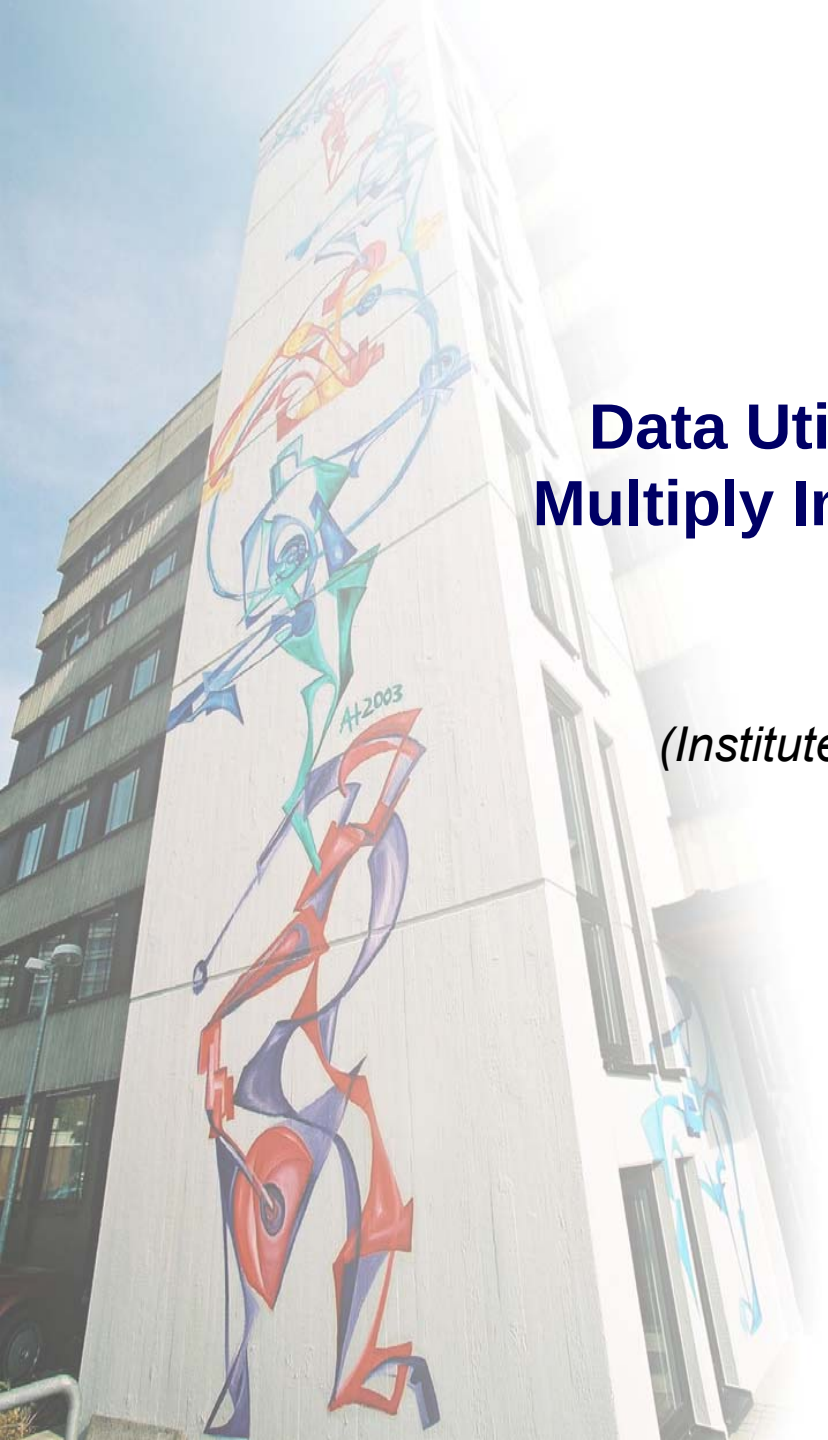


Data Utility vs. Disclosure Risk for Multiply Imputed Synthetic Datasets

Jörg Drechsler
(Institute for Employment Research, Germany)

3rd IAB Workshop on Confidentiality and Disclosure - SDC for Microdata,
Nürnberg, 20. November 2008

Overview

- Background
- Synthetic Datasets for The IAB Establishment Panel
- Fully vs. Partially Synthetic Datasets
  - Data Utility
  - Disclosure Risk
- The Impact of The Number of Imputations on
  - Data Utility
  - Disclosure Risk
- Imputation in Two Stages
- Conclusions/Future Work

Project Background

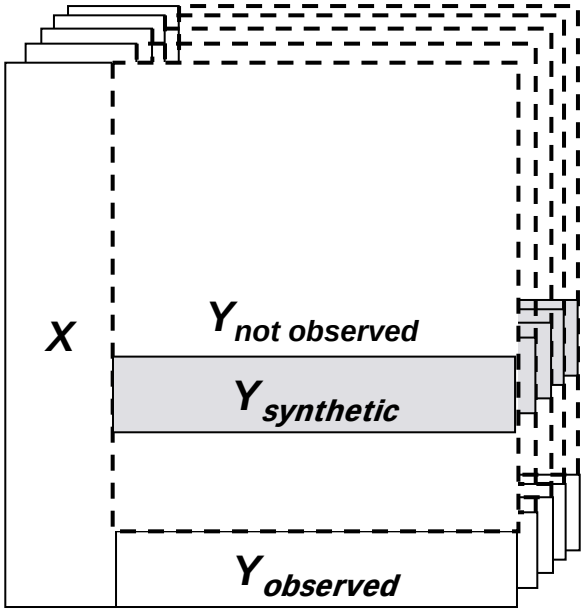
- Yearly conducted establishment survey (IAB Establishment Panel)
- Strong demand for access from external researchers
- Only on-site and remote access possible so far
- High costs in terms of time and money
- Project-goal: Generate synthetic datasets of the survey for release as scientific use file

- Project start: summer 2006

The IAB Establishment Panel

- Annually conducted establishment survey
- Since 1993 in Western Germany, since 1996 in Eastern Germany
- *Population*: All establishments with at least one employee covered by social security
- *Source*: Official Employment Statistics
- Sample of more than 16.000 establishments in the last wave
- *Contents*: employment structure, changes in employment, business policies, investment, training, remuneration, working hours, collective wage agreements, works councils

Fully synthetic datasets (Rubin 1993)

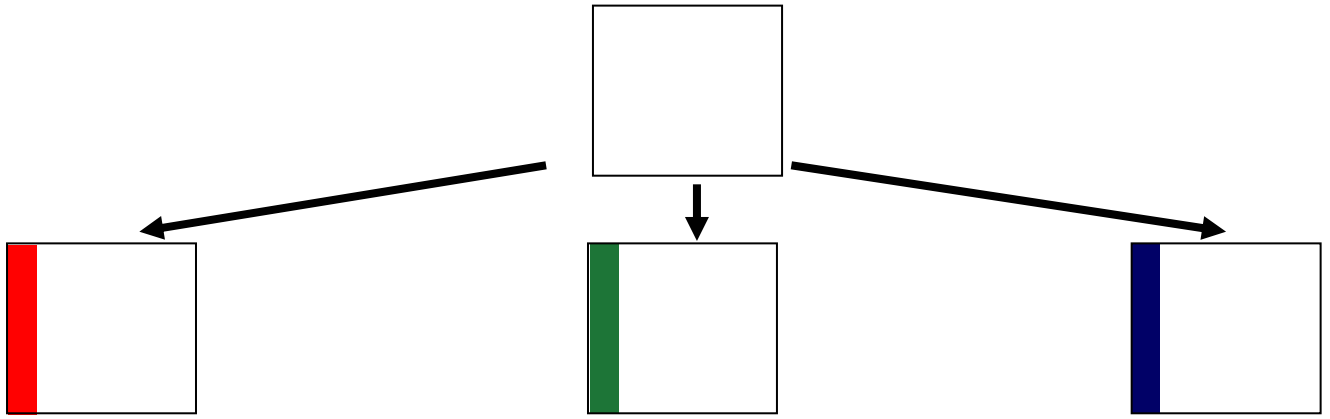


- advantages: - data are fully synthetic
  - re-identification of single units almost impossible

- disadvantages: - strong dependence on the imputation model
  - setting up a model might be difficult/impossible

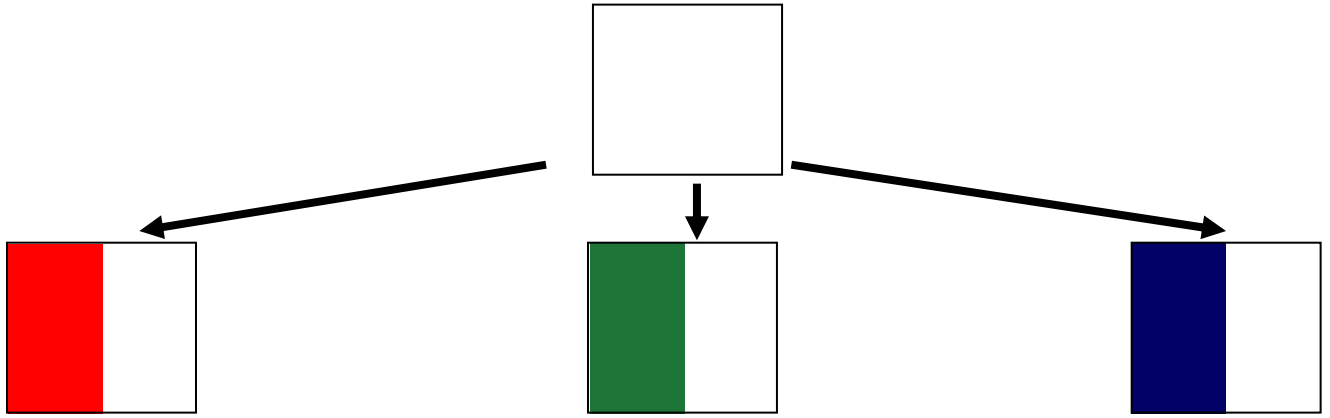
Partially synthetic datasets (Little 1993)

- only potentially identifying or sensitive variables are replaced



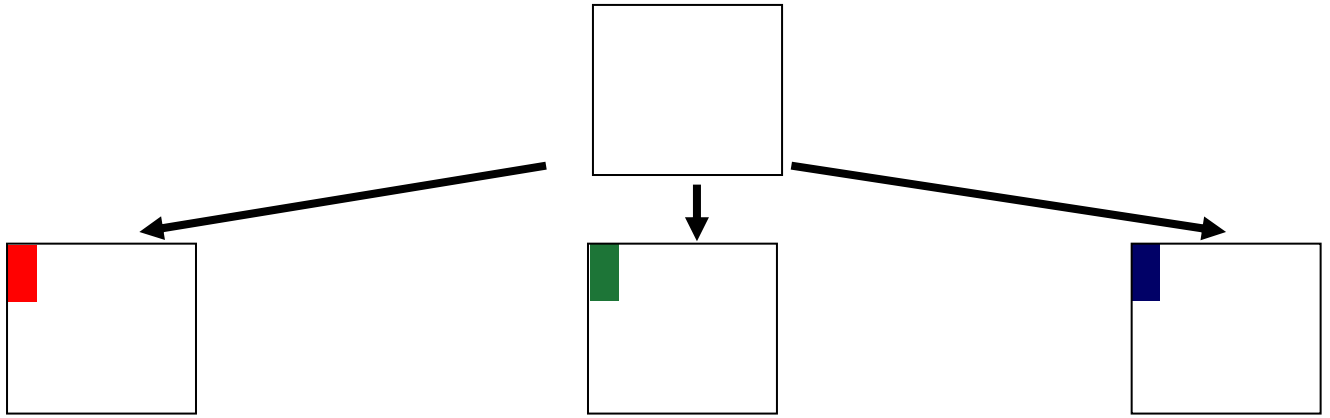
Partially synthetic datasets (Little 1993)

- 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

Quantities needed for inference from synthetic datasets

- in each imputed dataset

$$q_i = Q(d_i) \qquad u_i = U(d_i)$$

$$\bar{q}_m = \sum_{i=1}^{m} q_i / m$$

$$b_m = \sum (q_i - \bar{q}_m)^2 / (m-1)$$

$$\bar{u}_m = \sum_{i=1}^{m} u_i / m$$

Inference From Synthetic Datasets

- Point estimate:

$$\bar{q}_m = \sum_i q_i$$

- Variance estimate for partially synthetic datasets

$$T_p = \bar{u}_m + b_m / m$$

- Variance estimate for fully synthetic datasets

$$T_f = (1+1/m)b_m - \bar{u}_m$$

- Adjustments for negative T_f

$$T_f^* = T_f + \frac{n_{syn}}{n} * \bar{u}_m$$

Overview

- Background
- Synthetic Datasets for The IAB Establishment Panel
- Fully vs. Partially Synthetic Datasets
  - Data Utility
  - Disclosure Risk
- The Impact of The Number of Imputations on
  - Data Utility
  - Disclosure Risk
- Imputation in Two Stages
- Conclusions/Future Work

Generating fully synthetic datasets for the IAB Establishment Panel

- Create a synthetic dataset for selected variables from the wave 1997 from the Establishment Panel
- Draw 10 new sample from the Official Employment Statistics using the same sampling design as for the Establishment Panel (Stratification by industry, size, and region)
- The number of observations in each sample equals the number of observations in the panel $n_s=n_p=7332$
- Every sample is imputed ten times using sequential regression
- Number of variables from the establishment panel: *48*
- Imputations are generated using *IVEware* by Raghunathan, Solenberger and Hoewyk (2001)

Imputation procedure for partially synthetic data

- Only two variables are synthesized: - number of employees
  - industry (16 categories)
- Same variables for the imputation models
- Imputation by sequential regression
- Imputation model: - multinomial logit for the industry
  - linear model for the cubic root of the nb of employees
  - 4 independent linear models defined by quartiles for the establishment size
- Imputations based on own coding in *R*.

Overview

- Background
- Synthetic Datasets for The IAB Establishment Panel
- Fully vs. Partially Synthetic Datasets
  - Data Utility
  - Disclosure Risk
- The Impact of The Number of Imputations on
  - Data Utility
  - Disclosure Risk
- Imputation in Two Stages
- Conclusions/Future Work

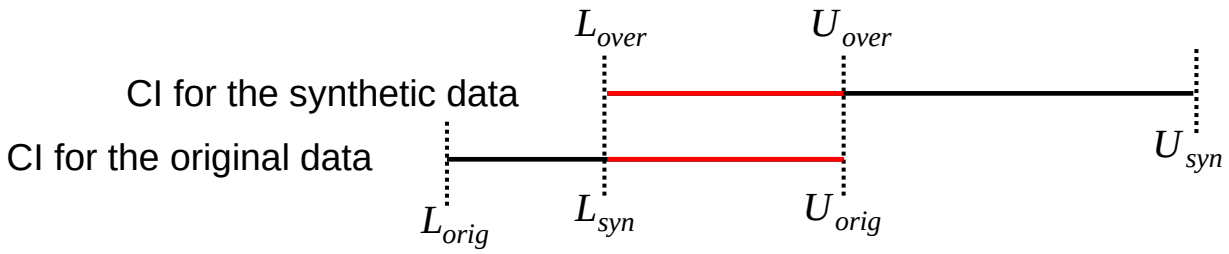
Analytical validity

- Compare results from the original data with results from the synthetic data
- **Regression analysis:**
  - Zwick (2005) analyses the productivity effects of different continuing vocational training forms in Germany
  - Probit regression to explain, why firms offer vocational training
  - 13 Explanatory variables including: Share of qualified employees, establishment size, industry, collective wage agreement, high qualification needs expected…
- **Descriptive analysis:**
  - Average number of employees in 16 industry classes
- Two data utility measures:
  - Comparison of the point estimates from the original dataset and the synthetic datasets
  - confidence interval overlap

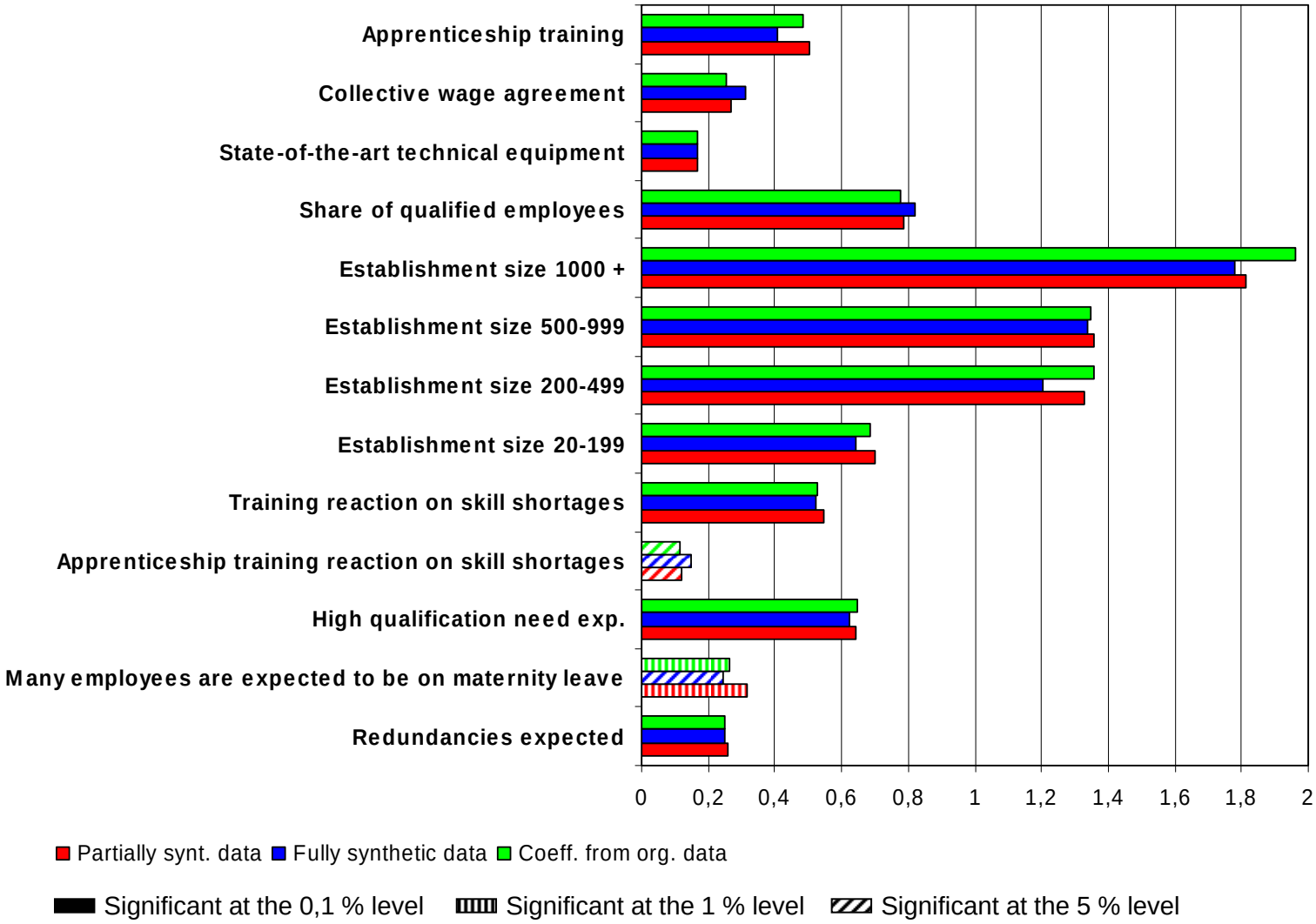
Confidence interval overlap

- 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 β_k

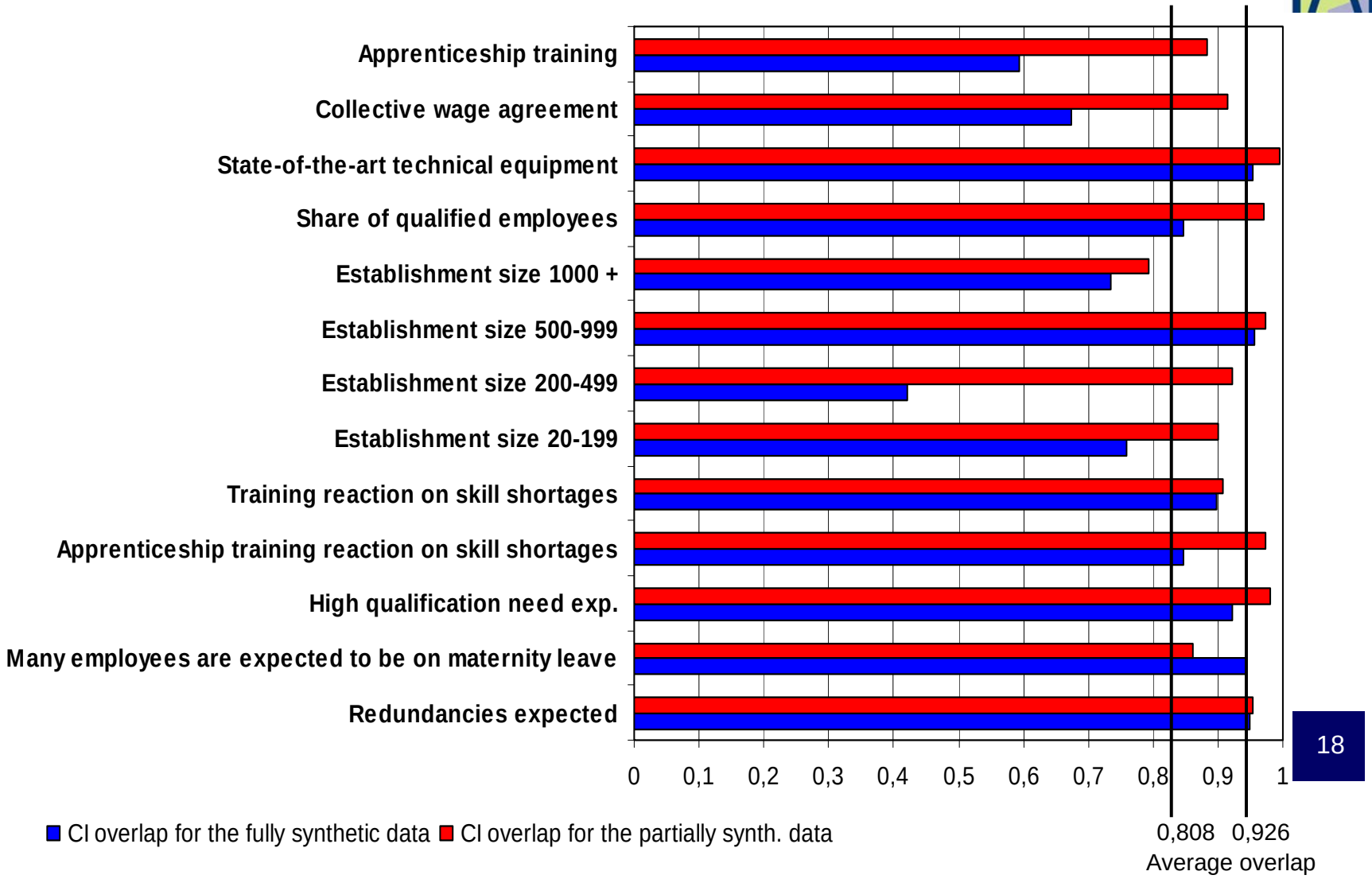
$$J_k = \frac{1}{2}\left[\frac{U_{over,k} - L_{over,k}}{U_{orig,k} - L_{orig,k}} + \frac{U_{over,k} - L_{over,k}}{U_{syn,k} - L_{syn,k}}\right]$$



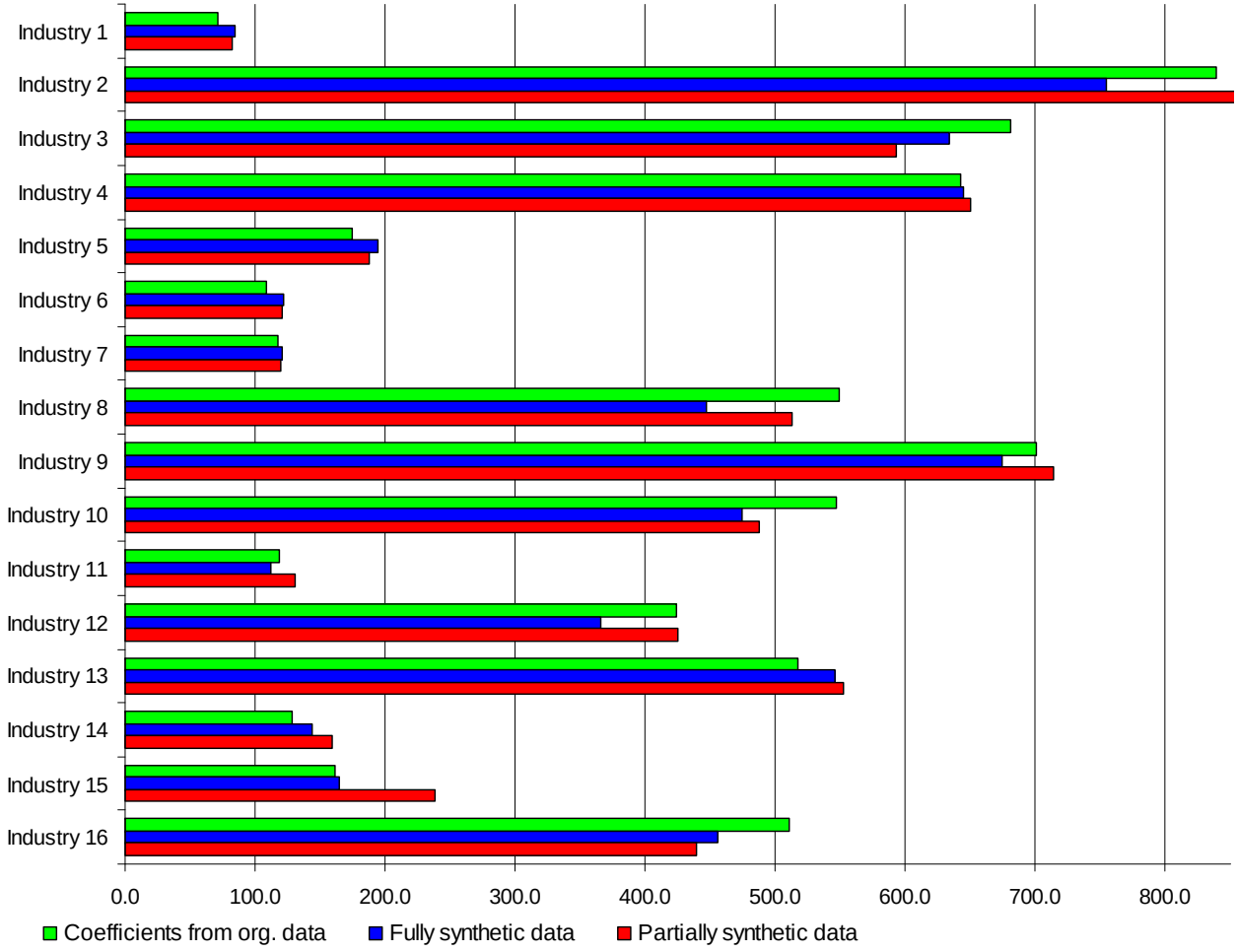
Results from the regression (Zwick 2005)



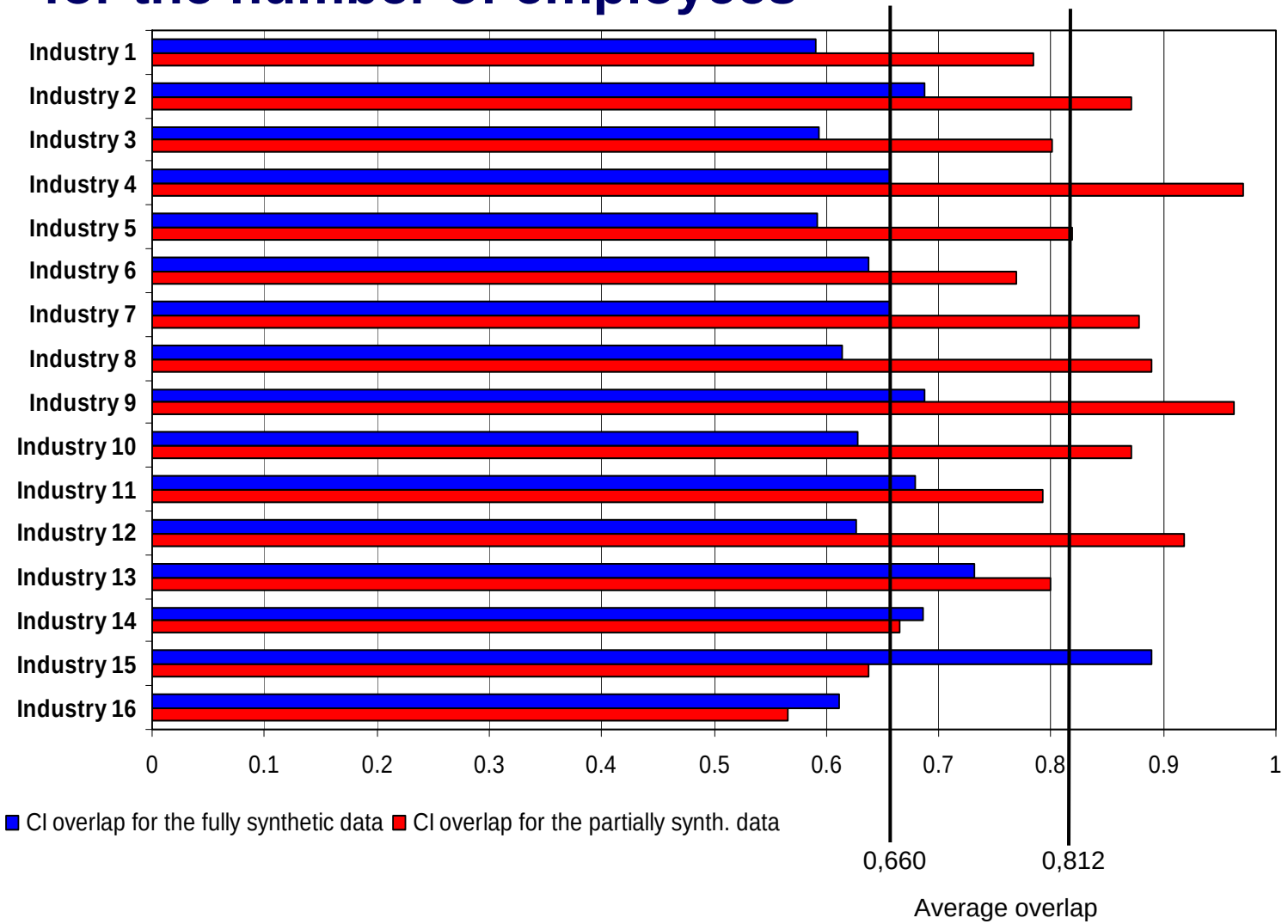
Average confidence interval (CI) overlap for the estimates from the regression



Average number of employees in 16 industry classes



Average confidence interval (CI) overlap for the number of employees



Overview

- Background
- Synthetic Datasets for The IAB Establishment Panel
- Fully vs. Partially Synthetic Datasets
  - Data Utility
  - Disclosure Risk
- The Impact of The Number of Imputations on
  - Data Utility
  - Disclosure Risk
- Imputation in Two Stages
- Conclusions/Future Work

Disclosure risk

- Difficult to compare between partially and fully synthetic datasets
- Disclosure risk is low for fully synthetic datasets, although not zero
- DR is higher for partially synthetic datasets, because:
  - True values remain in the dataset
  - Only survey respondents are included
- For partially synthetic datasets a careful disclosure risk evaluation is necessary

Measures for Disclosure Risk

- disclosure risk measures based on Reiter & Mitra (to appear)
- Compute probabilities of reidentification for each record *j* *(j=1,…,n)* in the released dataset

- Assumptions: - Intruder has exact information for some target records *t* from external databases
  - intruder knows who participated in the survey

- Let t_0 be the unique identifier for the target record
- Let d_{j0} be the identifier for record *j* in the released data *D*, *j=1,…,s*
- Intruders goal: match if $t_0=d_{j0}$; don't match if $t_0 \neq d_{j0}$

Measures for Disclosure Risk II

- Let *J* be a random variable with

$$J = \begin{cases} j & \text{for } d_{j0} = t_0 \text{ and } j \in D \\ s+1 & \text{for } d_{j0} = t_0 \text{ and } j \notin D \end{cases}$$

$$\Pr(J = j \mid t, D, M)$$

with: *D* set of released synthetic datasets

M any additional information about the generation of *D*

- Intruder does not know actual values in Y_{rep}
- Integrate over its possible values
- Monte Carlo approach to estimate $\Pr(J = j \mid t, D, M)$

Assumptions for this study

- Assumptions about the intruder:
  - Intruder has exact information on the number of employees and the industry code for all establishments in the survey
  - No information about the generation of the data is released

- The intruder treats simulated values as plausible values of Y_{rep}

- Observation from the survey is considered a match if

  - $industry_{obs} = industry_{target}$

$$nb.emp_{\text{target}} - sd_s(nb.emp_{\text{target}}) \leq nb.emp_{obs} \leq nb.emp_{\text{target}} + sd_s(nb.emp_{\text{target}})$$

$sd_s(nb.emp_{\text{target}})$ standard dev. for the nb of emp. in cell *s*

- 20 different cells *s* defined by quantiles

Measures for Disclosure Risk III

- Average matching probability

$$p_{match,i} = (1/m)\sum_k (1/N_k) I_i$$

with N_k = nb of records that fulfill the matching criteria in dataset k
I_i = 1 if record i is among the N_k records, 0 otherwise
m = number of synthetic datasets

- Intruder selects record j with highest value of $Pr(J=j|t,D,M)$

- Further definitions:

c_j = number of records with $max(p_{match,i})$ for target t_j
I_j = 1 if true match is among the c_j units, 0 otherwise
K_j = 1 if $c_j I_j$=1, 0 otherwise

- Expected match risk $\sum_j (1/c_j) I_j$

- True match risk $\sum_j K_j$

Results for this study

- Expected match risk = True match risk = 139
- 7,330 single matches for the 7,332 records
- No common pattern for identified records
- No difference in the distributions for region and industry between the identified records and the underlying data
- Small and large establishments are identified
- Disclosure risk increases with establishment size

Establishment size	Percentage of identified records
<100	1.38
100-1,000	1.87
>1,000	5.21

Summary of Results

- Generating synthetic datasets can be a useful method for SDC
- Advantages for partially synthetic datasets:
  - Higher data validity
  - Imputation models easier to set up
  - Lower risk of biased imputations
- Disadvantages for partially synthetic datasets:
  - Higher risk of disclosure
    - True values remain in the dataset
    - Only survey respondents are included
  - Careful disclosure risk evaluation necessary
- The IAB will release partially synthetic datasets

Overview

- Background
- Synthetic Datasets for The IAB Establishment Panel
- Fully vs. Partially Synthetic Datasets
  - Data Utility
  - Disclosure Risk
- The Impact of The Number of Imputations on
  - Data Utility
  - Disclosure Risk
- Imputation in Two Stages
- Conclusions/Future Work

Data Utility vs. Disclosure Risk

- users benefit from large numbers of synthetic datasets
  - Increasing *m* will drive down the variability introduced by the imputation
- large number of imputations can increase disclosure risk
  - intruders can use the additional information for re-identification purposes
- Agencies have to address this trade-off

The Impact of Increasing *m* on Data Utility

Average number of employees by industry					
	original data	*m=3*	*m=10*	*m=50*	*m=100*
industry1	71.47	84.23	84.17	82.61	82.36
industry2	839.11	919.37	851.19	870.21	852.88
industry3	681.07	557.72	574.52	594.42	593.1
industry4	642.86	639.9	644.76	643.5	649.64
industry5	174.46	179.77	176.04	183.51	187.38
industry6	108.89	132.45	121.77	120.82	120.69
industry7	117.08	111.59	112.89	117.1	119.61
industry8	548.67	455.35	504.27	514.21	512.99
industry9	700.7	676.92	689.44	711.79	713.39
industry10	546.97	402.41	490.32	499.3	487.68
industry11	118.64	142.74	130.2	132.14	130.98
industry12	424.31	405.63	414.9	424.5	425.21
industry13	516.74	526.14	549.06	550.16	551.92
industry14	128.09	185.79	167.11	160.04	158.99
industry15	161.98	292.76	233.37	221.87	238.09
industry16	510.84	452.85	449.9	441.51	439.33

The Impact of Increasing *m* on Data Utility

CI overlap for different numbers of imputation				
	m=3	*m=10*	*m=50*	*m=100*
industry 1	0.778	0.77	0.777	0.782
industry 2	0.844	0.893	0.853	0.874
industry 3	0.73	0.776	0.797	0.8
industry 4	0.983	0.992	0.995	0.971
industry 5	0.92	0.935	0.863	0.817
industry 6	0.605	0.749	0.764	0.767
industry 7	0.809	0.82	0.863	0.876
industry 8	0.692	0.862	0.894	0.89
industry 9	0.926	0.966	0.968	0.963
industry 10	0.66	0.876	0.897	0.871
industry 11	0.609	0.804	0.773	0.792
industry 12	0.903	0.912	0.916	0.918
industry 13	0.946	0.814	0.809	0.799
industry 14	0.408	0.589	0.655	0.664
industry 15	0.586	0.639	0.654	0.638
industry 16	0.666	0.645	0.583	0.566
average	0.754	0.815	0.816	0.812

Data Utility over 10 simulation runs

Average CI overlap for differnt numbers of imputation (10 Simulation runs)	*m=3*	*m=10*	*m=50*	*m=100*
simulation 1	0.754	0.815	0.816	0.812
simulation 2	0.818	0.818	0.808	0.82
simulation 3	0.812	0.813	0.82	0.816
simulation 4	0.854	0.796	0.819	0.817
simulation 5	0.823	0.808	0.808	0.824
simulation 6	0.796	0.801	0.823	0.807
simulation 7	0.787	0.778	0.819	0.819
simulation 8	0.785	0.799	0.815	0.821
simulation 9	0.77	0.829	0.823	0.821
simulation 10	0.808	0.804	0.821	0.809
average	0.801	0.806	0.817	0.817

Overview

- Background
- Synthetic Datasets for The IAB Establishment Panel
- Fully vs. Partially Synthetic Datasets
  - Data Utility
  - Disclosure Risk
- The Impact of The Number of Imputations on
  - Data Utility
  - Disclosure Risk
- Imputation in Two Stages
- Conclusions/Future Work

The Impact of Increasing *m* on Disclosure Risk

Average match risk for different numbers of imputation				
	m=3	***m=10***	***m=50***	***m=100***
expected match risk	67.82	94.8	126.93	142.55
true match risk	35.2	82.5	126.1	142.4

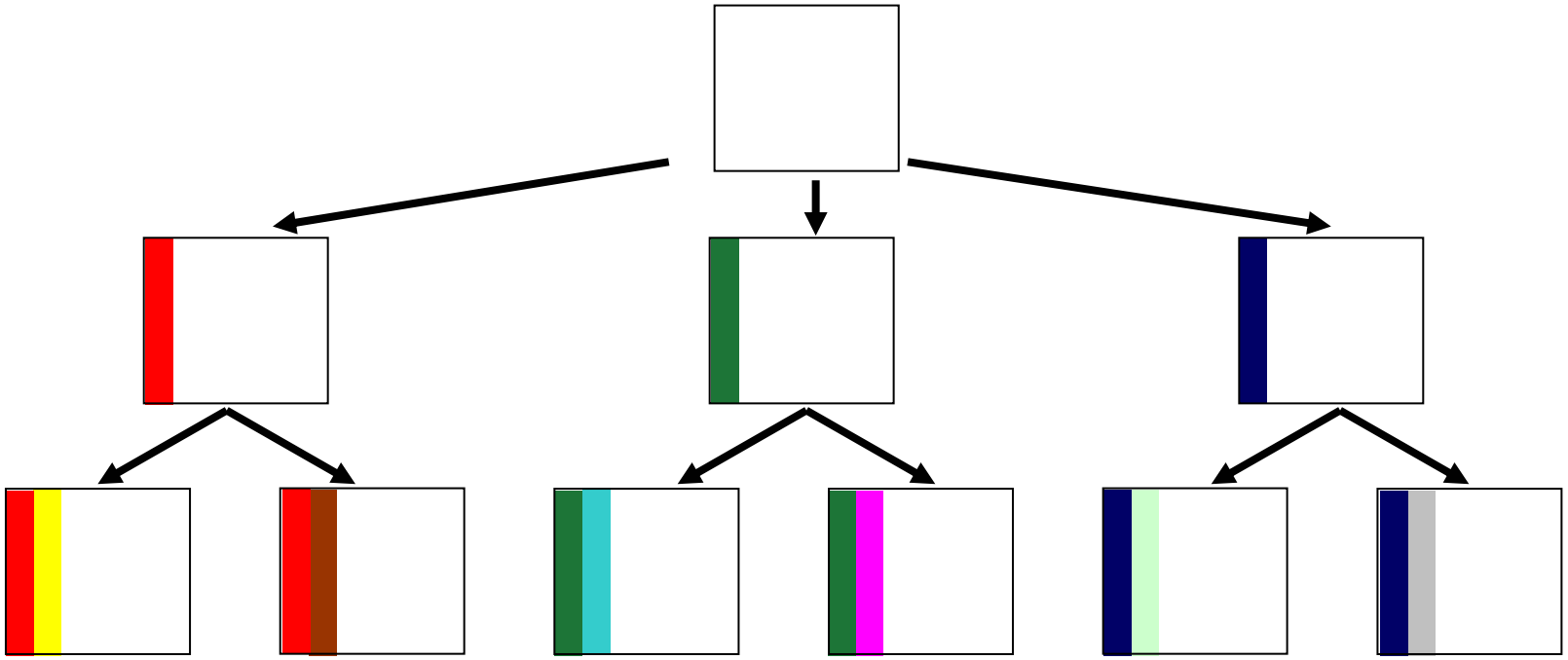
- Correctly identified records vary across the 10 replicates
  - m=50: - 619 records are correctly identified
    - more than half of them only once
    - only 10 establishments identified in all 10 replicates
  - m=10: - no establishment identified more than 7 times
- Increase in the number of imputations leads to increase in data utility and disclosure risk
- Obvious trade-off between data utility and disclosure risk

Overview

- Background
- Synthetic Datasets for The IAB Establishment Panel
- Fully vs. Partially Synthetic Datasets
  - Data Utility
  - Disclosure Risk
- The Impact of The Number of Imputations on
  - Data Utility
  - Disclosure Risk
- Imputation in Two Stages
- Conclusions/Future Work

Multiple Imputation in Two Stages

- some variables are imputed more often than others



Motivation for Two Stage Imputation

- reduction in labor costs

- address the trade-off between data utility and disclosure risk for partially synthetic datasets
    - two stage approach allows for treating variables separately depending on their individual disclosure risk

Data Utility and Disclosure Risk for Two Stage Imputation

Average CI overlap and match risk for different two stage imputations (10 simulation runs)			
	average J	*expected m.r.*	*true m.r.*
m=3,r=3	0.819	83.12	67.6
m=3,r=16	0.819	97.99	91.8
m=5,r=10	0.823	106.14	101.2
m=10,r=5	0.824	113.79	109.4
m=16,r=3	0.824	119.93	116.4

- Data utility is slightly increased (m_{10}=0.806, m_{50}=0.817)
- Disclosure risk is reduced ($m_{10,exp}$=94.8, $m_{10,tr}$=82.5, $m_{50,exp}$=126.93, $m_{50,tr}$=126.1)
- New combining rules necessary (Reiter & Drechsler (2007))

Overview

- Background
- Synthetic Datasets for The IAB Establishment Panel
- Fully vs. Partially Synthetic Datasets
  - Data Utility
  - Disclosure Risk
- The Impact of The Number of Imputations on
  - Data Utility
  - Disclosure Risk
- Imputation in Two Stages
- Conclusions/Future Work

Conclusions

- Generating synthetic datasets can be a useful method for SDC
- Partially synthetic datasets tend to have higher data utility but for the price of a higher disclosure risk
- A two stage imputation approach can address the trade-off between data utility and disclosure risk

Future Work

- Release a SUF of the latest wave of the survey in 2009
- Long term goal: release complete longitudinal data
- Develop methods for selecting *m* and *r* based on risk-utility measures
- Develop methods to decide which variables should be imputed on stage one and which variables should be imputed on stage two

Thank you for your attention