INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

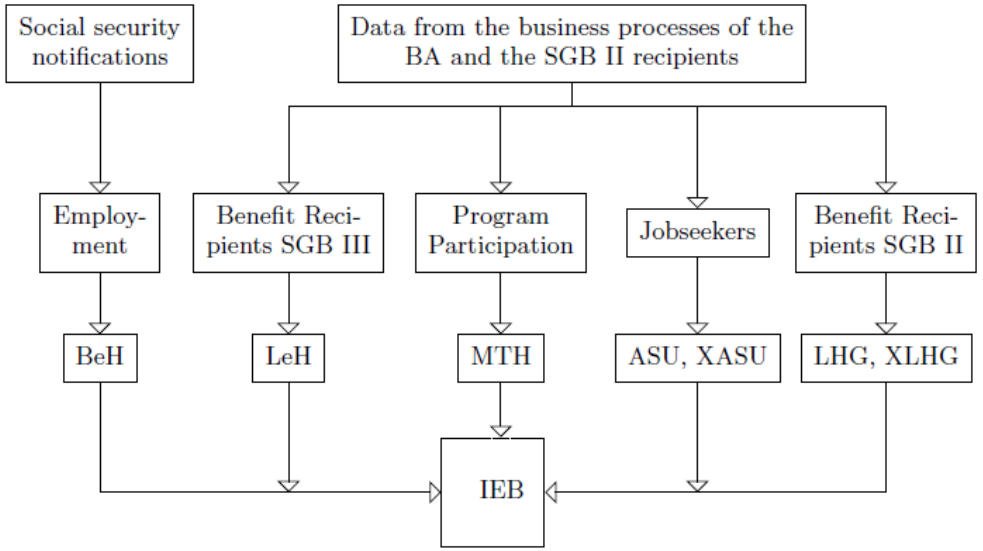
DIFFERENTIAL PRIVACY FOR GEODATA

Jörg Drechsler (IAB, University of Maryland)
Jordi Soria-Comas (Univerisity Rovira i Virgili)

World Statistical Congress
July 16, 2021

MOTIVATION – THE GEOREFERENCED IEB

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



- Recent addition of detailed geographical information motivated this entire research

DIFFERENTIAL PRIVACY

- Provides formal privacy guarantees
- Motivation: bound the risk, if an individual decides to become part of the database
- Differential privacy ensures that change of a single row in the database has (almost) no impact on the reported results
- Implies that individuals can participate without risk
- Privacy is ensured through randomization

A randomized function gives ε-differential privacy if and only if for all datasets D_1 and D_2 differing on at most one element, and for all $S \subset Range(\kappa)$,

$$P(\kappa(D_1) \in S) \leq \exp(\varepsilon) P(\kappa(D_2) \in S)$$

APPLICATION

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

Variables included in the dataset used for the evaluations	
Exact gecoding information	recorded as distance in meters from the point 52° northern latitude, 10° eastern longitude
Sex	male/female
Foreign	yes/no
Employed	yes/no
Unemployment benefits	yes/no
Skills	low/medium/high
Wage	low/medium/high
Distance to work	5 categories (≤1, 1-5, 5-10, 10-20, >20 km)

THE GEOMETRIC MECHANISM

- Simple mechanism to ensure differential privacy for frequency tables
- Adds random draws from two-sided geometric distribution to each cell count
- Descrete version of the Laplace mechanism
- Can be useful for generating differentially private microdata
  - All variables need to be treated as categorical
  - Full cross-classification of all variables
  - Random draw from two-sided geometric distribution is added to each cross-classification cell
  - Noisy data are turned back into microdata to be released
- Utility for aggregated statistics could be improved by applying the algorithm at different geographical levels and enforcing consistency (TopDown approach of the U.S. Census Bureau)
- Not implemented yet

OBTAINING VALID INFERENCES

- Computed means and totals based on noisy data will be unbiased
- But we need to quantify the uncertainty from noise infusion
- Variance estimation straightforward for totals
- Let $Var(U) = \frac{2e^{-\varepsilon}}{(1-e^{-\varepsilon})^2}$, be the variance of the noise term
- For any total $\tilde{t}$ computed on the noisy data, the variance is given as

$$Var(\tilde{t}) = m \cdot Var(U),$$

with *m* number of cells in the cross-classified table, over which we need to aggregate to obtain $\tilde{t}$.

ESTIMATING THE VARIANCE OF A MEAN

- Mean is always a ratio of two random variables (even size of the database is random)
- Variance can no longer be obtained in closed form
- Can use first order Taylor series expansion
- Let $\tilde{\tilde{X}}$ be the noisy estimate for the mean of interest
- First order Taylor series expansion around the mean leads to

$$Var(\tilde{\tilde{X}}) = Var\left(\frac{\hat{t}_x}{\hat{t}_y}\right) \approx \frac{1}{\hat{t}_y^2}\left(Var(\hat{t}_x) + \frac{\hat{t}_x^2}{\hat{t}_y^2}Var(\hat{t}_y) - 2\frac{\hat{t}_x}{\hat{t}_y}d \cdot Var(U)\right),$$

with d number of cells in the fully cross-classified table that contribute to both t_x and t_y

ILLUSTRATIVE SIMULATION

- Using only 4 variables and a small geographical subset (40 cells in the cross-classified table)

	# not employed	# not employed foreigners	% not employed	% not emp. among foreigners
True value	8023	2258	18.81	40.09
Average est.	8023.17	2258.12	18.81	40.09
Var. ratio	1.02	1.01	1.02	0.99
CI coverage (%)	94.5	94.7	95.34	94.78
avrg. CI length	23.79	16.82	0.05%	0.22%

DEALING WITH THE GEOCODES

- Releasing detailed geocodes not realistic
- We consider the geocoding information to be prior knowledge
- Geographical information itself is not sensitive
- Only relationship to the other variables might be sensitive
  - We can use the actual geocodes to partition the geocoding domain
  - Advantage over aggregation using fixed grids: improved accuracy because small cell counts are avoided
- We use a micro aggregation algorithm for the partitioning
- We only cluster records with same zipcode to improve accuracy
- Cluster size is a tuning parameter to choose between geographical detail and accuracy of results

DEALING WITH NEGATIVE CELL COUNTS

- Fully cross-classified table will be sparse
- Noisy table will contain negative values
- Especially problematic if microdata should be released based on the noisy table
- Possible solutions discussed in the literature
  - Set all negative cell counts to zero
  - Use Fourier transformation and add small amount to the first coefficient (Lp mechanism)
  - Use linear programming to find solution that is as close as possible to the noisy table subject to a non-negativity constraint
- Disadvantages of the proposed solutions
  - DP estimates no longer unbiased
  - Unclear how to obtain valid variance estimates

POSSIBLE ALTERNATIVE: CALIBRATION

- We cannot use information from the original data without spending some privacy budget
- We can use unbiased counts from the noisy table instead
- Proposed solution consists of the following steps
  - Generate dp table $\widetilde{T}$ from the original data
  - Set all negative cell counts in $\widetilde{T}$ to zero
  - Generate dp microdata based on adjusted table
  - Use $\widetilde{T}$ to generate calibration weights to be released together with the microdata
- Calibration step should use all cells from $\widetilde{T}$ as benchmarks

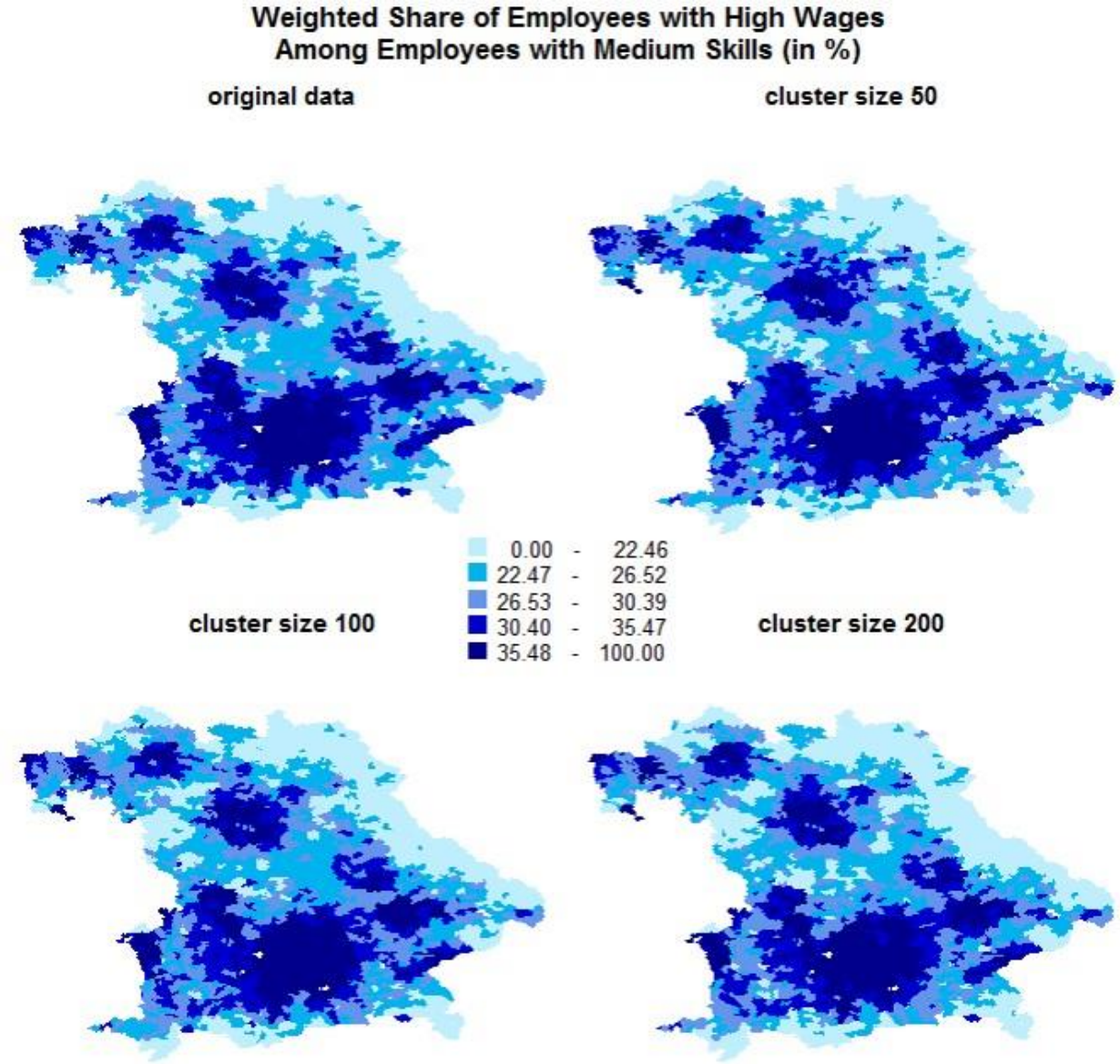
SIMULATION RESULTS

- Same simulation using all variables (160,364 cells in the table, *N*=42,665)
- All negative noisy counts are set to zero

	# not employed	# not employed foreigners	% not employed	% not emp. among foreigners
True value	8023	2258	18.8	40.09
Average est.	10778.41	3999.96	7.66	6.75
Var.ratio	1.8	2.05	1.89	2.11
CI coverage (%)	0	0	0	0
Avrg. CI length	439.41	310.71	0.31%	0.52%

SIMULATION RESULTS

	Calibration approach			
	# not employed	# not employed foreigners	% not employed	% not emp. among foreigners
True value	8023	2258	18.8	40.09
Average est.	8023.74	2258.13	18.81	40.28
Var.ratio	0.97	0.99	0.97	1.01
CI coverage (%)	94.36	94.66	94.78	95.14
Avrg. CI length	439.41	310.71	1.24%	11.14%


Weighted Share of Employees with High Wages
Among Employees with Medium Skills (in %)
original data
cluster size 50
cluster size 100
cluster size 200
0.00 - 22.46
22.47 - 26.52
26.53 - 30.39
30.40 - 35.47
35.48 - 100.00

CONCLUSIONS

- Geometric mechanism as a convenient tool for DP microdata release
- Seems to produce acceptable results for this application
- Only useful if interest lies on detailed geographical results
- Approaches akin to TopDown algorithm needed otherwise
- More difficult to obtain valid inferences

CONTACT

Jörg Drechsler

joerg.drechsler@iab.de

ILLUSTRATIVE SIMULATION STUDY

- Assume that only information on sex, foreign status, employment status, and Zip code are released
- User is the major of the city of Fürth and he is interested in
  - the total number of unemployed
  - the total number of unemployed foreigners
  - the unemployment rate
  - the unemployment rate among foreigners
- Will he get valid results based on the protected data?
- Repeated simulation design
- dp dataset is generated and analyzed 5,000 times
- We set $\varepsilon=1$

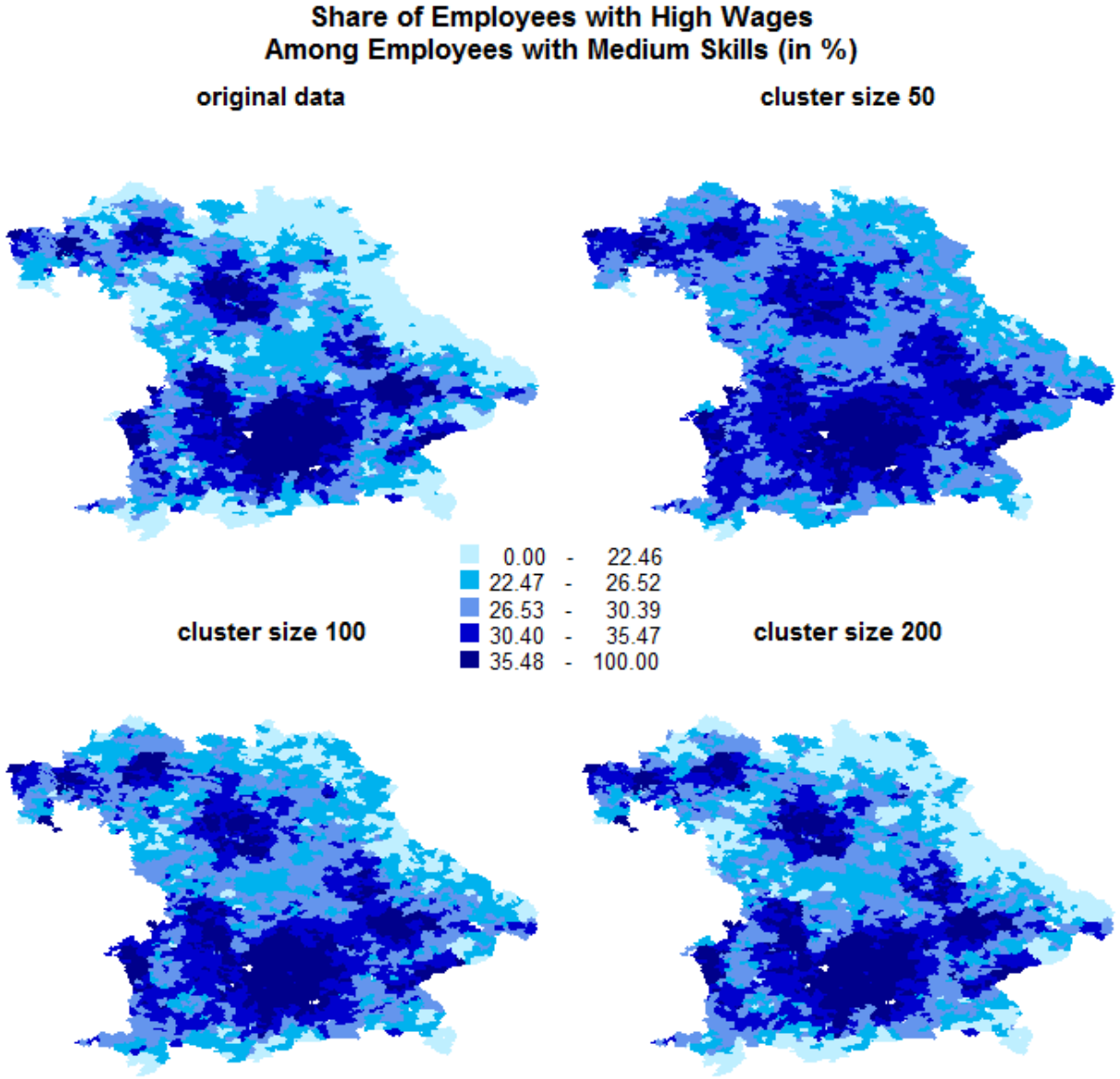
SIMULATION STUDY FOR THE CITY OF FÜRTH

- Same simulation design, but using all variables from the previous table
- Exact geocodes cannot be used
- We use a differentially private microaggregation approach to built clusters of size 50
- Size of the database N=42,665
- Fully cross-classified table contains 160,364 cells
- All negative cell counts in the noisy table are set to zero

DISCUSSION OF THE CALIBRATION APPROACH

- Advantages
  - Estimates from the dp microdata are unbiased
  - Calibrating to all cells in $\widetilde{T}$ ensures that calibration has no effect on uncertainty
  - Variance estimates are still valid
- Disadvantages
  - Totals and means at low level of aggregation will still be negative
  - Might not be accepted by untrained users of the data

- Negative estimates could be avoided by calibrating only to aggregated values from $\widetilde{T}$
- Unbiasedness will only be guaranteed at the selected level of aggregation


Share of Employees with High Wages
Among Employees with Medium Skills (in %)
original data
cluster size 50
cluster size 100
cluster size 200
0.00 - 22.46
22.47 - 26.52
26.53 - 30.39
30.40 - 35.47
35.48 - 100.00