POTSDAM INSTITUTE FOR CLIMATE IMPACT RESEARCH

Generating Spatial Synthetic Populations

Distribution of carbon footprints and effects of carbon prices

Johannes Többen

GWS Osnabrück & Potsdam Institute for Climate Impact Research

toebben@gws-os.com

Leibniz Association

Agenda

1. Motivation
2. Research questions
3. How to make sense of the data?
4. What are synthetic populations?
5. Methodology
6. Outlook

Motivation

- Carbon-intensities and burdens of carbon prices unevenly distributed across household
- Mixed directions of effects found for many factors
- Many households only have discretionary control over of their total carbon footprint:
  - Tenants cannot change their heating system
  - Use of public transport depends on availability
  - Electricity mix driven mostly by national policy
- Most driving factors have a spatial dimension and depend on the local context

Summary of factors driving household's footprints studied so far*

	Factors	Direction of effect	Reasoning	Sources
Socio-economic	Income (INC)	+	Income directly determines household capacity to consume. The direction of the effect is more difficult to predict on product level, e.g. there exist inferior goods whose consumption goes down as income rises	Wilson *et al* 2013b, Tukker *et al* 2010, Peters and Hertwich 2008, Jackson and Papathanasopoulou (2008), Lenzen *et al* (2006)
	Household size (HHSIZE)	—	Household members share electrical appliances and require less individual living space Economies of scale in different consumption domains	Tukker *et al* (2010), Lenzen *et al* (2006), Wilson *et al* (2013b), Minx *et al* (2013)
	Urban-rural typology (URBAN)	+/−	Urban typology is associated with more compact development and larger availability of public transport, but studies have also found urban inhabitants to have higher impacts associated with food, leisure travel and manufactured products	Marcotullio *et al* (2014), Tukker *et al* (2010), Lenzen *et al* (2006), Minx *et al* (2013), Wiedenhofer *et al* (2013)
	Tertiary education (EDUC)	+/−	Education and social status redesign individual preferences towards more or less emission-intensive lifestyles	Chancel and Piketty (2015)
	Basic need spending (BASIC)	—	Spending on necessities (food, shelter, clothing) may be associated with lower emissions per unit of expenditure compared to that of transport and manufactured products	Ivanova *et al* (2015), Steen-Olsen *et al* (2016)
	Dwelling size (NROOMS)	+	Housing size determines the requirements of space heating/cooling and building material use	Lenzen *et al* (2006), Newton and Meyer (2012)
Geographic	Temperature (HDD)	+/−	Lower average temperatures (north) and low-quality, poorly isolated homes (south) are associated with higher emissions. Rising temperatures may also drive energy use for cooling.	Minx *et al* (2013), Wiedenhofer *et al* (2013), Chancel and Piketty (2015)
	Landscape (FORESTAREA)	+/−	Access to forest and semi-natural area may foster low-carbon leisure activities, but also encourage the consumption of available resources	Ivanova *et al* (2015)
Technical	Electricity mix intensity (EMIX)	+	The local electricity mix directly determines the carbon intensity of products produced and consumed locally (e.g. housing emissions)	Tukker *et al* (2010)

*Ivanova et al. (2017) Mapping the carbon footprint of EU regions. Environmental Research Letters, 12

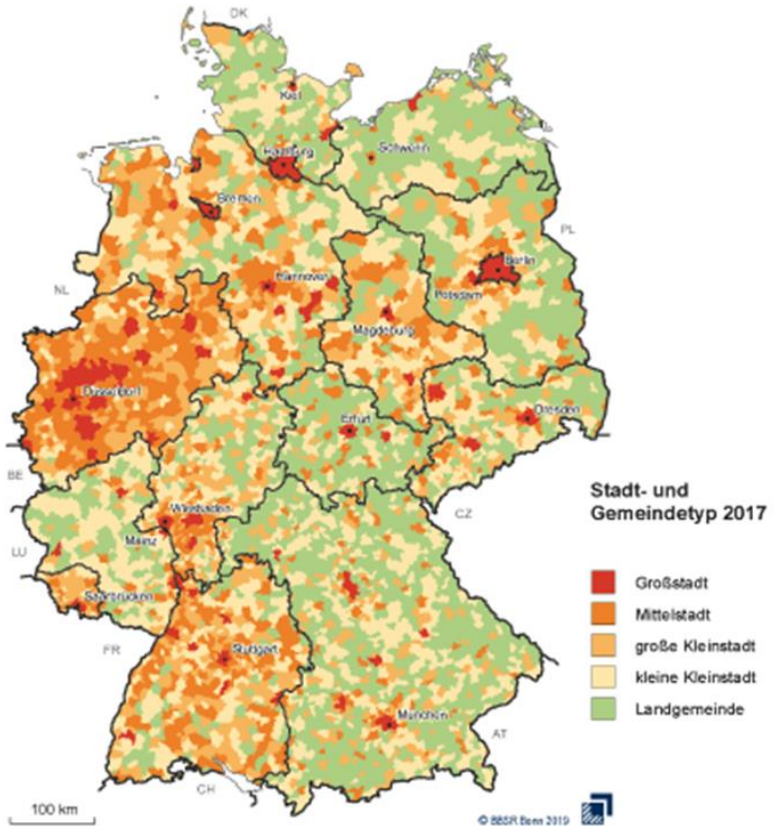
Research questions

Map income and consumption pattern of heterogenous households taking local context of their residence communities into account

Research Questions:

- Which household and location specific factors drive direct and indirect carbon emissions of households?
- How are costs and benefits of different carbon tax regimes distributed across households and communities (taking the ability to respond to price signals into account)?
- What would be an optimal tax regime reaching climate targets while minimizing regressive impacts?

11.300 German communities by size class



Main Data Sources I

Income and expenditure survey 2013 (EVS): Stratified sample of 43.000 households with 90.000 household members

- **Geographic:** Federal states, size class of communities
- **Demographics**: Age, gender, marital status, household size, family type
- **Socio-economics**: education, economic activity status, occupation, industry
- **Income and deductions from income**: Income from employment, public transfers, and different types of capital, taxes and social security
- **Expenditures:** COICOP3
- **Living conditions**: dwelling characteristics, household appliances and transportation equipment

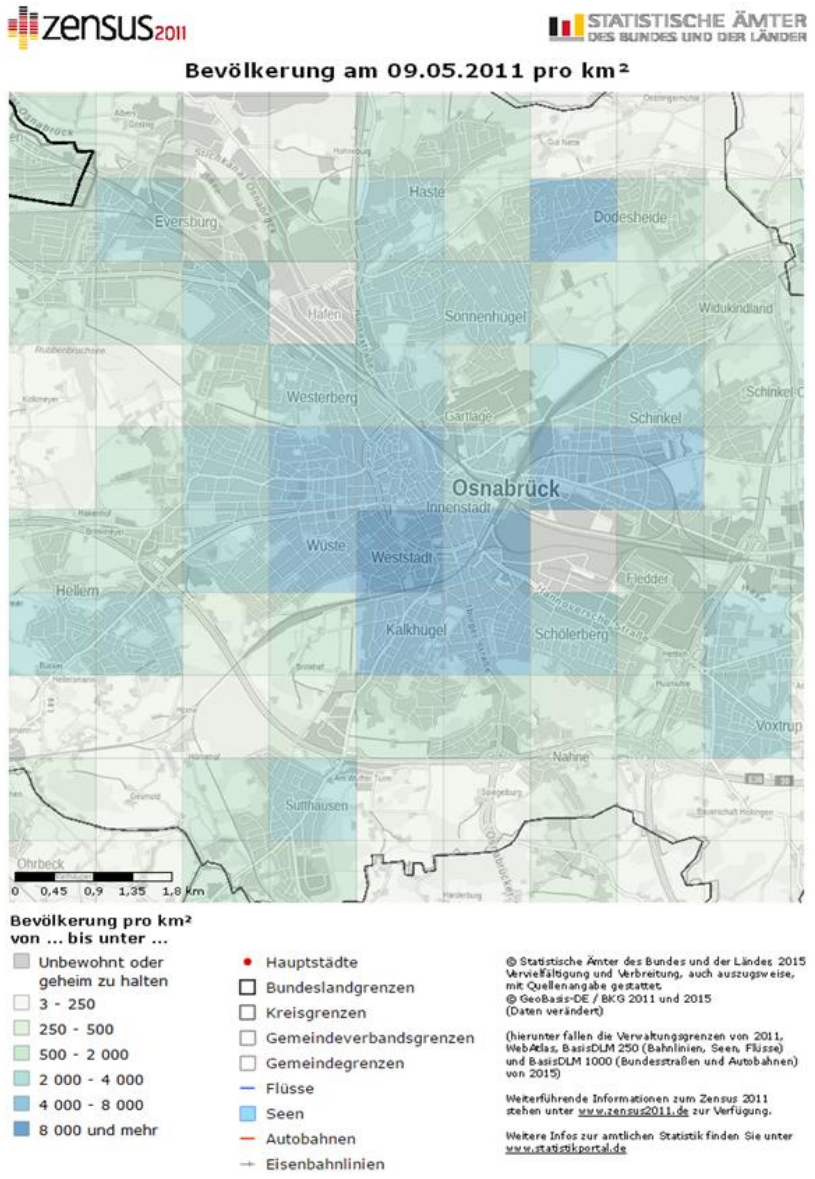
Main Data Sources II

Census 2011: Number of persons, households and dwellings by various categories

- **Persons:** age, gender, marital status, citizenship, economic activity status, socio-economic status, education, industry
- **Households:** size, family type
- **Dwelling:** type, size, age, heating system, owner

➢ Available for various hierarchical level of administrative regions (up to community level)

➢ Georeferenced data at 100m x 100m resolution



How to make sense of the data?

- We take the heterogeneity of households **and** the local context into account.
- How to combine data in a smart way?

Problem: Too many relevant household types to make useful groupings

- Relevant household variables could be:
  - Income (e.g. deciles)
  - Household size (e.g. 1,2,3,4,5+)
  - Household type (e.g. single, couple, couple with kids, lone parents, other)
  - Educational attainment (e.g. non, vocational, tertiary)
  - Economic activity status (e.g. employed, unemployed, inactive)
  - Dwelling size (e.g. 5 size classes)

  → 11.250 combinations

How to make sense of the data?

- If we follow the “typical” IO thinking, we would
  1. Form groups of households with similar characteristics
  2. Estimate shares of household types in a community based on census data
  3. Assign them “typical” consumption structures estimated from consumer surveys
  4. Disaggregate final demand vector of MRIO

➢ Easily yields more household type-community combinations than there are households in Germany

➢ **Creating a synthetic population is much more reasonable**

What are synthetic populations?

Synthetic population: Simulated individual-level data having the same statistical properties as observed in a sample & (hopefully) the actual population

H.ID	P.ID	Region	Person Variables				Household Variables			Income & deductions			Expenditures	
			Age	Eduction	...	Industry	H.Size	...	H.Owner	Labour	...	Pensions	Food	...
1	1	1	[15,30)	ISCED5	...	BtC	1	...	tenant	2470	...	-	230	...
2	2	1	[30,50)	ISCED7	...	BtC	3	...	owner	3101	...	-	342	...
2	3	1	[30,50)	ISCED3	...	GtH	3	...	owner	1251	...	-	342	...
2	4	1	[0,6)	-	...	-	3	...	owner	-	...	-	342	...
3	5	1	[50,65)	ISCED7	...	ItJ	1	...	owner	1894	...	-	177	...
4	6	1	[15,30)	ISCED4	...	F	1	...	tenant	1954	...	-	162	...
5	7	1	[65,inf)	ISCED5	...	-	2	...	tenant	-	...	1302	214	...
5	8	1	[65,inf)	ISCED5	...	-	2	...	tenant	-	...	890	214	...
6	9	1	[30,50)	ISCED4	...	F	2	...	tenant	1623	...	-	154	...
6	10	1	[30,50)	ISCED6	...	BtC	2	...	tenant	2743	...	-	154	...
...	...	...	...	...	...		...	...	...		...	-		...
39.326.225	80.219.694	11.339	[15,30)	ISCED4	...	A	2	...	tenant	2130	...	-	269	...
39.326.225	80.219.695	11.339	[30,50)	ISCED8	...	BtC	2	...	tenant	4510	...	-	471	...

Links between s

Synthetic population: Simul statistical properties as observed in a s

Final Demand:
- Carbon embodied indirectly in consumption expenditures computed by MRIO
- Applying carbon taxes rate translates this into increased costs of living

Caution:
- Price and quality differences of products purchased are difficult to account for

H.ID	P.ID	Region	P							ductions	Expenditures	
			Age	Ed	...			...		Pensions	Food	...
1	1	1	[15,30)	ISC	...			...		-	230	...
2	2	1	[30,50)	ISCED7	...	BtC	3	...	owner 3101	...	-	342	...
2	3	1	[30,50)	ISCED3	...	GtH	3	...	owner 1251	...	-	342	...
2	4	1	[0,6)	-	...	-	3	...	owner -	...	-	342	...
3	5	1	[50,65)	ISCED7	...	ItJ	1	...	owner 1894	...	-	177	...
4	6	1	[15,30)	ISCED4	...	F	1	...	tenant 1954	...	-	162	...
5	7	1	[65,inf)	ISCED5	...	-	2	...	tenant -	...	1302	214	...
5	8	1	[65,inf)	ISCED5	...	-	2	...	tenant -	...	890	214	...
6	9	1	[30,50)	ISCED4	...	F	2	...	tenant 1623	...	-	154	...
6	10	1	[30,50)	ISCED6	...	BtC	2	...	tenant 2743	...	-	154	...
...	...	...	...	...	...		...	...	...	...	-		...
39.326.225	80.219.694	11.339	[15,30)	ISCED4	...	A	2	...	tenant 2130	...	-	269	...
39.326.225	80.219.695	11.339	[30,50)	ISCED8	...	BtC	2	...	tenant 4510	...	-	471	...

Links between synthetic populations

Synthetic population: Simulated individual-level data having the same statistical proper[illegible] actual population

Income & deductions from income:

- Carbon Footprints: Relationship between income levels and carbon intensity of household's consumption
- Carbon tax regimes:
  - Relationship between cost of carbon and income tells us something about regressiveness
  - Income and deductions are angle to simulate impacts tax revenue recycling schemes

H.ID									Variables	Income & deductions			Expenditures	
									H.Owner	Labour	…	Pensions	Food	…
1									tenant	2470	…	-	230	…
2									owner	3101	…	-	342	…
2									owner	1251	…	-	342	…
2									owner	-	…	-	342	…
3									owner	1894	…	-	177	…
4	6	1	[15,30)	ISCED4	…	F	1	…	tenant	1954	…	-	162	…
5	7	1	[65,inf)	ISCED5	…	-	2	…	tenant	-	…	1302	214	…
5	8	1	[65,inf)	ISCED5	…	-	2	…	tenant	-	…	890	214	…
6	9	1	[30,50)	ISCED4	…	F	2	…	tenant	1623	…	-	154	…
6	10	1	[30,50)	ISCED6	…	BtC	2	…	tenant	2743	…	-	154	…
…	…	…	…	…	…		…	…	…		…	-		…
39.326.225	80.219.694	11.339	[15,30)	ISCED4	…	A	2	…	tenant	2130	…	-	269	…
39.326.225	80.219.695	11.339	[30,50)	ISCED8	…	BtC	2	…	tenant	4510	…	-	471	…

Link between synthetic populations

Possible Future Applications:

- Link persons/households to global value chains in their roles as

1. Consumers
2. Suppliers of primary inputs (especially labour)

→ Study e.g. uneven gains from trade

…vidual-level data having the same statistical …(hopefully) the actual population

							Household Variables			Income & deductions			Expenditures	
						Industry	H.Size	…	H.Owner	Labour	…	Pensions	Food	…
1	1	1	[15,30)	ISCED5	…	BtC	1	…	tenant	2470	…	-	230	…
2	2	1	[30,50)	ISCED7	…	BtC	3	…	owner	3101	…	-	342	…
2	3	1	[30,50)	ISCED3	…	GtH	3	…	owner	1251	…	-	342	…
2	4	1	[0,6)	-	…	-	3	…	owner	-	…	-	342	…
3	5	1	[50,65)	ISCED7	…	ItJ	1	…	owner	1894	…	-	177	…
4	6	1	[15,30)	ISCED4	…	F	1	…	tenant	1954	…	-	162	…
5	7	1	[65,inf)	ISCED5	…	-	2	…	tenant	-	…	1302	214	…
5	8	1	[65,inf)	ISCED5	…	-	2	…	tenant	-	…	890	214	…
6	9	1	[30,50)	ISCED4	…	F	2	…	tenant	1623	…	-	154	…
6	10	1	[30,50)	ISCED6	…	BtC	2	…	tenant	2743	…	-	154	…
…	…	…	…	…	…		…	…	…		…	-		…
39.326.225	80.219.694	11.339	[15,30)	ISCED4	…	A	2	…	tenant	2130	…	-	269	…
39.326.225	80.219.695	11.339	[30,50)	ISCED8	…	BtC	2	…	tenant	4510	…	-	471	…

What are synthetic populations?

Synthetic populations are typically created in two phases:

1. **Estimation:**
   a. Joint distribution of variables of interest is approximated using microdata
   b. Approximated joint distribution is adjusted to known marginal distributions of population characteristics.
2. **Sampling:**
   a. Individuals are randomly drawn from the joint distribution and added to the synthetic population
   b. Fit synthetic population to known marginals if necessary

What are synthetic populations?

Who creates synthetic populations and what are they good for?

a) **Statistical offices**:

- Anonymization of survey data to create Public Use Micro Samples (PUMS)
- Simulation studies to evaluate the behavior of statistical methods sampling designs

b) **Researchers**: Main data source for (spatial) microsimulation

- Analysis of individual-level phenomena over geographical space e.g. travel behavior for urban and transportation planning (e.g.
- Impacts of as tax policy (e.g. tax-benefit microsimulation model EUROMOD)

What are synthetic populations?

Who creates synthetic populations and what are they good for?

a) **Statistical offices**:

- Anonymization of survey data to create Public Use Micro Samples (PUMS)
- Simulation studies to evaluate the behavior of statistical methods sampling designs

b) **Researchers**: Main data source for (spatial) microsimulation

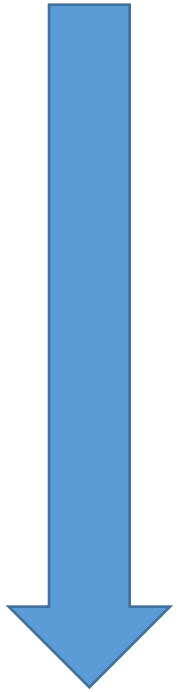
- Analysis of individual-level phenomena over geographical space e.g. travel behavior for urban and transportation planning
- Impacts of as tax policy (e.g. tax-benefit microsimulation model EUROMOD)

Methodology

- Stepwise procedure loosely following the Meindl et al. (2014) who created a Synthetic population of Austria based on EU SILC
- Implementation in R using the SimPop package commissioned by the International Household Survey Network (IHSN) and the World Bank Development Data Group

Stepwise Approach: At each step variables from the previous steps are used to predict variables to be added

1. **Estimate basic household and population structure in each community**
2. **Simulate additional demographic and socio-economic variables**
3. Simulate income & deductions from income
4. Simulate dwellings characteristics and possession of durable goods
5. Simulate consumption for non-durable goods and services



1. Estimate basic household and population structure in each community

Example of a 3-D contingency table

male

Household size	Household type: single	couple w/o kids	couple with kids	other	Total
1	10	0	0	0	10
2	0	7	0	6	13
3	0	0	7	5	12
4	0	0	5	2	7
5+	0	0	2	0	2
Total	10	7	14	13	44

female

Household size	Household type: single	couple w/o kids	couple with kids	other	Total
1	5	0	0	0	5
2	0	6	0	3	9
3	0	0	5	2	7
4	0	0	4	1	5
5+	0	0	2	1	3
Total	5	6	11	7	29

Total

Household size	Household type: single	couple w/o kids	couple with kids	other	Total
1	15	0	0	0	15
2	0	13	0	9	22
3	0	0	12	7	19
4	0	0	9	3	12
5+	0	0	4	1	5
Total	15	13	25	20	73

Marginal frequencies

- At community/county level from Census 2011
- Used as constraints on the synthetic population
- Example: Number of tenant vs. owner households

$$f(x = tenant) = 44$$

Conditional frequencies

- Available for some person attributes from Census 2011, especially age, marital and economic activity status by gender
- Used as constraints on synthetic population
- Example: Number of tenant households among couples with kids

$$f(x = tenant | y = couple\ with\ kids) = 14$$

Joint frequencies

- Information available from the microdata but not for specific communities
- Example: Number of couples with 2 kids (i.e. size = 4) who are tenants

$$f(x = tenant, y = couple\ with\ kids, z = size\ 4) = 4$$

- **Objective:** Estimate community specific joint frequencies by those observed in the survey to meet marginals/conditionals from Census

1. Estimate basic household and population structure in each community

Iterative Proportional Fitting (IPF):

- Standard approach in Synthetic Population Literature and equivalent to RAS/minimizing cross-entropy
- Main idea: Iteratively adjust weights of households in the sample such that marginal/conditional frequencies at community level are satisfied
- Problems:
  - Too many variables of interest and most of them not categorial (i.e., income, expenditures)
  - Just replicates households observed in sample (incl. empirical zeros)
  - Curse of dimensionality: Number of possible combinations of person, household and dwelling characteristics in Census data much larger than number households observed in the sample

➢IPF is only used for setting up the basic household and population structure

1. Estimate basic household and population structure in each community

Estimation

- Household characteristics are assigned to persons living in the respective household
- Person weights equal household weights divided by the number of household member
- Households with identical characteristics are added up

H.ID	P.ID	H.weight	P.Age	P.Gender	P.Marital	H.Size	H.Type	weight(0)
1	1	23.64	[15,30)	male	single	1	single	23.64
2	2	43.23	[30,50)	female	married	3	couple with kids	14.41
2	3	43.23	[30,50)	male	married	3	couple with kids	14.41
2	4	43.23	[0,6)	female	single	3	couple with kids	14.41
3	5	37.81	[50,65)	female	widowed	1	single	37.81
4	6	15.18	[15,30)	male	divorced	1	single	15.18
5	7	6.67	[65,inf)	male	married	2	couple without kids	3.335
5	8	6.67	[65,inf)	female	married	2	couple without kids	3.335
6	9	37.75	[30,50)	male	single	2	other	18.875
6	10	37.75	[30,50)	male	single	2	other	18.875
…	…	…	…	…	…	…	…	…

$$\min D = \sum_i weight_i(0) \ln \frac{weight_i(1)}{weight_i(0)}, \text{ s.t.}$$

$$\sum_i weight_i(1)\ P.Age_{ij} = N.Age_{rj}$$

$$\sum_i weight_i(1)\ P.Gender_{ij} = N.Gender_{rj}$$

$$\sum_i weight_i(1)\ H.Size_{ij} = N.Size_{rj}$$

$$\sum_i weight_i(1)\ H.Type_{ij} = N.Type_{rj}$$

where i denotes persons and j denotes value of respective variable

2. Simulate additional demographic and socio-economic variables

Objective: simulate categorial variables allowing for combinations that do not (or rarely) occur in the sample but are likely to occur in the true population

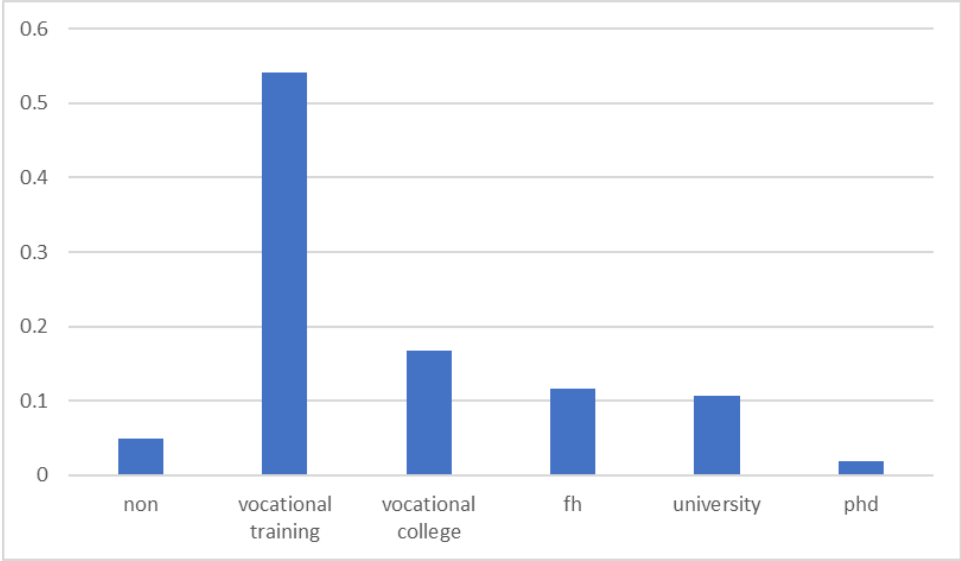
Variable	#Categories	Marginals?	Values
Citizenship	3	communities	German, EU27, other
Economic Activity Status	3	counties/ some communities	Employed, unemployed, not economically active
Educational attainment	6	counties/ some communities	Non, vocational training, vocational college, FH, university, PhD
Socio-economic status	3	counties/ some communities	Dependent employee, self-employed, official
Industry (NACE rev. 2)	10	counties/ some communities	A, BtC, DtE, F, GtH, ItJ, K, LtN, O, PtU

2. Simulate additional demographic and socio-economic variables

General Approach:

1. Estimate conditional distributions from sample using multinomial logistic regression models
2. Apply coefficient estimates from sample to variables already available in the synthetic population
   a. 1. variable → basic household variables used as covariates,
      i.e. $p(y_{i1}|x_{i1}, \dots, x_{iM})$
   b. 2. variable → basic household variables and 1. variable used as covariates,
      i.e. $p(y_{i2}|y_{i1}, x_{i1}, \dots, x_{iM})$
   c. N[th] variable → basic household variables and N-1 variable used as covariates
      i.e., $p(y_{iN}|y_{i1}, \dots, y_{iN-1}, x_{i1}, \dots, x_{iM})$
3. Adjust simulated variables to marginals available at county and community level

Conditional probability of educational attainment of a married male between 30 and 50 living with his wife and one kid



2. Simulate additional demographic and socio-economic variables

Problem: Using variables as covariates estimated in the previous step means that the order of the additional categorical variables may be relevant

How to overcome the problem that the order of estimation matter?

- Once all additional variables are available in the synthetic population continue iteratively with estimating conditional from sample population and predicting variables in the synthetic one
  - *a.* y_{i1} is predicted by $y_{i2}, \dots, y_{iN}$ and $x_{i1}, \dots, x_{iM}$
  - *b.* y_{i2} is predicted by $y_{i1}, y_{i3}, \dots, y_{iN}$ and $x_{i1}, \dots, x_{iM}$
  - c. …
- Markov Chain Monte Carlo: Sampling from an unknown, highly dimensional joint distribution is approximated by interactively drawing samples from the conditional distributions*

*Hastings, W. K. (1970): *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*. *Biometrika*, 57

Outlook

- Continue general approach of
    1. Estimating relationships between variables already available in synthetic population and variable to be added from EVS sample
    2. Predict new variable in synthetic population based on estimates from step 1
    3. Adjust estimates to known marginals
- Specific models necessary to simulate remaining variables
    - Income: Sampling from a Generalized Pareto Distribution conditioned on demographic and socio-economic characteristics
    - Dwelling and durables: mix of linear regression
    - Consumption of non-durables and services: Demand System

No Conclusion, yet…

Critical Feedback and Ideas for further applications highly appreciated!

POTSDAM INSTITUTE FOR CLIMATE IMPACT RESEARCH

Thanks for listening!

Johannes Többen

GWS Osnabrück & Potsdam Institute for Climate Impact Research

toebben@gws-os.com