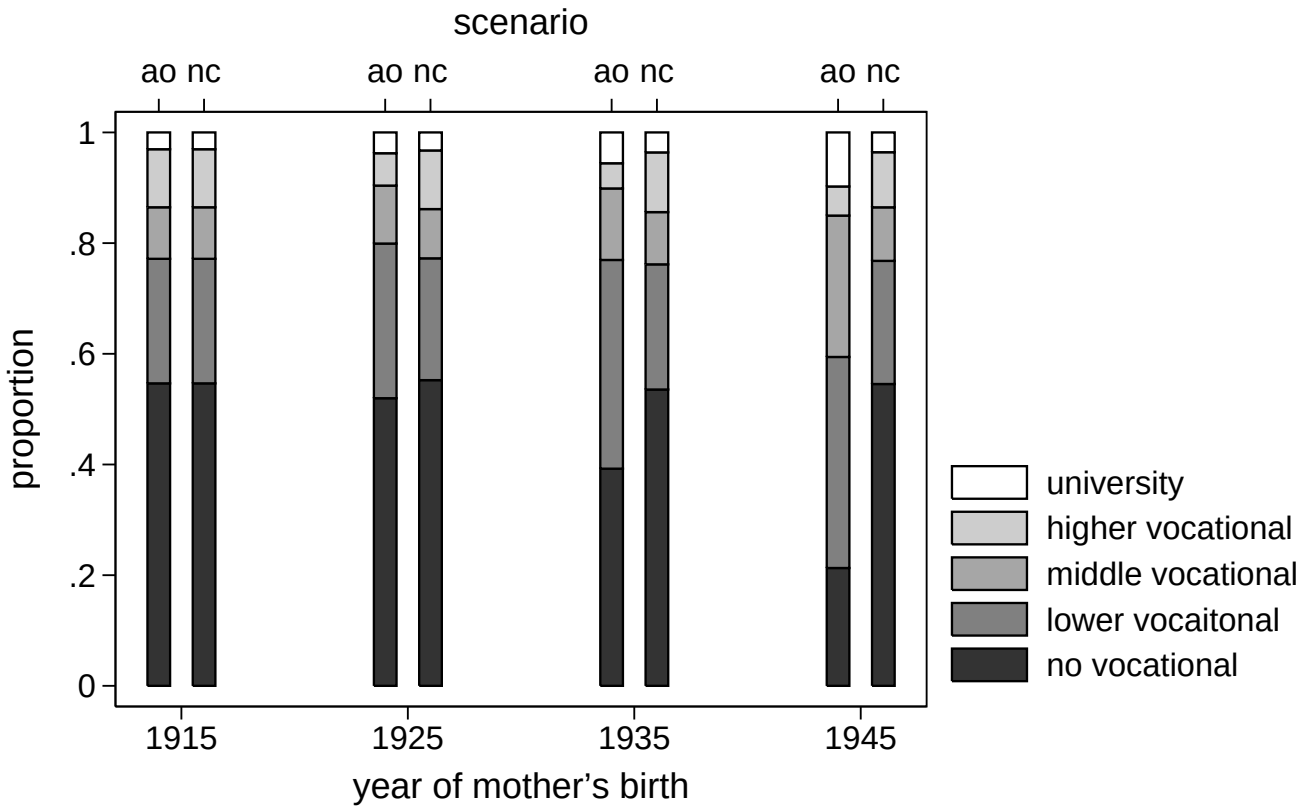
- In order to get an idea of what Educational expansion did we will look at the following scenario's
  - Educational expansion in the mother's generation took place as observed.
  - No educational expansion took place in the mother's generation.
- The outcome is the educational attainment of the offspring

Education

The following categorization for education will be used:

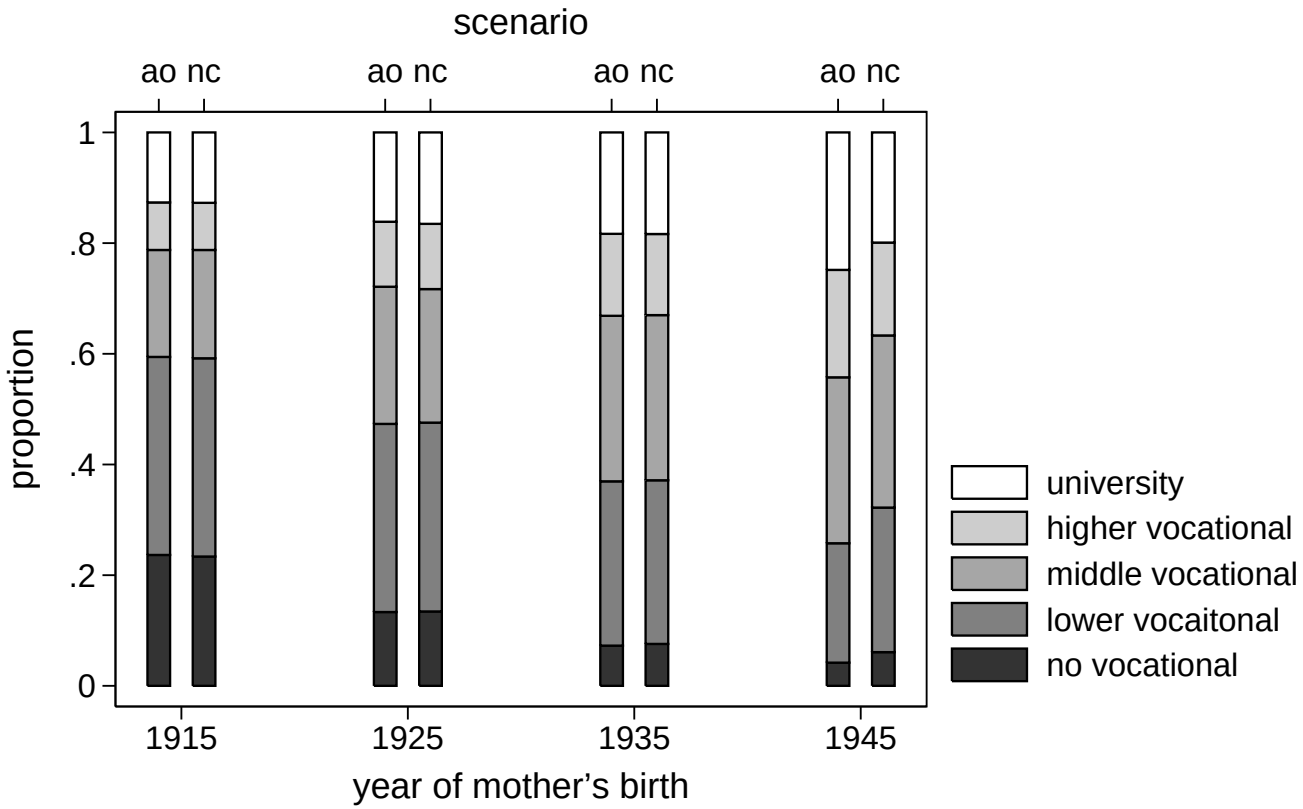
1. No vocational education
2. Lower vocational education (vocational education after Volksschule/Hauptschule/Grundschule)
3. Middle vocational education (vocational education after Realschule)
4. Higher vocational education (vocational education after Gymnasium)
5. University (Fachhochschule and Hochschule)

Scenarios



ao = as observed, nc = no change

The distribution of education of offspring in each scenario



ao = as observed, nc = no change

The big picture

1. start with an empty dataset
2. create 100,000 new ‘women’ for each cohort
3. given their cohort assign them an education
4. given their cohort and education assign them kids
5. given their cohort and education assign them a partner’s education
6. given the mother’s and father’s education, the number of siblings, and the year of birth assign the kids an education

steps 1/3

a. Start with data on women's education and year of birth
b. create a table of the proportion of women in each education category by cohort

cohort	pr1	pr2	pr3	pr4	pr5
1915	0.499	0.245	0.108	0.106	0.042
1925	0.467	0.302	0.125	0.057	0.049
1935	0.356	0.393	0.145	0.043	0.063
1945	0.183	0.386	0.272	0.053	0.106

c. save this table as a dataset and create 100,000 copies of each row (`expand`).
d. Give each 'women' at random an education based on the probabilities from the table.

Intermezzo: simulating from a multinomial distribution

- Say we have three educational categories with probabilities .3, .5, .2.
- We draw a number `u` from a standard uniform distribution
  - Give category 1 if `u` $<$ .3
  - Give category 2 if .3 $<$ `u` $<$ (.3 + .5)
  - Give category 3 if `u` $>$ (.3 + .5)

In Stata:

```
gen u = runiform()
gen ed = cond(u < .3, 1, ///
cond(u < .3 + .5, 2, 3))
```

Step 4: Assigning kids

a. Start with data on when women get their children, and the education and year of birth of the mother.
b. For each year between ages 15 and 40 estimate probabilities that women from a given cohort, age, and education get a child.
c. make a table of wherein for each cohort, age, education combination the predicted probability is given.
d. Go back to out simulated dataset
e. expand each observation 35 times and assign them ages 15 to 40
f. merge our table of birth probabilities in this dataset.
g. For each year randomly assign a women a birth based on these probabilities.
h. Keep only those mother-years where a birth occurred.

Step 5: Assigning a partner

a. Start with data on the education of mothers and fathers, year of birth of the mother, and whether there is a father present.
b. estimate the probability for each father's educational category (make "no father" an educational category) given mother's education and mother's cohort.
c. create a table of these probabilities for each combination of mother's education and cohort.
d. Go back to the simulated dataset
e. Merge the table of probilities into this dataset.
f. Randomly assign each mother a partner's education based on these probabilities.

Step 5: Assigning an education to the child

a. Start with data on the education of mothers, fathers and children, year of birth of the child, and the number of siblings.
b. estimate the probability for each child's educational category given the parents' education, cohort, and number of siblings.
c. create a table of these probabilities for each combination of mother's and father's education, number of siblings, and cohort.
d. Go back to the simulated dataset
e. Merge the table of probilities into this dataset.
f. Randomly assign each child an education based on these probabilities.

data

- Life History Study
- ALLBUS
- more to come...

Conclusion

- Educational expansion was mainly an effect of ‘real’ changes in the educational system.
- There has hardly been a feedback loop from educational expansion on expansion in the next generation.
- These results are based on a simulation because the necessary data is not sufficiently available.
- This basically consists of sequentially creating a dataset by randomly assigning characteristics to ‘individuals’ based on estimated conditional probabilities.