Exploiting the integration of businesses micro-data sources

Giovanni Seri – Daniela Ichim – Valeria Mastrostefano – Alessandra Nurra

National Accounts and Business statistics Department

Italian National Statistical Institute (Istat)

{seri,ichim,mastrost,nurra}@istat.it

Poznan
September 8, 2015

Outline

1. Introduction
2. Data sources
3. Methods
4. Results
5. Conclusions and further works

Introduction

- An exhaustive archive covering the whole population defined by the SBS Regulation (FRAME) primary source for SBS statistics
- Sample survey data: Community Innovation Survey (CIS), Information and communication technologies Survey (ICT)
- Combining register and survey data to produce economic indicators exploiting the interaction between two data sources
- Value added (VA) per person employed (PE) for the subpopulations of enterprises defined by ICT or CIS indicators

Data sources

Frame SBS 2012 core variables (Turnover, Purchases of goods and services, Personnel costs, …)

Sample surveys: ICT and CIS

- Enterprises with at least ten persons employed
- Domains of interest: combination of economic activity (Nace), size class (Number of persons employed) and region (Nuts)
- Sampling design: one stage stratified random sampling
- Stratification according to the domains of interest
- Estimates through calibration methodology (Deville, Särndal, 1992; Casciano et al., 2006)
- Italian Statistical Business Register – ASIA
  - 2011 for ICT
  - 2012 for CIS

Data sources

FRAME theoretical target population :

ICT 196186 units (101,6% of ASIA 2011 target population)

CIS 160909 units (Frame do not include Financial services sector)

ICT linked dataset: 17667 units (around 93% of the sample dataset)

CIS linked dataset: 17760 units (around 99% of the sample dataset)

- changes occurred in the number of persons employed or in the NACE
- demographic events

Data sources

ICT indicators:

- downloading speed of Internet connection declared by businesses (**e_speed**), intensity of use of the network in terms of persons employed using Pc connected to the Internet for work reasons, dematerialization and integration of organizational processes, levels of maturity reached by the company in e-commerce (from those only buying on line to those firms selling and buying on line or having also own website offering opportunities to place on line orders for goods and services)

CIS indicators:

- Product innovation, Process innovation, Organizational innovation, Marketing innovation, R&D driven or not, Product or Process innovation (**PPI**)

Methods

Macro integration:

- Balancing
- Iterative proportional fitting: IPF (ICT, CIS)

Micro integration:

- Calibration (weighting):
  - Calibration applied to the linked file (ICT, CIS)
  - Calibration applied to the survey dataset (ICT)

Methods

A) IPF applied to bidimensional tables:
(VA/PE) X (NACE) X (ICT or CIS indicator)

(VA/PE) X (NACE) margin by FRAME;

Calibration:

Survey population totals: number of enterprises and number of persons employed in domains given by combinations of
(NACE) X (Size class) X (Region)

B) and C) population totals: number of enterprises, number of persons employed **and value added** in domains given by combinations of
(NACE) X (Size class)

D) population totals: number of enterprises, number of persons employed, **value added and e_speed** in domains given by combinations of
(NACE) X (Size class)

Results

Value added (VA) per person employed (PE) for ICT and non-ICT NACE and e_speed values: comparison of methods A, B, C and D

VA/PE	IPF (A)			Linked dataset (B)			Survey dataset (C)			Table (D)		
	e_speed			e_speed			e_speed			e_speed		
NACE	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT
Outside ICT sector	49082	62457	55065	48905	62976	55065	48529	63435	55065	50520	61363	55056
Inside ICT sector	52313	123658	104070	51801	122924	104070	50770	124932	104070	54442	125040	104265
Tot_e_speed	49168	67433	57600	48977	68006	57600	48588	68483	57600	50625	66725	57600

Results

Value added per person employed for Pavitt categories and values of PPI: comparison of methods A and B

VA/PE	IPF			Linked dataset		
	PPI			PPI		
PAVITT	0	1	Tot_CIS	0	1	Tot_CIS
Not elsewhere classified	66945	110452	81831	67515	112120	81831
High-technology	89509	88624	88837	90627	88231	88837
Medium-high-technology	54347	71570	67341	56533	70933	67341
Medium-low-technology	50065	61042	56180	50603	60703	56180
Low-technology	41953	61195	52800	43984	59747	52800
Knowledge-intensive services	64292	114504	95853	65103	115239	95853
Lessknowledge-intensive services	47237	58403	51877	47879	58302	51877
Tot_PPI	52489	73223	63332	53423	73000	63332

Results

Value added out Turnover (%) for ICT and non-ICT NACE and e_speed values: comparison of methods A, B, C and D

VA/PE	IPF (A)			Linked dataset (B)			Survey dataset (C)			Table (D)		
	e_speed			e_speed			e_speed			e_speed		
NACE	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT	0	1	Tot_ICT
Outside ICT sector	20,9	19,6	20,2	21,4	19,9	20,6	21,2	20,4	20,8	21,4	19,4	20,4
Inside ICT sector	30,9	43,5	41,1	31,1	43,6	41,4	30,5	41,2	39,3	28,9	42,9	39,9
Tot_e_speed	21,1	21,4	21,2	21,6	21,7	21,6	21,4	22,1	21,8	21,5	21,2	21,4

Conclusions and future works

- Macro level or micro level integration
- Calibration keeping into account different variables and/or domains
- Different calibration for different indicators
- Analysis of the results
- Statistical matching for the manufacturing sector
- Suggestions are welcome ☺