Institute for Employment Research

The Research Institute of the Federal Employment Agency

IAB

Another Piece in the Data Dissemination Puzzle - Synthetic Scientific-Use-Files for the IAB Establishment Panel

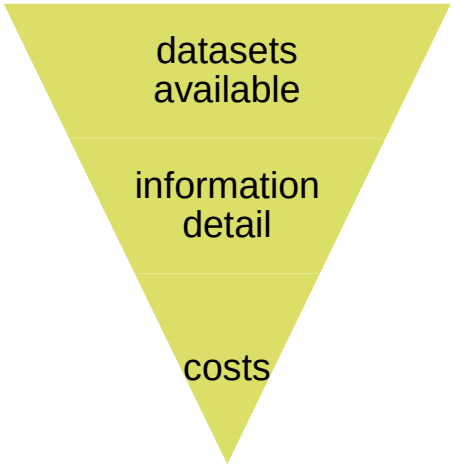
4th User Conference of the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB)

08. April 2011, Nürnberg

Jörg Drechsler

Current Situation in Germany

- Simplified data access for academic researchers
- Data are required to be de facto anonymous
- Three access channels
  - Onsite Access
  - Remote Execution
  - Scientific-Use-Files



Access to Business Data

- Almost no Scientific-Use-Files for business data available in Germany
- Higher risk of disclosure
  - smaller populations
  - skewed distributions
  - all kinds of information available on businesses
  - Identification more attractive for intruders
  - high sampling rates for large businesses
- Standard methods like coarsening, top coding, and dropping variables is not sufficient to protect large businesses in the database
- Data perturbation methods like swapping, adding noise, or microaggregation would have to be applied on a very high level
- Release of high quality data is very difficult
- Multiply imputed synthetic datasets as a possible solution

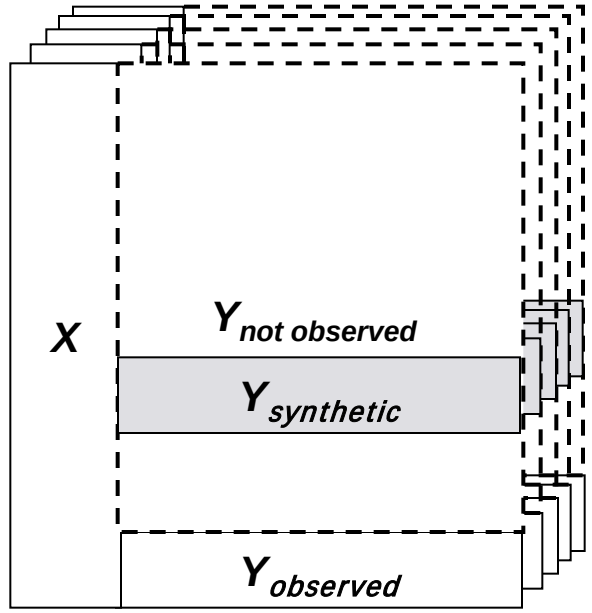
Synthetic Datasets

- Idea is closely related to multiple imputation for nonresponse
- Generate synthetic datasets by drawing from the posterior predictive distribution for new data given the observed data
- Not the missing values but the sensitive values are replaced with a set of plausible values given the original data
- Generate multiple draws from the posterior predictive distribution to be able to obtain valid variance estimates from the synthetic data

Synthetic Datasets

- Three steps necessary for data release:
  - Fit model to the original data
  - Repeatedly draw from that model to generate multiple synthetic datasets
  - Release these datasets to the public
- Over the years different designs for generating synthetic data evolved
- Two main approaches: Fully synthetic datasets and partially synthetic datasets

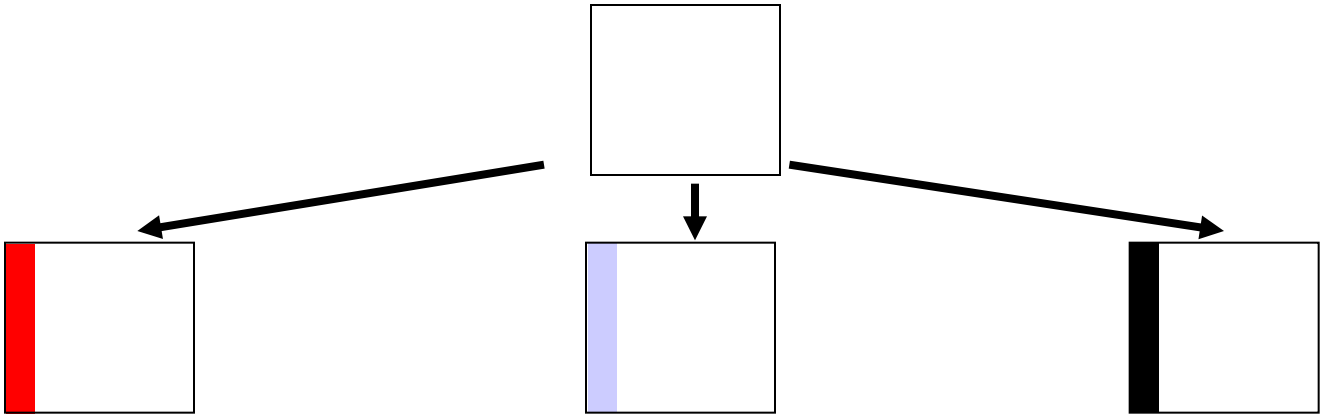
Fully synthetic datasets (Rubin 1993)



- advantages:
  - data are fully synthetic
  - re-identification of single units almost impossible
  - all variables are still fully available
- disadvantages:
  - strong dependence on the imputation model
  - setting up a model might be difficult/impossible

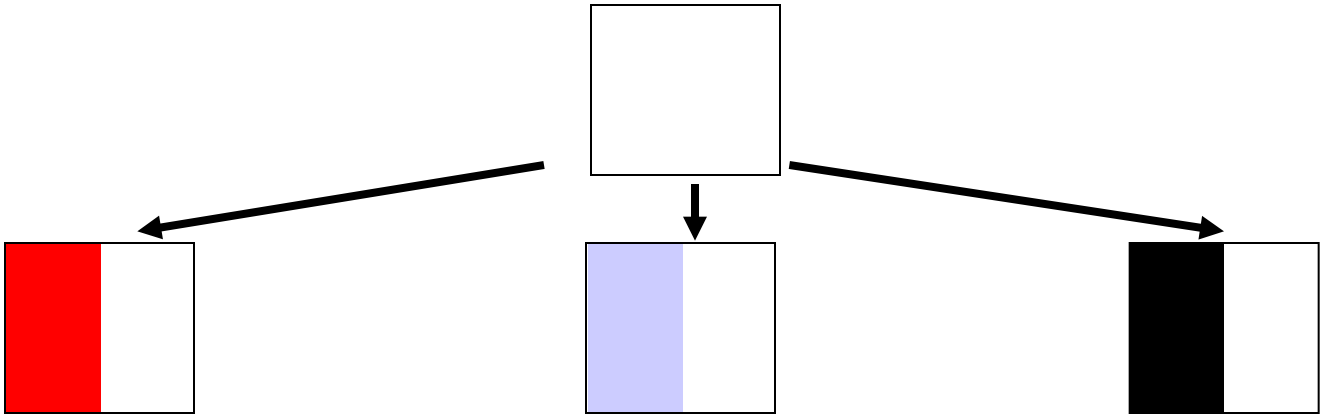
Partially synthetic datasets (Little 1993)

- only potentially identifying or sensitive variables are replaced



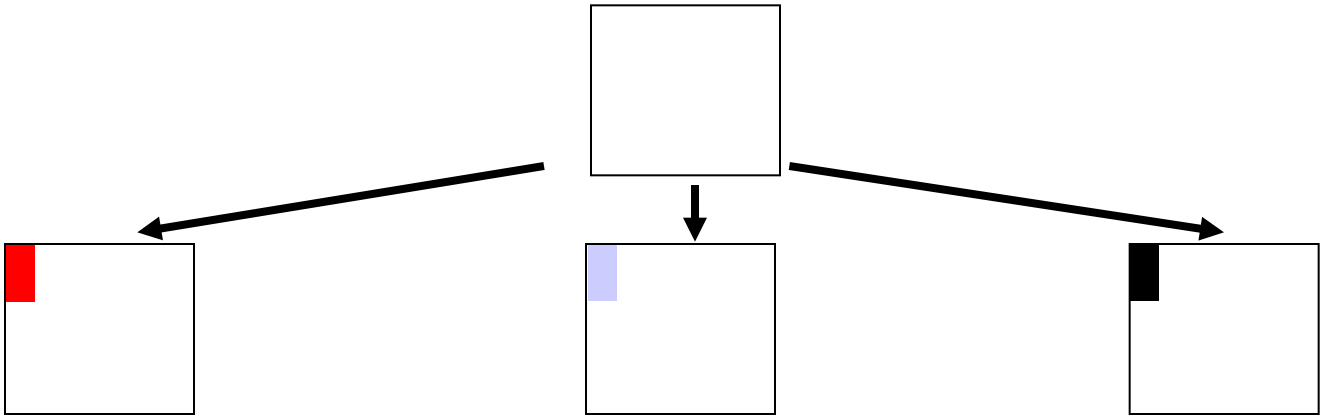
Partially synthetic datasets (Little 1993)

- only potentially identifying or sensitive variables are replaced



Partially synthetic datasets (Little 1993)

- only potentially identifying or sensitive variables are replaced



- advantages: - model dependence decreases
  - models are easier to set up
- disadvantages: - true values remain in the dataset
  - disclosure might still be possible

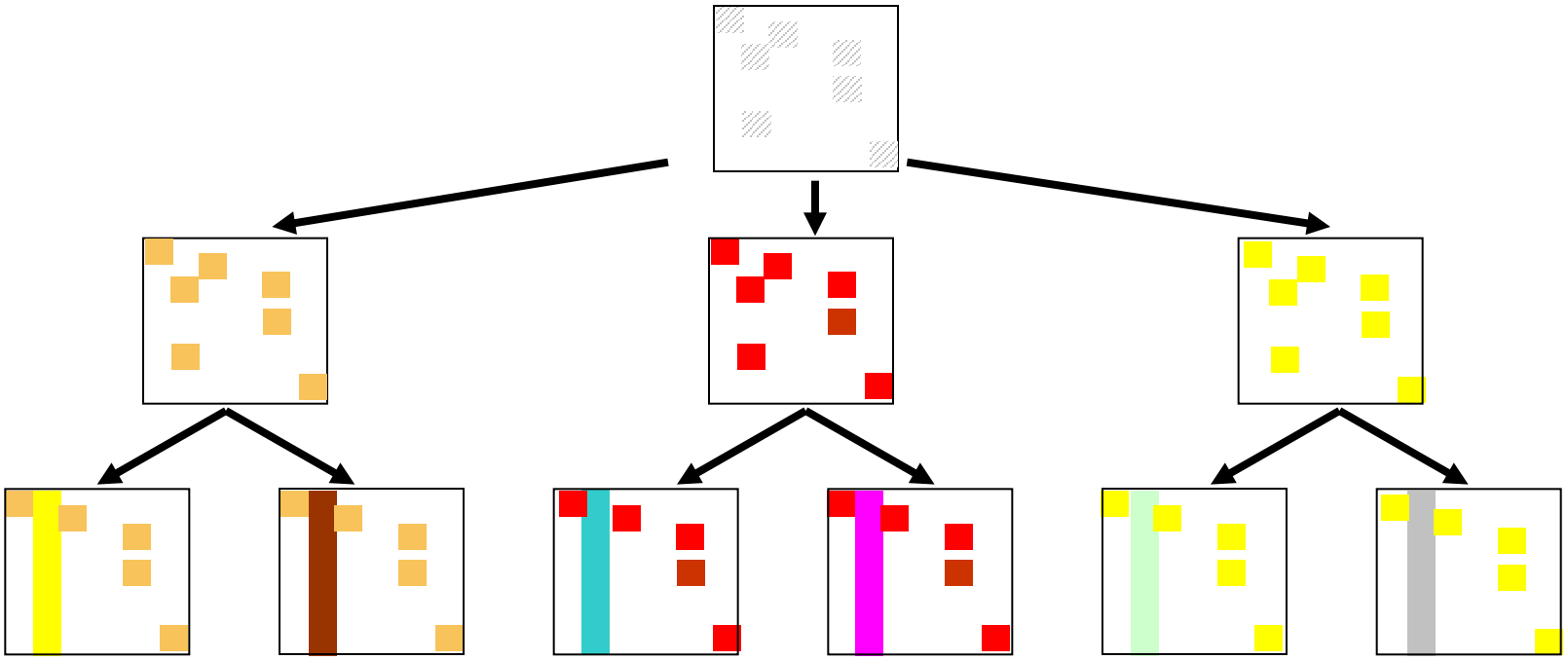
Synthetic data analysis

- Analysis based on the synthetic data is straight forward for the user
  - Analyse each synthetic dataset separately
  - Combine the results from the different datasets to obtain final estimates

Synthesis of the IAB Establishment Panel

- Only de facto anonymity is required for SUFs
- Partially synthetic data to leave as much data unchanged as possible
- Restriction to one wave (2007)
- Original data are subject to nonresponse
- Need to impute all missing values first

Two stage imputation for nonresponse and SDC



The synthesis task

- Synthesis of the complete wave 2007 of the IAB Establishment Panel
- Almost all continuous variables are synthesized
- Combination of variables that could be used for re-identification purposes (e.g. region, industry, establishment size) and sensitive variables (e.g. turnover, subsidies)
- All records are synthesized for each variable
- Missing values are imputed before the synthesis (m=5)
- r=5 synthetic datasets for every imputed dataset (m*r=25 datasets)

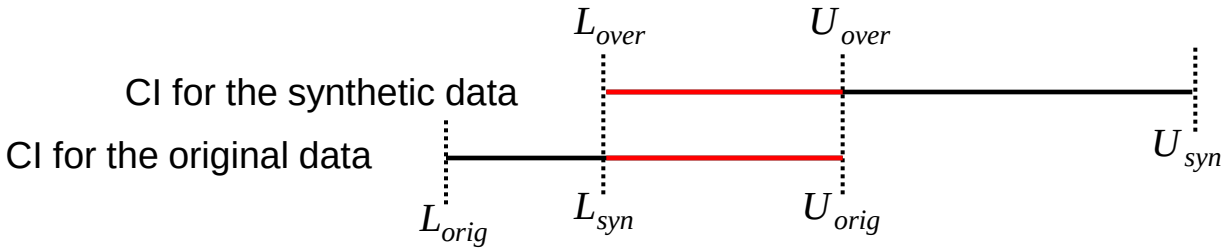
Analytical validity of the synthetic datasets

- Two regressions suggested by colleagues at the IAB
- First regression:
  - dependent variable: part-time yes/no
  - probit regression on 19 explanatory variables + industry dummies
- Second regression:
  - Dependent variable: expected employment trend (decrease, no change, increase)
  - ordered probit on 38 variables + industry dummies
- Both regressions are computed separately for West and East Germany

Confidence interval overlap

- Suggested by Karr et al. (2006)
- Measure the overlap of CIs from the original data and CIs from the synthetic data
- The higher the overlap, the higher the data utility
- Compute the average relative CI overlap for any estimate of interest

$$J_k = \frac{1}{2}\left[\frac{U_{over,k} - L_{over,k}}{U_{orig,k} - L_{orig,k}} + \frac{U_{over,k} - L_{over,k}}{U_{syn,k} - L_{syn,k}}\right]$$

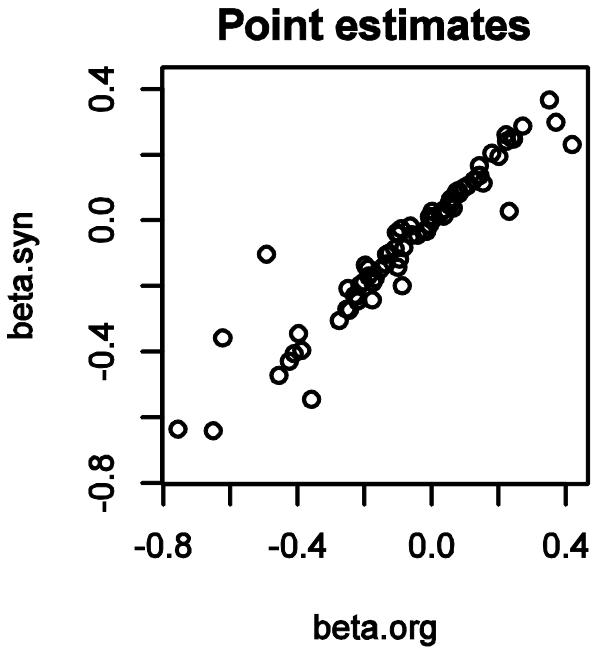


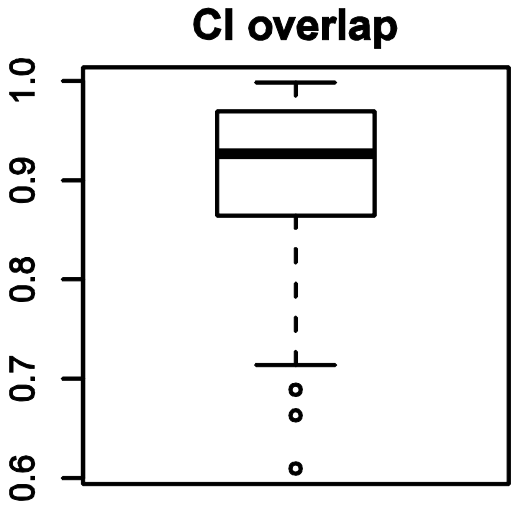
Regression results for West Germany

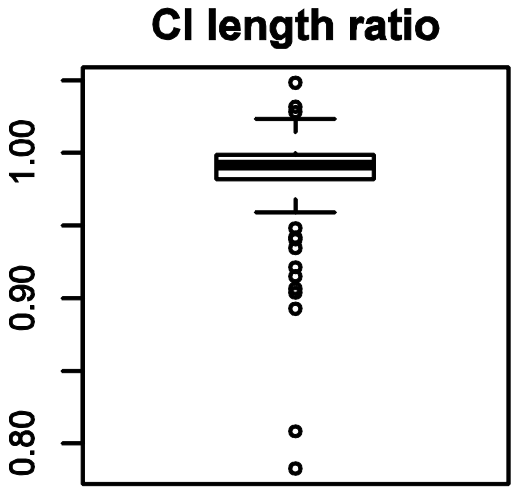
	beta org.	*beta syn.*	*J.k.beta*	*z-score org.*	*z-score syn.*	*CI length ratio*
Intercept	-0.809	-0.752	0.87	-7.23	-6.85	0.99
5-10 employees	0.443	0.437	0.97	8.52	7.99	1.06
10-20 employees	0.658	0.636	0.90	11.03	10.88	0.98
20-50 employees	0.797	0.785	0.95	13.02	12.36	1.04
100-200 employees	0.892	0.908	0.96	9.23	9.48	0.99
200-500 employees	1.131	1.125	0.99	9.99	9.87	1.01
>500 employees	1.668	1.641	0.97	8.22	8.33	0.97
growth in employment exp.	0.010	0.006	0.98	0.18	0.12	0.99
decrease in emp. expected	0.087	0.100	0.96	1.11	1.27	1.00
share of female workers	1.449	1.366	0.73	17.63	18.71	0.89
share of employees with university degree	0.319	0.368	0.91	2.18	2.59	0.97
share of low qualified workers	1.123	1.148	0.93	12.17	11.87	1.05
share of temporary employees	-0.327	-0.138	0.75	-1.74	-0.71	1.05
share of agency workers	-0.746	-0.856	0.88	-3.09	-4.24	0.84
employment in the last 6 month	0.394	0.369	0.87	8.33	7.82	1.00
dismissal in the last 6 months	0.294	0.279	0.92	6.38	6.03	1.00
foreign ownership	-0.113	-0.117	0.99	-1.33	-1.38	0.99
good or very good profitability	0.029	0.033	0.98	0.72	0.82	0.99
salary above collective wage agreement	0.020	0.031	0.95	0.35	0.54	0.99
collective wage agreement	0.016	0.007	0.95	0.31	0.13	0.97

- Average CI overlap: 0.92

Results for the second regression







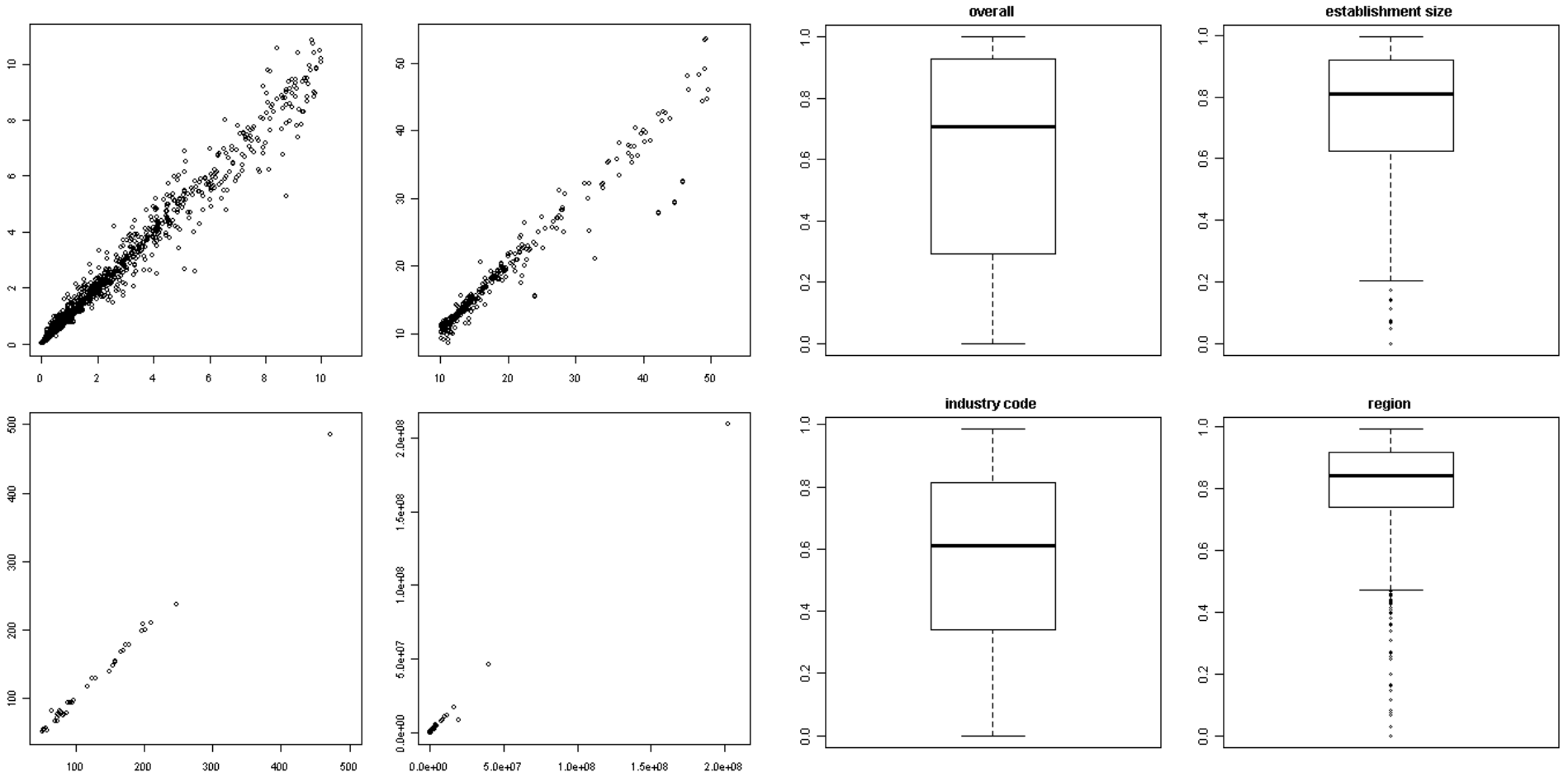
- Average CI overlap: 0.91

Minimum CI overlap: 0.61

Some Descriptive Statistics

- Compute the weighted overall mean and the mean in different subgroups for all continuous variables
- Subgroups defined by:
  - region (16 categories)
  - industry (17 categories)
  - establishment size (10 categories)
- All cells with $\geq$ 100 observations above zero are evaluated
- 2,170 estimands

Point Estimates and Confidence Interval Overlap



Limitations

- Generating disclosure protected data that will provide valid results for any possible query is impossible
- Variables derived from other variables should be treated with care
- Useful SDC method should provide information for the user, which analyses might provide valid results
- Detailed meta-data on imputation models
- Guarantee to run all obtained results on the original data

Conclusions

- Datasets ready and available through the FDZ
- Detailed documentation on the FDZ homepage
- First try but promising initial results
- Looking forward for feedback
- Hopefully the data can be released as Campus Files in the future
- One wave of limited use
- Next step: Develop longitudinal synthetic data

Institute for Employment Research

The Research Institute of the Federal Employment Agency

IAB

Thank you for your attention