Institute for Employment Research

The Research Institute of the Federal Employment Agency

IAB

Generating synthetic geocoding information for public release

06. October 2015,
Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Helsinki

Jörg Drechsler
(Institute for Employment Research)

&

Monika Jingchen Hu
(Vassar College)

Background

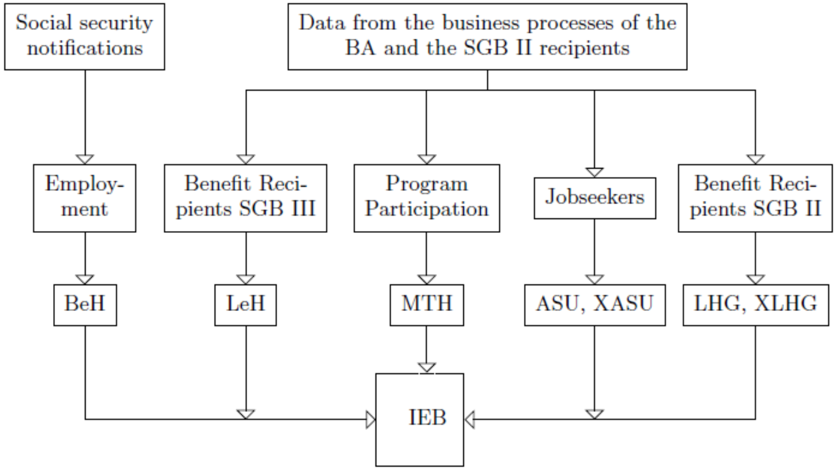
- more and more agencies collect detailed geocoding information
- information can be useful for various purposes
  - to allow for detailed analyses on a user defined geographical level
  - to link information from other sources
- sharing of detailed geocoding information problematic
- geocodes not necessarily sensitive information
- but detailed geographical information increases the risk of re-identification
- usually very limited access to detailed geocodes for external researcher

Synthetic data for statistical disclosure control

- proposed by Rubin (1993) and Little (1993)
- especially useful if high level of protection is required
- idea is closely related to multiple imputation for nonresponse
- generate synthetic datasets by drawing from a model fitted to the original data
- not the missing values but the sensitive values are replaced with a set of plausible values given the original data
- if models are carefully selected, important relationships found in the original data are preserved

Application – The Georeferenced IEB

- Integrated Employment Biographies (IEB): large database constructed from different administrative data sources of the German Federal Employment Agency


Social security notifications
Data from the business processes of the BA and the SGB II recipients
Employment
Benefit Recipients SGB III
Program Participation
Jobseekers
Benefit Recipients SGB II
BeH
LeH
MTH
ASU, XASU
LHG, XLHG
IEB

- detailed geocoding information has been added recently
- data should be disseminated to the scientific community if possible

Application – The Georeferenced IEB

- goal: evaluate whether a useful synthetic dataset could be generated for a small set of variables from the IEB

variable	characteristics
exact geocoding info	recorded as distance in meters from the point 52 northern latitude (Y), 10 eastern longitude (X)
sex	male/female
foreign	yes/no
age	6 categories
education	6 categories
occupation level	7 categories
occupation	12 categories
industry of the employer	15 categories
wage	10 categories defined by quantiles
distance to work	5 categories
ZIP code	2,063 ZIP code levels (not used as predictor)

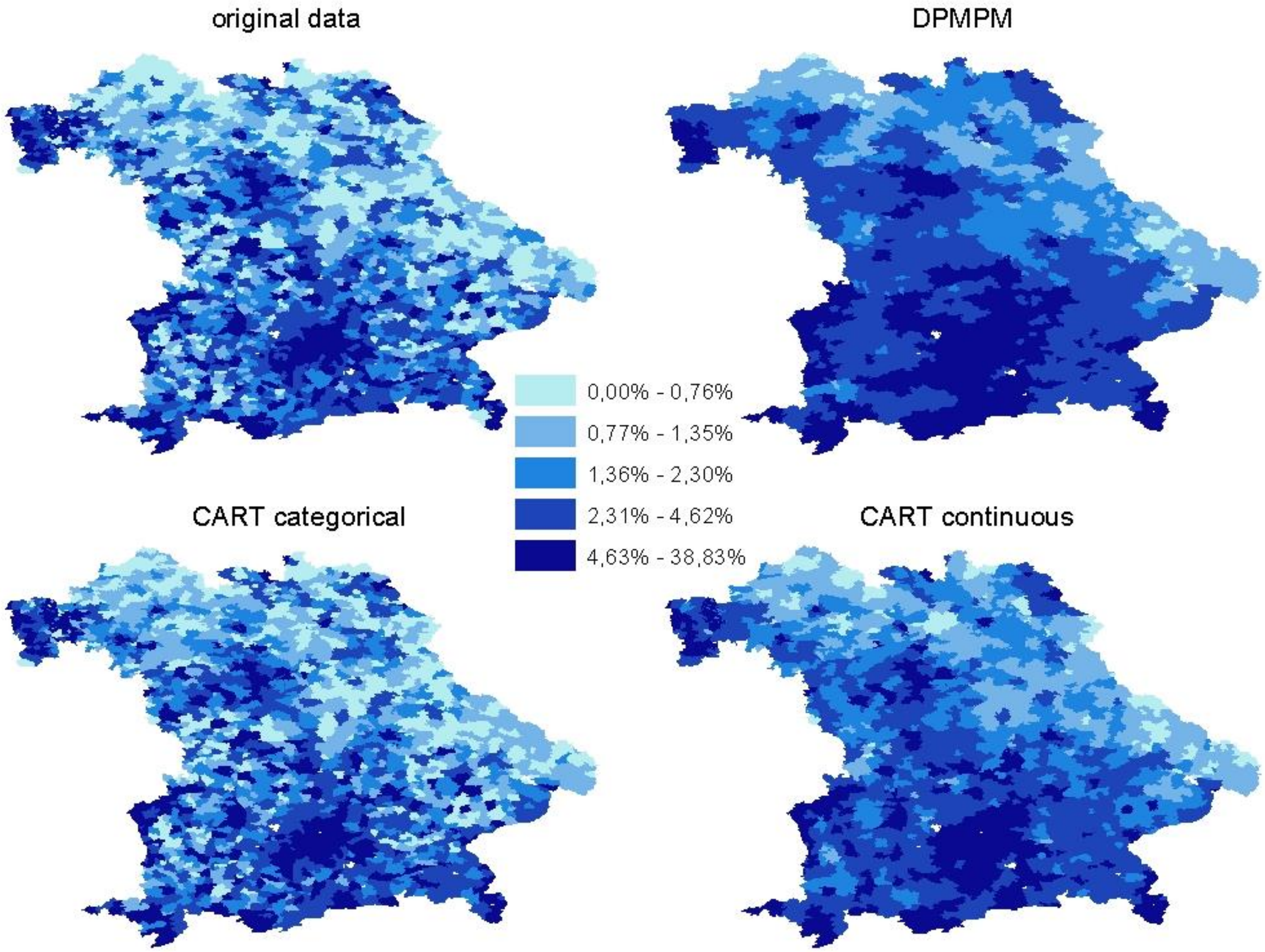
- dataset limited to fully observed records in Bavaria (~ 4 Mio records)

Application – The Georeferenced IEB

- we only synthesize the geocoding information
- 3 different synthesis models
- Dirichlet process mixture of products of multinomials (DPMPM)
  - Bayesian version of latent class model for unordered categorical data
- CART models
- two versions of the CART models
  - treat geocodes as categorical
  - treat geocodes as continuous
- data divided into clusters of size 15,000 for computational reasons
- m=5

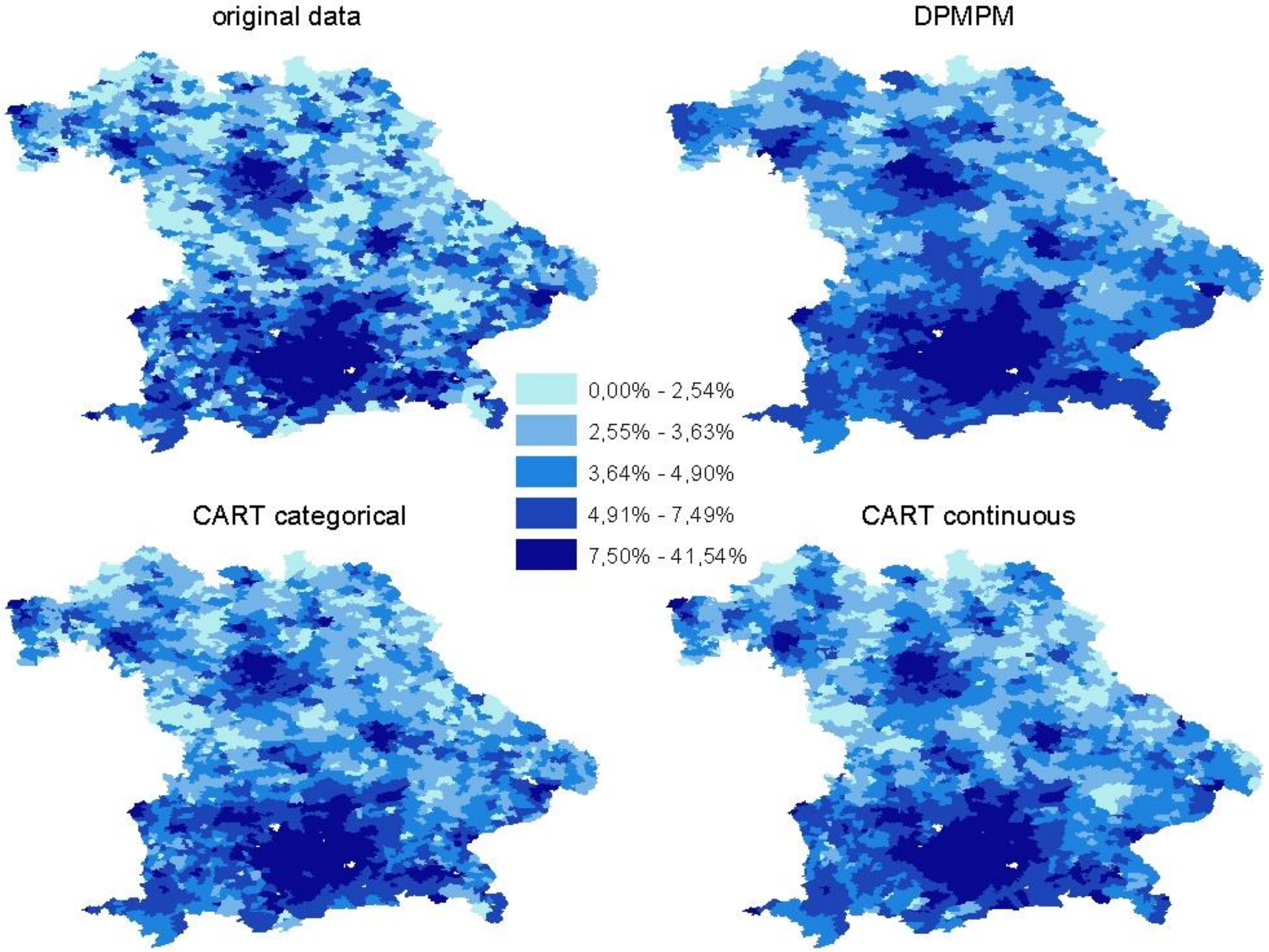
Utility evaluations – specific measures

Share of foreigners in Bavaria by ZIP code level



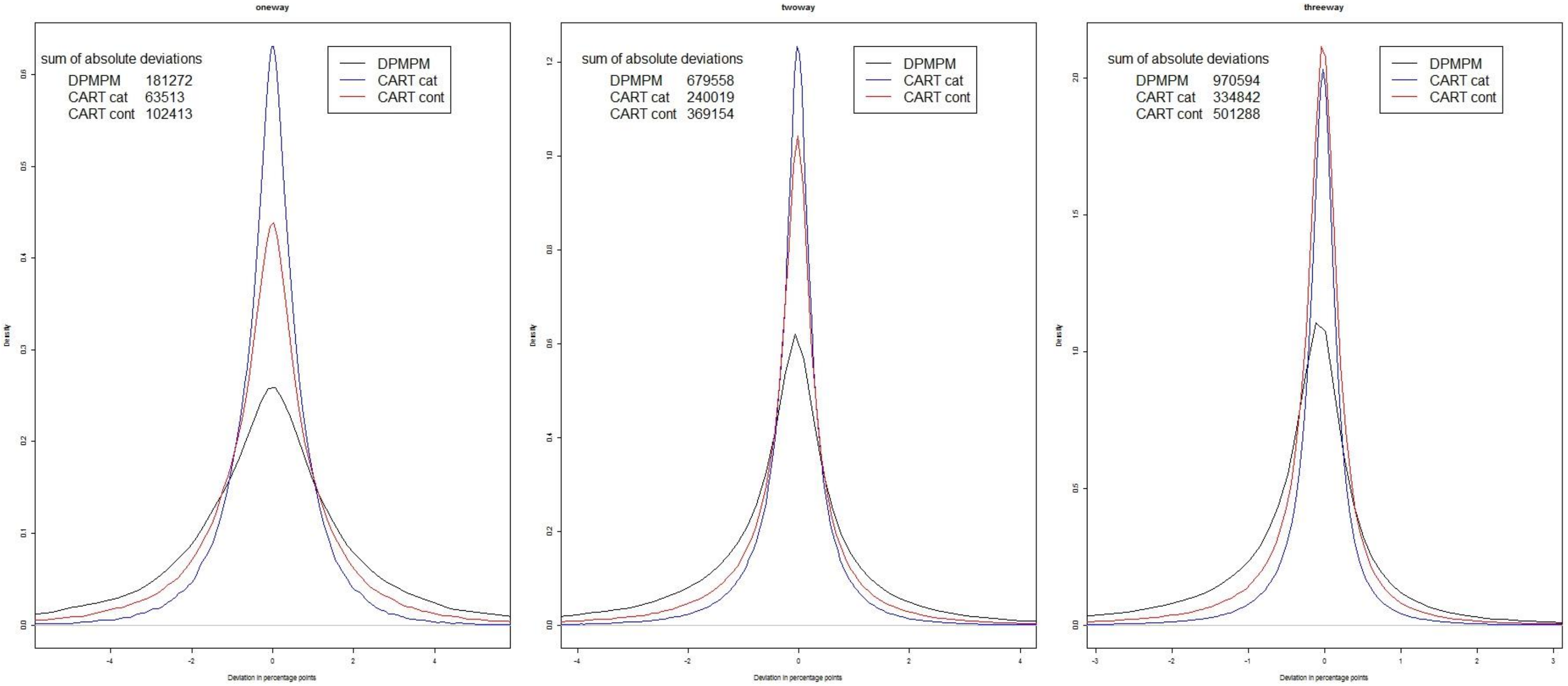
Utility evaluations – specific measures

Share of females with university degree or similar among all employed females by ZIP code level



Utility evaluations – global measure

- compute relative frequencies of all possible interactions (up to threeway) of all variables on the ZIP code level
- compute absolute distance of these frequencies between the original data and the synthetic data



Disclosure risk evaluations

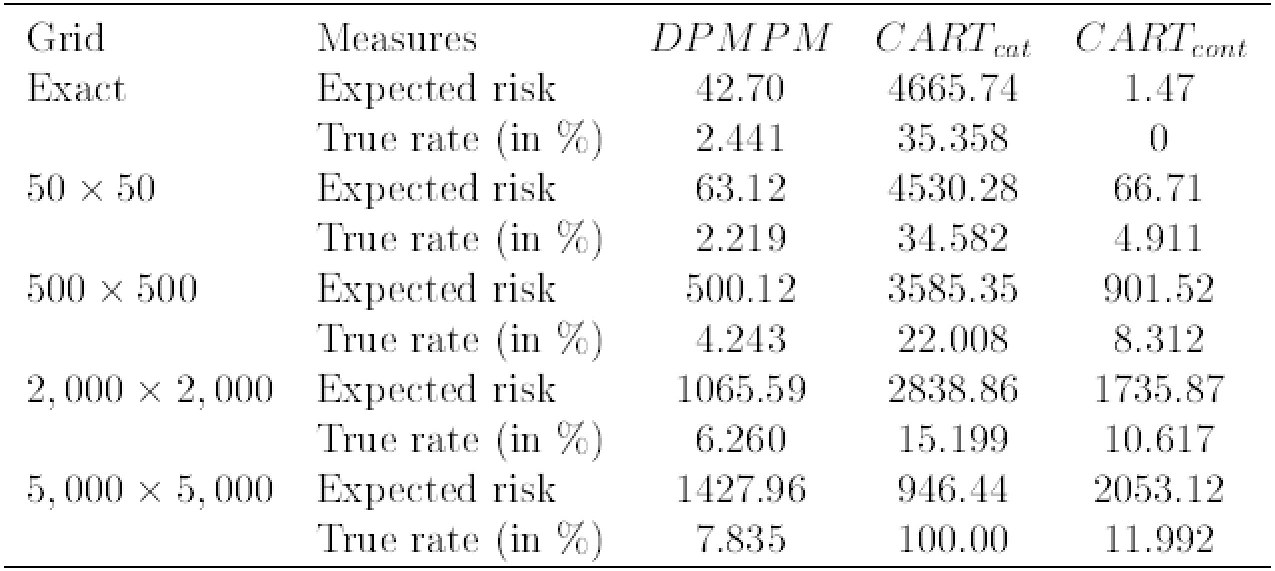
- use risk measures suggested by Reiter and Mitra (2009)
- assume that intruder has some background knowledge on some target variables
- tries to find these targets in the released data to learn sensitive information
- intruder computes matching probabilities for each record in the released file
- declares the record with the highest matching probability to be the match
- risk measures evaluate how often this strategy is successful

Application of disclosure risk measures

- we assume intruder knows sex, age, industry, occupation, foreign (y/n), geocode
- target records: sample of 100 records from each cluster
- total number of target records 22,200
- intruder matches on all variables and uses various grids for the geocode

exp. risk org data:
21971.14

exp. risk w/o geocodes:
1821.16



Grid	Measures	$DPMPM$	$CART_{cat}$	$CART_{cont}$
Exact	Expected risk	42.70	4665.74	1.47
	True rate (in %)	2.441	35.358	0
50×50	Expected risk	63.12	4530.28	66.71
	True rate (in %)	2.219	34.582	4.911
500×500	Expected risk	500.12	3585.35	901.52
	True rate (in %)	4.243	22.008	8.312
$2,000 \times 2,000$	Expected risk	1065.59	2838.86	1735.87
	True rate (in %)	6.260	15.199	10.617
$5,000 \times 5,000$	Expected risk	1427.96	946.44	2053.12
	True rate (in %)	7.835	100.00	11.992

Conclusions and Outlook

- analytical validity
  - all methods smooth the geospatial effects
  - DPMPM shows very low analytical validity
  - CART categorical performs best
  - CART continuous can generate unreasonable geocodes
- disclosure risk
  - risks very high for CART categorical
  - DPMPM shows lowest risks
- future plans
  - synthesize other variables
  - tune CART synthesizers

Thank you for your attention

joerg.drechsler@iab.de