Quality of linked firm-level and micro-aggregated datasets: The example of the ESSLait Micro Moments Database

Diana Iancu – Statistics Norway
Eva Hagsten – Statistics Sweden
Patricia Kotnik – University of Ljubljana

European Conference on Quality in Official Statistics
June 2014, Vienna

Introduction and roadmap

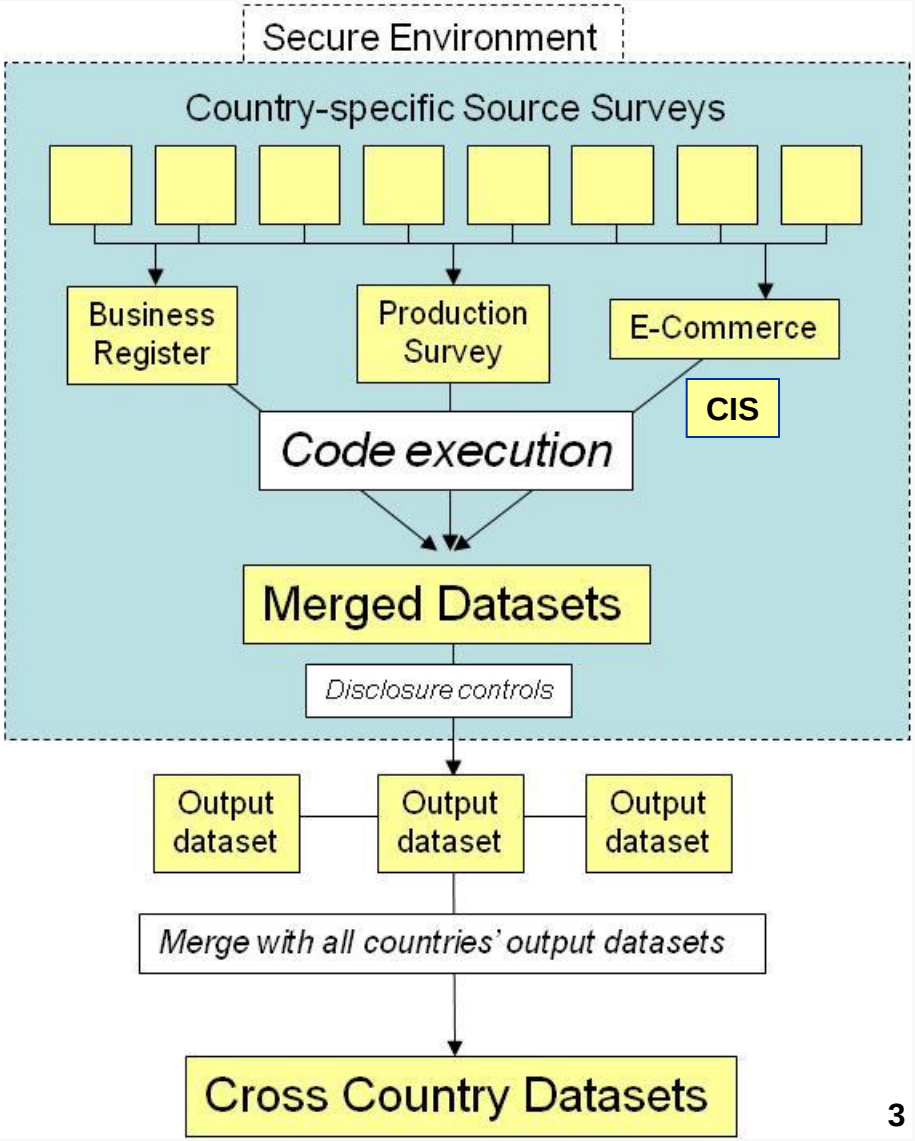
- The impact of stepwise linking and aggregation of information from firm-level datasets in several countries on the representativeness and usefulness of indicators from the ESSLait Micro Moments Database (MMD)
- Distributed microdata research (DMD) and sources
- Data linking and statistical properties of linked datasets
- Overlap across samples and over time
- Representativeness – ex-post re-weighted variables
  - Use in descriptive statistics
  - Use for marginal analyses
- Final remarks

Distributed microdata research (DMD) and sources

ICT Impacts (2006)
14 European countries
DMD Method with “Common Code” software

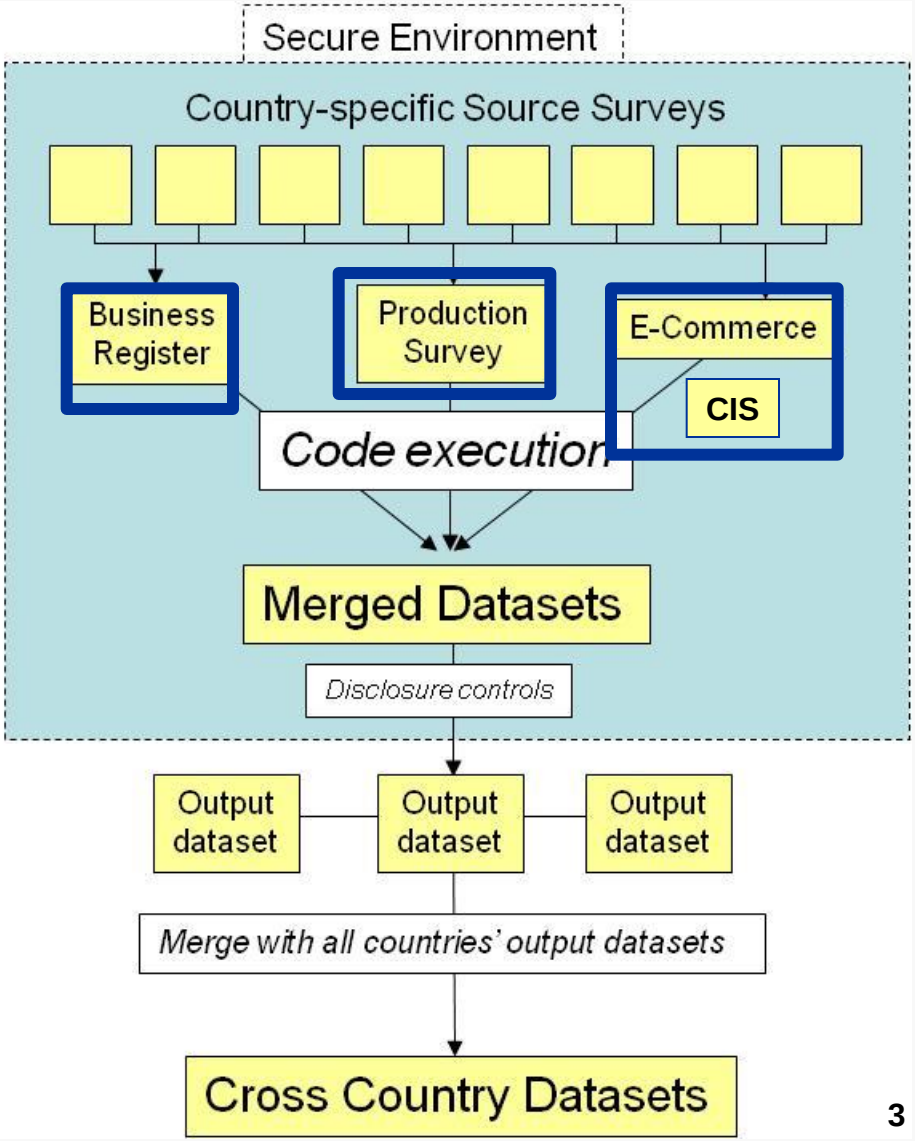
Distributed microdata research (DMD) and sources

ICT Impacts (2006)
14 European countries
DMD Method with “Common Code” software



Distributed microdata research (DMD) and sources

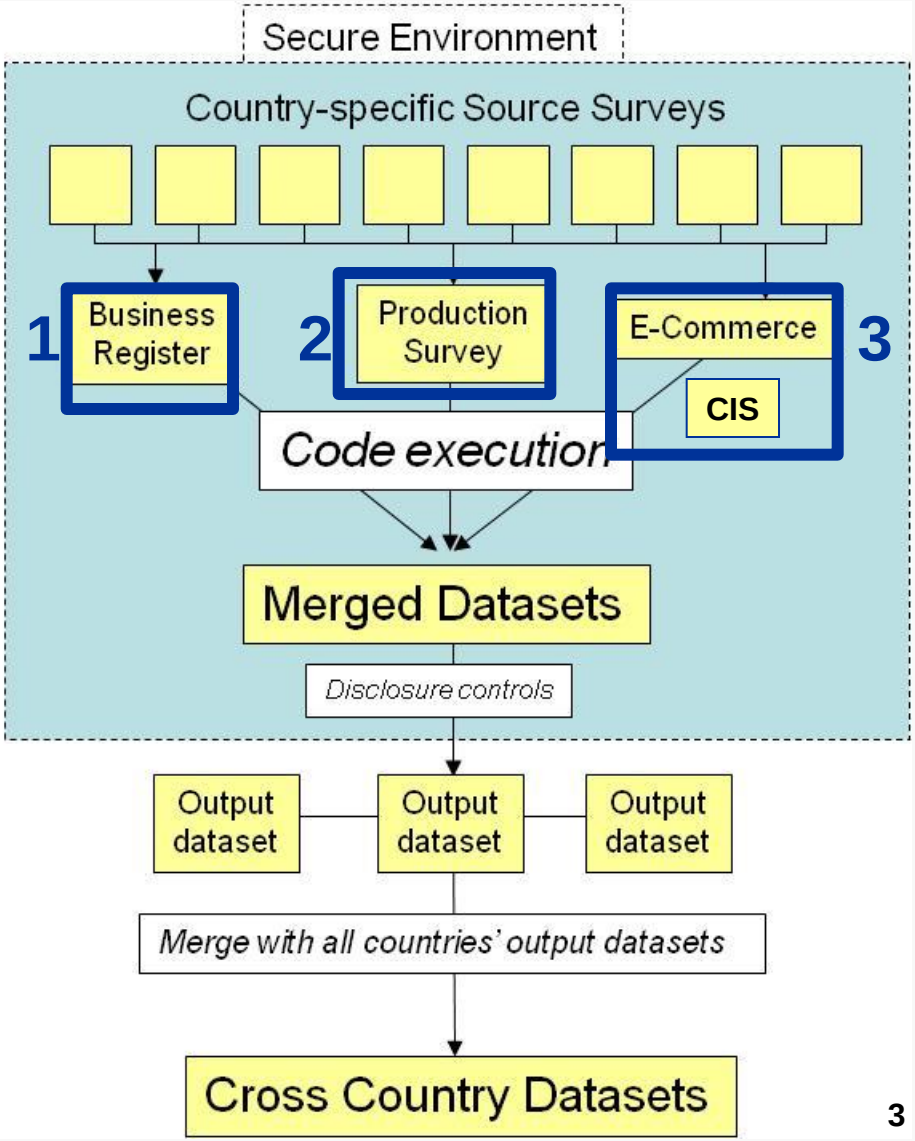
ICT Impacts (2006)
14 European countries
DMD Method with “Common Code” software



Distributed microdata research (DMD) and sources

ICT Impacts (2006)
14 European countries
DMD Method with “Common Code” software

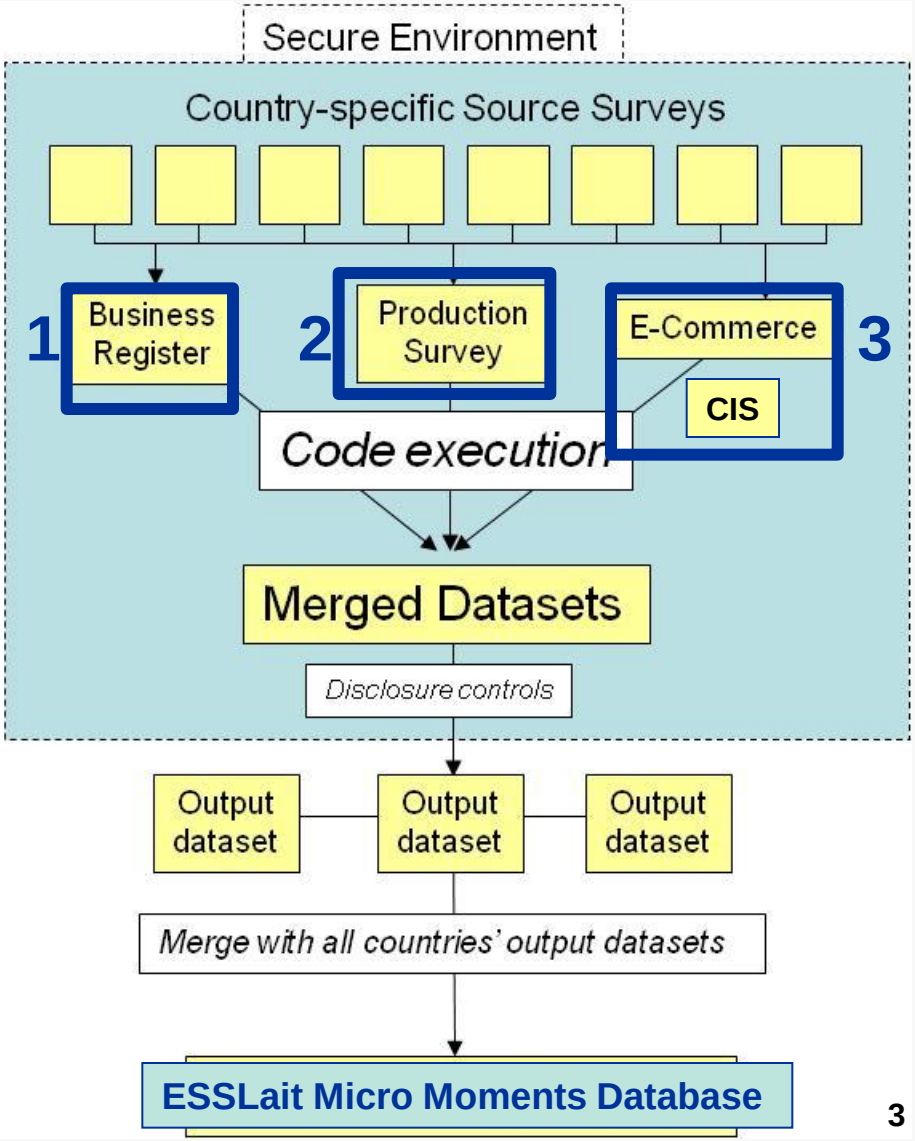
1. **Business Register BR: industry code, age, employment**
2. **Production Statistics PS: production values, exports, capital, employment, pay, educational achievement, ownership, affiliation**
3. **E-commerce Survey EC and Community Innovation Survey IS**



Distributed microdata research (DMD) and sources

ICT Impacts (2006)
14 European countries
DMD Method with “Common Code” software

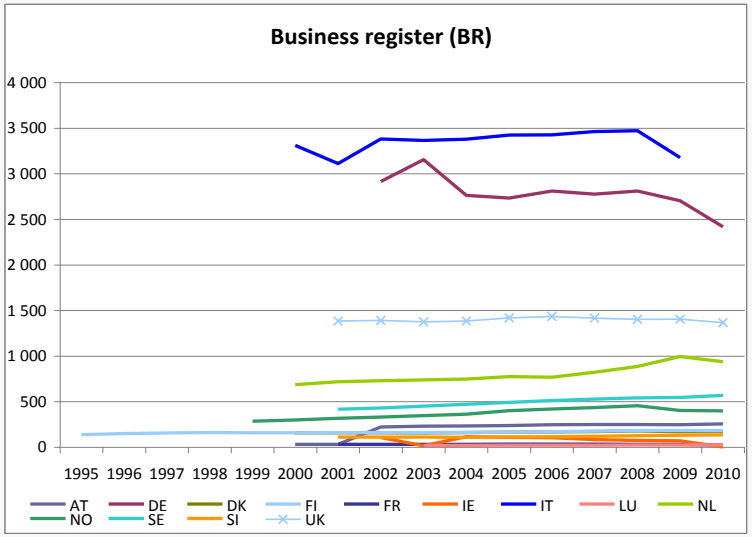
1. **Business Register BR: industry code, age, employment**
2. **Production Statistics PS: production values, exports, capital, employment, pay, educational achievement, ownership, affiliation**
3. **E-commerce Survey EC and Community Innovation Survey IS**

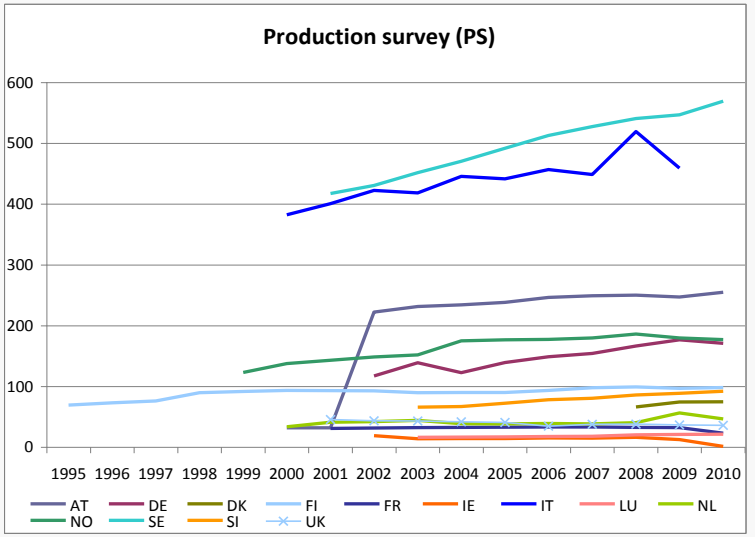


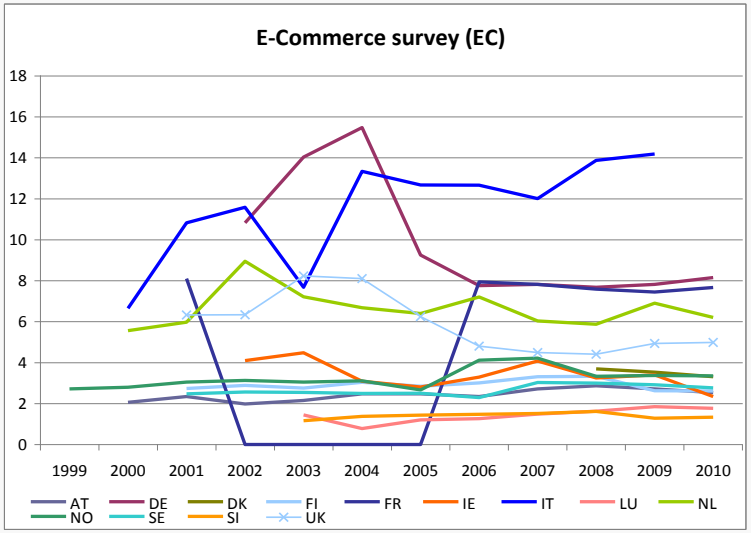
Data linking and statistical properties of linked datasets

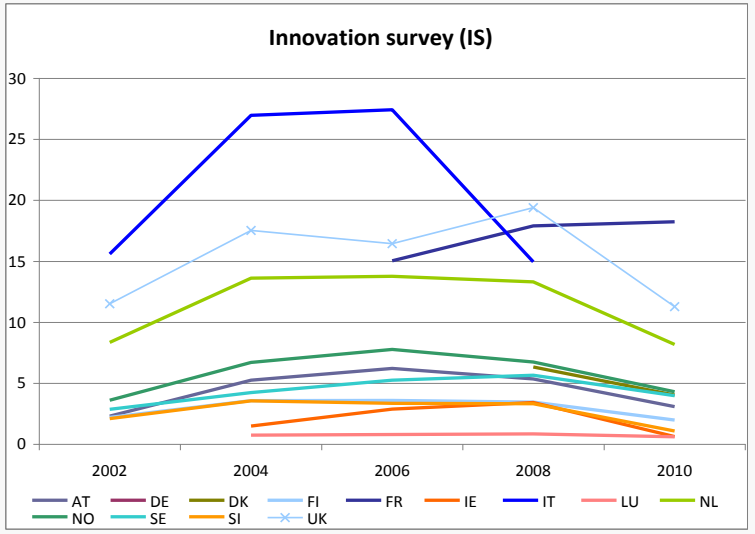
- Papers dealing with selection bias and sample representativeness in linked datasets: Chesher and Nesheim (2006), Ritchie (2004), Fazio et al. (2006)
- Multiple sources of bias
- Longitudinal data integrity issues
- Long-term solutions to dealing with sample bias in linked datasets
- Fazio et al. (2006) – short-term approaches:
  - Re-weighting
  - Conditioning variables
  - Banded regressions
- Our paper deals with the representativeness of single indicators in the process of linking different microdata sets in the short-run

Coverage over time – Number of firms per sample, for the source datasets (in thousands)

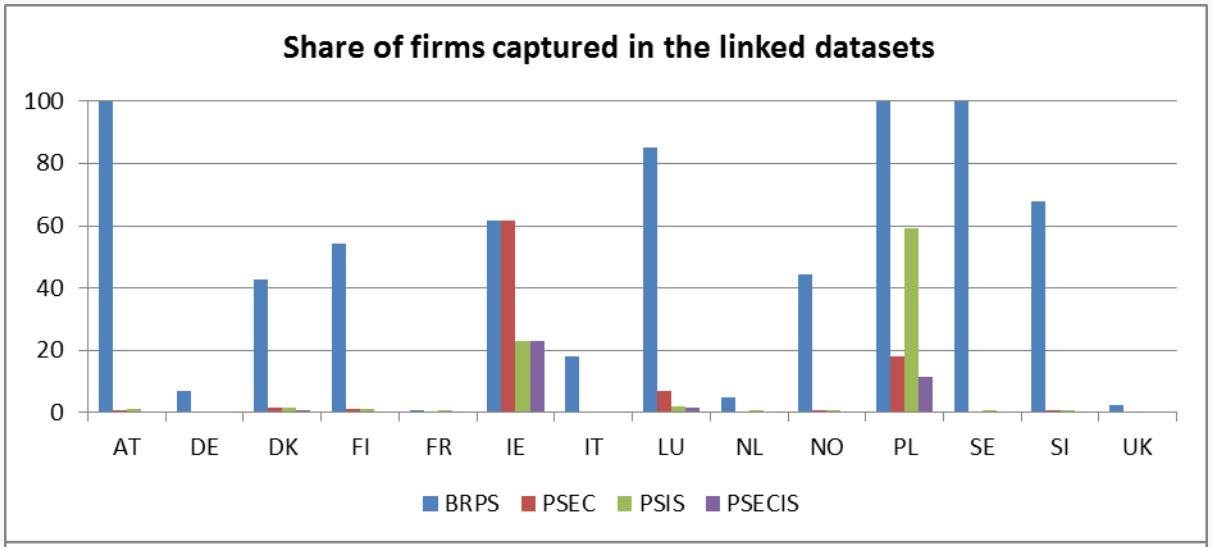


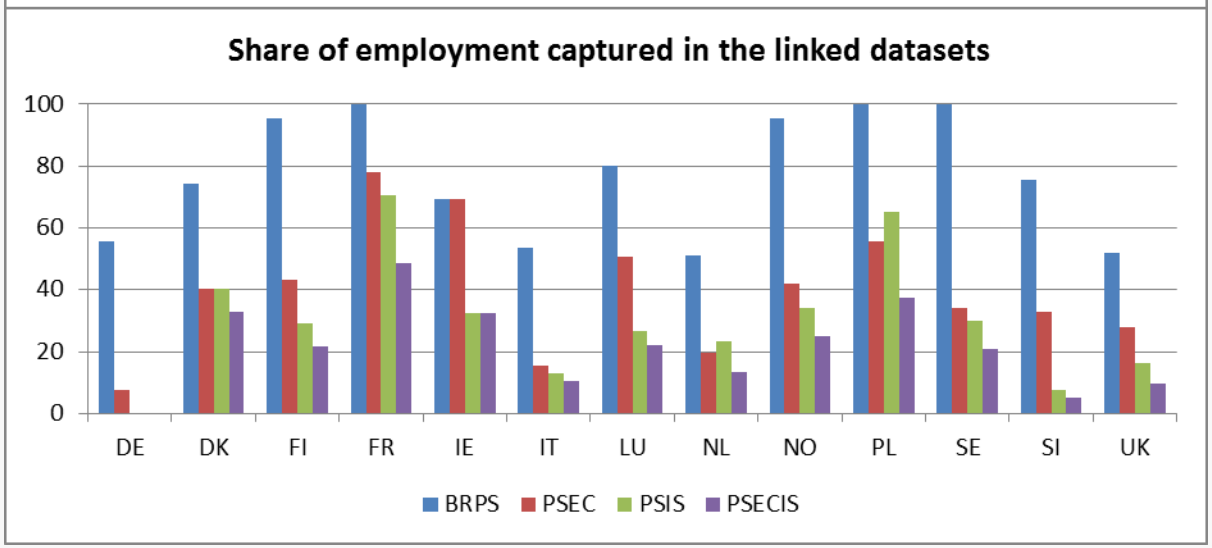






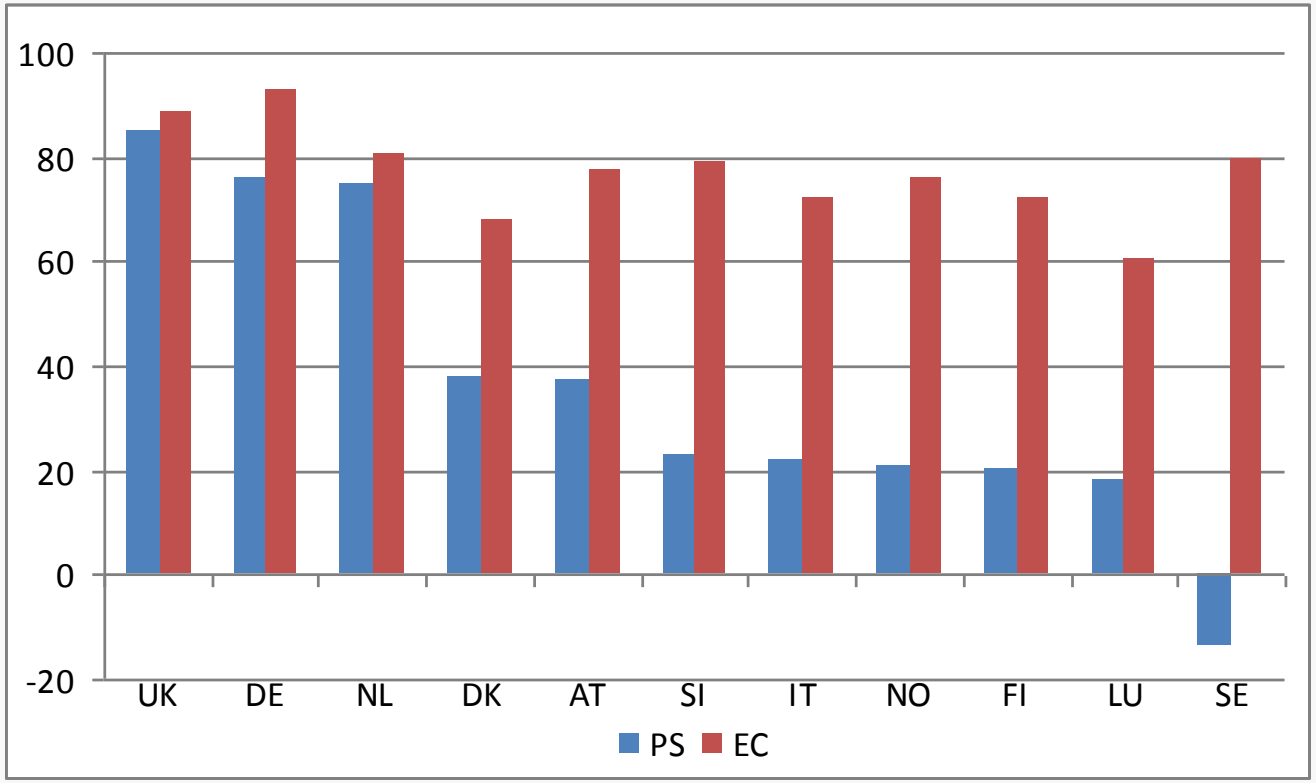
Coverage across samples, throughout the merging procedure (proportion of BR)





Attrition

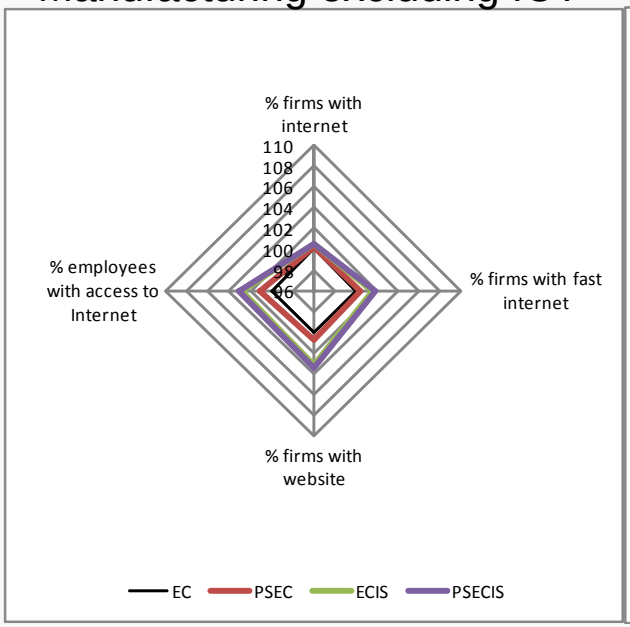
- Non-survival rates between 2003 and 2010 in the PS and EC



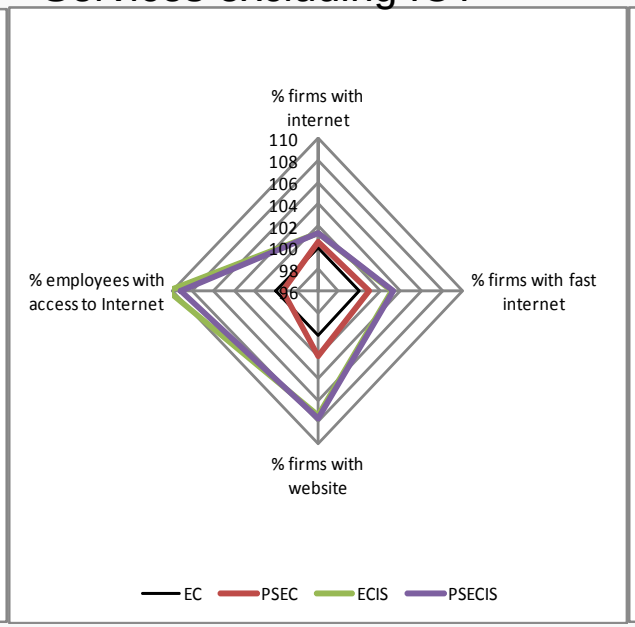
ICT indicators across samples

- Average ICT intensities in merged datasets, by industry across countries
- EC 2010=100

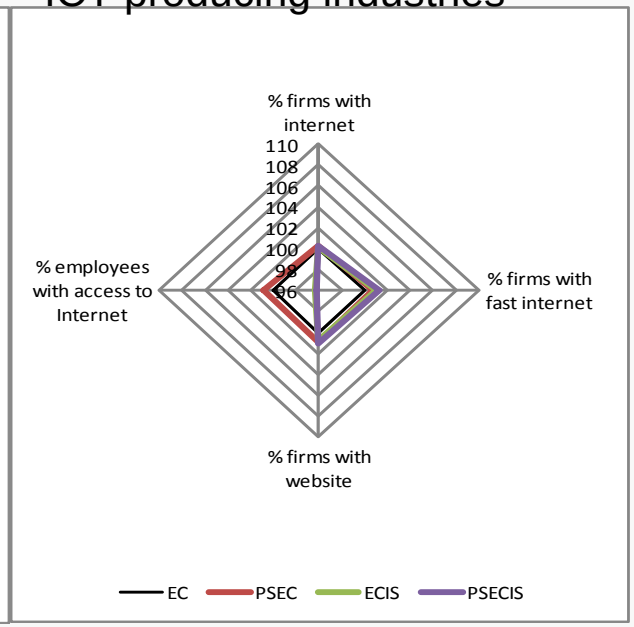
Manufacturing excluding ICT



Services excluding ICT



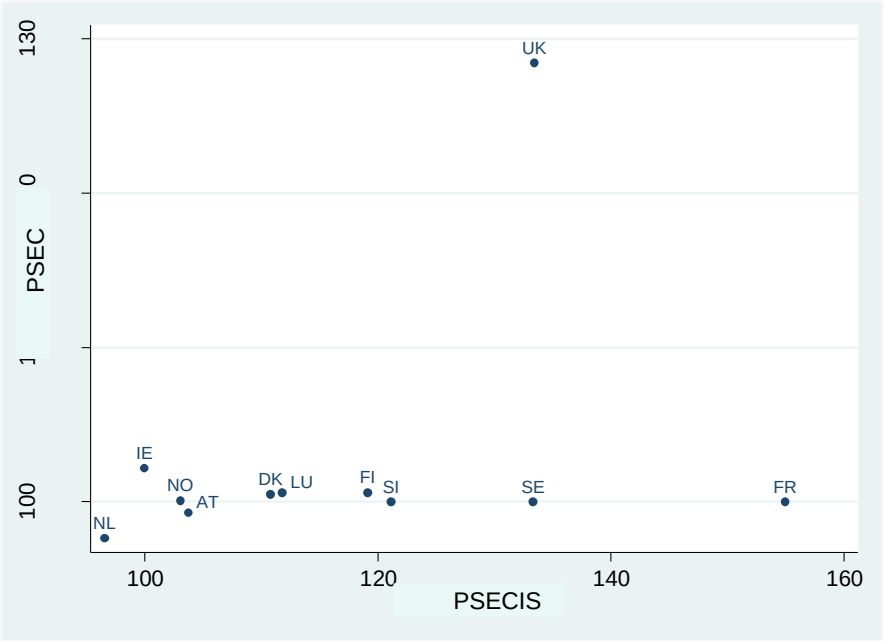
ICT producing industries



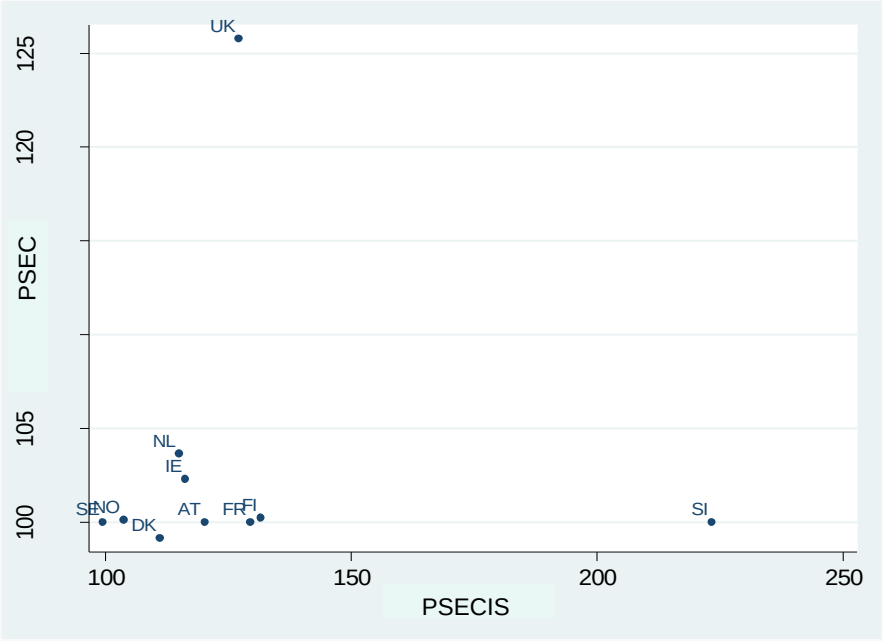
ICT indicators across samples

- Average ICT use in manufacturing (excluding ICT) across samples, as share of firms with e-sales

2006. Index, EC 2006=100



2010. Index, EC 2010=100



Ex-post control of selection bias

- Original weights become inappropriate after linking
- Reweighting – the variables become more representative of the underlying universe of firms
- Each descriptive Micro Moments dataset includes the aggregated average value of each variable as well as three different sets of re-weights for key variables:
  - First set: based on data available in the business register
  - Second set: constructed using firm size (measured as number of employees) at the sample and population levels
  - Third set: a combination of the business register and the firm size weights

Ex-post control of selection bias

- Comparison of mean values for employment for the PS and PSEC samples, by different reweighting approaches (in thousands):

Country	PS, BR reweighting	PSEC, no reweighting	PSEC, BR reweighting	PSECIS, no reweighting	PSECIS, BR reweighting
DK	1 010	509	1 070	207	413
IE	766	237	741	69	308
NO	1 110	468	1 140	221	635

- Comparison of mean values for AESELL for the PS and PSEC samples, by different reweighting approaches:

Country	EC, BR reweighting	PSEC, no reweighting	PSEC, BR reweighting	PSEC, empl. reweighting	PSEC, BR & empl. rewg.
FI	0.20	0.31	0.20	0.55	0.44
NO	0.27	0.34	0.27	0.46	0.40
SE	0.25	0.35	0.25	0.60	0.47

Industry-level analysis

- Consider the type of relationship examined when deciding which set of weights should be used (if any):
  - firm-level relationships – un-weighted variables can be used
  - macroeconomic relationships – employment-based weights seem the best at emphasizing the relevance of larger firms
- Comparison of reweighting schemes in pooled regressions:

Dependent variable: Labour productivity (appropriately weighted)

Reweighting scheme \ Sample	PS	PSEC	PSIS	PSECIS
HKpct, no reweighting	-0.18	0.12	0.24	0.13
(t-stat)	(1.51)	(0.84)	(1.57)	(0.84)
HKpct, BR reweighting	-0.46	-0.14	-0.39	-0.60
(t-stat)	(3.38)	(0.56)	(1.35)	(2.55)
HKpct, empl. reweighting	0.38	0.65	0.72	0.84
(t-stat)	(3.40)	(5.11)	(5.65)	(5.96)
HKpct, BR & empl. reweighting	0.28	0.39	0.35	2.55
(t-stat)	(2.55)	(3.06)	(2.45)	(3.28)

Firm-level analysis

- Firm-level regressions with ICT intensive human capital across samples

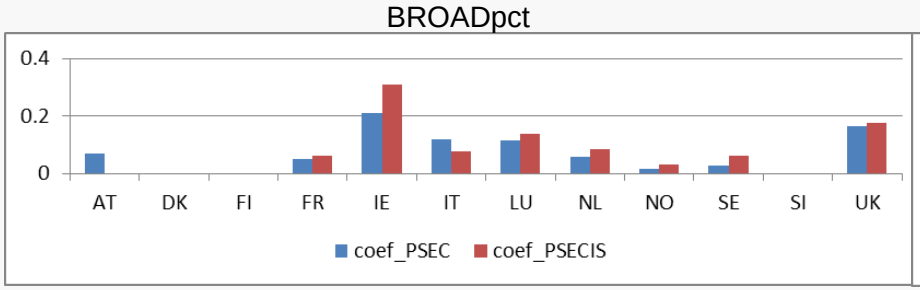
Dependent variable: (log) Labour productivity						
	Sample					
		PS			PSEC	
Country	FI	NO	SE	FI	NO	SE
HKITpct	0.260	0.178	0.135	0.280	0.307	0.318
t-stat	(43.31)	(30.70)	(26.46)	(13.00)	(5.17)	(8.06)
R-squared	0.884	0.751	0.602	0.879	0.899	0.808
Observations	171983	430460	551106	10651	3722	7344
BROADpct				0.045	0.041	0.101
t-stat				(4.57)	(2.54)	(8.32)
ECpct				0.015	0.016	-0.001
t-stat				(1.28)	(1.34)	(-3.26)

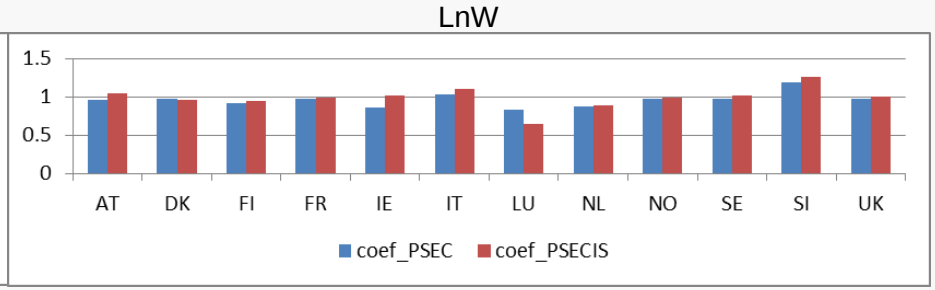
- Merging one smaller sample survey with a larger dataset or census does not seem to distort regression estimates, but may change them slightly

Firm-level analysis

- Firm level regressions with ICT intensity variable across samples

	Country	AT	DK	FI	FR	IE	IT	LU	NL	NO	SE	SI	UK
BROADpct	coef_PSEC	0.068	-0.001	0.011	0.049	0.209	0.117	0.113	0.058	0.015	0.028	-0.004	0.165
	coef_PSECIS	0.026	0.003	0.017	0.061	0.308	0.077	0.139	0.083	0.029	0.063	-0.004	0.175
LnW	coef_PSEC	0.964	0.969	0.915	0.974	0.862	1.024	0.83	0.872	0.978	0.973	1.185	0.979
	coef_PSECIS	1.049	0.96	0.942	0.989	1.021	1.095	0.646	0.885	0.99	1.013	1.253	1.006
R-squared	PSEC	0.92	0.93	0.93	0.95	0.83	0.89	0.79	0.92	0.94	0.94	0.91	0.87
	PSECIS	0.94	0.92	0.91	0.95	0.84	0.91	0.84	0.9	0.92	0.94	0.93	0.85





Conclusions

- Indicators become upward biased as more surveys are linked
- Specific values of ICT indicators appear less biased:
  - if the PS in a country is large or a census, if a sample co-ordination system is in use
  - for ICT and manufacturing firms
- Re-weighting can shift variable values from the smaller linked dataset closer to the larger dataset
- Inconclusive results for the use of re-weighting in industry-level regressions
- Firm-level estimations seem robust against selection bias (Fazio et al. (2006), Ritchie (2004))
- The major effect is a slightly higher estimate that does not significantly change the interpretation of results