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

Abstract

Efficient usage of different data sources is becoming increasingly important in times of cost-savings for both producers of statistics and respondents to surveys. Typically, secondary usage by linking of microdata from different sources allows information to be presented in dimensions not earlier available, and is also highly demanded by the research society. However, survey designs seldom target multiple purposes, resulting in potential selection bias in linked datasets. In this paper, we investigate how the stepwise linking and aggregation of information from firm-level datasets (business registers, production, ICT usage and innovation surveys etcetera) in 14 European countries affect the representativeness and usefulness of indicators from the unique ESSLait Micro Moments Database. We illustrate the overlap issue both across samples and over time, for all countries. The matter of representativeness is addressed by exploring the advantages of using ex post reweighted variables in analyses. Although this might be considered a good short term solution, the first best solution in the long run would be larger samples or increased sample co-ordination. Another main finding is that both descriptive and marginal ICT usage indicators become upward biased when the overlaps get smaller. This disturbs the interpretation of marginal results to a lesser extent than descriptive comparisons.

*The authors would like to thank Eric Bartelsman and Michael Polder for valuable support and comments during the course of work.

**Corresponding author, Statistics Sweden, Box 24300, SE-10461 Stockholm, eva.hagsten@scb.se

1. Introduction

During the different phases of the ICT impacts projects (ESSLait being the most recent), a set of national firm-level linked as well as cross-country micro-aggregated (the Micro Moments Database, MMD) datasets were created and developed.[1] The national datasets now available in 14 European ESSLait project countries consist of information from business registers (BR), production surveys (PS), education registers, and trade statistics as well as information from the EU-harmonised surveys on ICT usage (EC) and innovation activities (IS) in firms.[2] Data dimensions and indicators not previously available were created by the linking procedure.

The MMD consists of a suite of tables created in the stepwise linking procedure, reflecting combinations of all the above mentioned sources, PSEC, PSIS, ECIS and eventually the smallest and most unique dataset PSECIS (all of them allow the BR to be included in the PS unless otherwise stated). General statistics are available for each dataset as well as certain moments such as correlations, quartile distributions and joint adoptions. Additionally, there is micro-aggregated information on firm demographics and dynamics.[3] The sources are described in more detail by Denisova [5]. In order to facilitate comparisons across countries and over time, the underlying data have been converted to one industry code: the EUKLEMS NACE1 standard or alternative industry hierarchy.[4]

A firm-level linked dataset typically inherits not only the underlying biases in the original source data and their measurement errors, but it also risks new distortions through the linking process. Measurement errors cannot be dealt with in this context, since they relate to the data collection techniques that are generally beyond the purpose of the ESSLait project. However, they can still affect analyses of linked datasets.

As long as the data linking involves only administrative sources covering the whole population, errors can certainly appear but not in the guise of selection bias. However, when one or more sample surveys are introduced, specific attention needs to be paid. The sampling strategies used by most statistical offices are usually aimed at creating accurate macro

[1]Eurostat Grant agreements 49102.2005.017-2006.128 (ICT Impacts 2006-08), 50701.2010.001-2010.578 (ESSLimit 2010-12) and 50721.2013.001-2013.082 (ESSLait 2013).
[2]The following countries are members of the ESSLait project: Austria, Denmark, Finland, France, Germany, Ireland, Italy, Luxembourg, the Netherlands, Norway, Poland, Sweden, Slovenia and the United Kingdom.
[3]A more exhaustive description of the output datasets can be found in Bartelsman et al (2013).
[4]See www.euklems.net.

estimates, meaning that the focus is on recovery of most of the value added rather than the majority of firms. This implies a sample design where large firms are given a higher weight, and often firms above a certain threshold with respect to the number of employees are always sampled. Fortunately, and if need be, measures can be taken to control for the biases in the short term and in the long run.

Ideally, the long-term solution would be to collect all data in a way that simply hampers the appearance of any kind of selection bias by using censuses or administrative registers. While this is not a particularly realistic approach due to administrative costs and response burden issues, there are simple steps that can be taken to improve the situation. Within the frame of the Eurostat ESSLimit Project, Denisova [6] proposed an increased awareness of secondary data usages such as microdata research in the sampling design. (This is also one of the intentions of the funding body of the present and earlier project, the Eurostat MEETS programme.) A sample co-ordinating system prioritising multi-purpose datasets was found to be a good strategy to improve representativeness in linked microdata sets. In practice, this means that the system, besides providing a good macro estimate and reduction of the response burden of firms, acknowledges the demand for research purposes. Meanwhile, as such underlying changes most likely will take time to implement, this paper provides some guidance on the quality of the micro-linked and micro-aggregated ESSLait datasets together with suggestions of when ex-post measures might improve their representativeness and when they might be superfluous. Thus, this paper will mainly focus on a description of what happens to the representativeness of single indicators in the process of linking the different microdata sets and how this can be dealt with in the short run.

2. Statistical properties of linked datasets

Although a wide range of studies use merged data to examine firm behaviour, the literature on statistical properties of linked datasets is limited. According to Bartelsman and Doms [3], who reviewed papers that use longitudinal microdata to examine productivity growth patterns, most studies did not investigate linked data quality thoroughly because initial linking attempts of production statistics usually contained a large number of observations and therefore produced estimates with lower standard errors.

The particularities of the population, the sampling procedure, response rates, the linking process and correlation between variables make it difficult to find a universal solution to dealing with the microdata quality issues that arise in different cases. In the absence of an established framework for analysing output data quality, researchers apply various best practice methods when inspecting linked data for potential sample bias and measurement errors, although in-depth quality checks are usually omitted. An important task is to document the following:

- survey design,
- non-response patterns and measurement error for the source surveys,
- any treatments applied to the data (such as imputation) and
- the linking procedure itself, in order to identify the issues relevant for the research question.

However, such measures in assessing the quality of survey microdata certainly imply higher costs for statistical agencies.

Chesher and Nesheim [4] review the literature on the statistical properties of linked business datasets and investigate data quality issues that emerge when linking administrative data sources with sample surveys. In a paper on UK business data linking, Ritchie [12] also lists problems that arise when linking micro datasets and provides suggestions for overcoming them. Using ONS datasets, Fazio et al. [10] discuss sample representativeness and several approaches to dealing with selection bias. The authors exemplify by linking a production survey with the E-Commerce survey. However, unlike most other countries in our study, the ONS production survey does not target the whole population of firms, nor is it fully synchronised with other sample surveys, implying that the linking of the PS and EC surveys results in a far smaller overlap than the EC sample, making it impossible to generalise the conclusions for other countries.

The studies noted in the previous paragraph identify a series of issues related to input data quality that are beyond the scope of this paper. Output data may be affected by bias from several sources. For instance, in the case of surveys that include optional questions, such as the E-Commerce survey, topic saliency may influence response patterns, as firms for which ICT usage is of high importance are more inclined to respond, which generates an upward

bias in the estimates of ICT usage. In addition, there may be low response rates among unproductive firms, which could for instance lead to overestimated coefficients when assessing the impact of ICT on productivity [10].

Longitudinal data quality suffers due to corporate restructuring activities that lead NSIs to change the identifier of an existing firm. Although statistical agencies are often cautious with registering such statistical events, this makes it difficult in general to follow the evolution of the firm throughout the entire period. Small overlaps between survey samples, discontinuity in sampling of small firms, as well as changes in existing firms' identifiers may impede building representative panel data sets. Bartelsman and Doms [3] propose the use of aggregate measures, which may alleviate such noise in input and output microdata and describe firm dynamics more accurately.

In a number of countries, most surveys (including the CIS and E-Commerce) are designed to avoid high administrative burdens for small firms by preventing repeated sampling of the same firm. Unfortunately, the surveys supply only a limited amount of information regarding small firms as a result of such design features and concerns arise about the validity of inferences. Limiting the analysis only to large firms which are selected each year does not allow the generalisation of inferences outside the selected sample without a sample selection model – something that can be difficult to achieve for linked datasets [4]. Despite survey designs focusing on larger firms that are considered significant innovators, low response rates among large firms in voluntary surveys, such as the CIS, may bias the results towards underestimating the level of firm innovation [10], [12].

Fazio et al. [10] propose three approaches to deal with sample bias in linked datasets: reweighting, banded regressions and conditioning variables. The reweighting method consists of assigning a weight to each firm in the sample, based on the same characteristics of the sampled firms and of the firms in the NSI business registers. According to Eurostat [7], employment-based weights should be used for variables related to broadband access and usage, while a set of weights based on turnover is more appropriate for sales/purchase-related variables. Reweighting may be necessary, as the design weights of the source surveys might not reliably reflect the composition of the linked subsample. As Fazio et al. [10] caution, results may be influenced differently by the applied weighting scheme. Furthermore, the authors state that weighting is a powerful tool for eliminating bias in simple statistics such as

means or tabulations, whilst composite statistics affected by variable correlation may remain approximately unbiased even when not reweighted.

However, the use of weights is considered a sensitive issue. According to Chesher and Nesheim [4], weighted analysis is preferred for drawing sample inferences, while unweighted analysis should be performed only after testing the assumptions that justify it. If unweighted analysis is appropriate, the resulting estimates should be similar to those from weighted analysis. Fazio et al. [10] also signal a potential problem with weighting – it seems to have larger effects on variables with measurement error.

Conditioning variables – discrete variables with a small number of values or continuous variables split into several subsets – are more widely used in econometric studies than weighting techniques because they have an economic interpretation, making them more intuitive to use. Including dummy variables in model estimations not only increases the efficiency and robustness of estimates, it also reflects the weight of each stratum in the sample. The band regressions method involves running separate regressions for each stratum of the sample. The method is therefore more cumbersome to apply, as each band requires a different model according to the characteristics of firms in the respective bands.

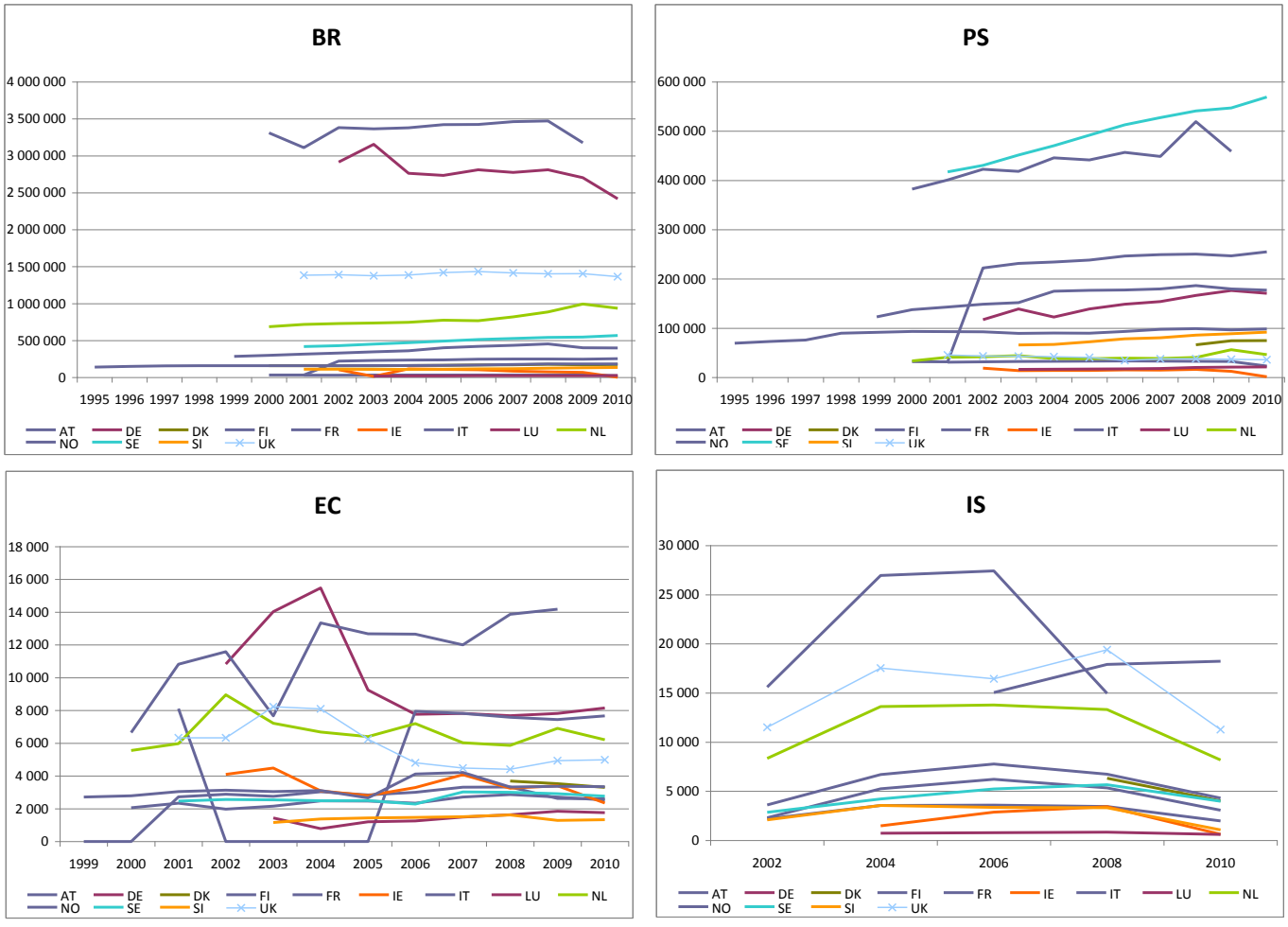
Unlike weighting, conditioning variables and banded regressions can be used without specifying the exact sample proportions when the dummies correspond to each subsample. When conditioning variables are used (as dummies in regressions), their impact should be assessed both individually and combined. Combining conditioning variables (such as size dummies) and weighting are useful when the model contains variables not interacting with the dummies, for which weighting diminishes the bias. Fazio et al. [10] recommend the use of conditioning variables for marginal analysis rather than the construction of weights.

3. Firms coverage across samples and over time

To illustrate the relevance of the sample overlap issue, we provide information about the coverage of firms and employment in various project datasets. Figure 1 shows the number of firms per country in each of the years with available information, for the four initial datasets: BR, PS, EC and IS. Important differences can be observed between countries at the BR level – Italy and Germany have the largest number of firms by far every year, followed by the

United Kingdom and the Netherlands, while Luxembourg has the fewest firms. The ranking of the countries is not the same with respect to the number of firms in the PS, where Sweden, Italy, Austria and Norway have the highest number of firms. For most countries, both the BR and PS tend to include a similar or larger number of firms from one year to the next. In contrast, response rates for the EC are more volatile over time in many of the countries. France, for instance, did not participate in the EC survey during 2002 to 2005. The IS has mainly been conducted every second year up to 2010, with different start years across countries. The intermediate years have been imputed in the ESSLait project for analytical purposes, although Figure 1 only illustrates the actual survey years.

Figure 1. Coverage over time - Number of firms per sample, for the source datasets



Source: ESSLait Micro Moments Database

Table 1 provides details about firm coverage across several samples throughout the linking procedure for the latest available year, 2010, or in the case of Italy, 2009. The first column shows the number of firms included in the BR of each country; the following four columns

display the share of BR firms present in several linked datasets (BRPS, PSEC, PSIS, PSECIS); and the last three columns reveal the share of firms in the BRPS sample that also appear in the PSEC, PSIS and PSECIS linked samples respectively. The empty cells are due to missing information about the number of firms in the PSIS and PSECIS in Germany. This simply follows from the fact that the German innovation survey has a legal status that does not allow linking to other surveys. The PS contains all firms in the BR only in three countries (Austria Poland and Sweden), leading to a 100 per cent coverage rate in the BRPS. In contrast, the PSEC, PSIS and PSECIS samples include at most 2 per cent of the firms in the BR for most countries, except for France and Ireland, where coverage rates are higher. The coverage is higher when considering rates as the share of firms in the BRPS also present in the PSEC, PSIS and PSECIS samples respectively.

Table 1. Coverage across samples, throughout the merging procedure - Share of firms captured in the linked datasets (population = BR)

		Percentage of BR				Percentage of BRPS		
Country	BR	BRPS	PSEC	PSIS	PSECIS	PSEC	PSIS	PSECIS
AT	254 962	100.0	1.0	1.2	0.2	1.0	1.2	0.2
DE	2 417 983	7.1	0.1			2.1		
DK	175 542	42.6	1.6	1.8	0.6	3.8	4.3	1.5
FI	182 009	54.1	1.4	1.1	0.4	2.6	2.0	0.7
FR*	23 253	0.8	0.3	0.6	0.1	33.0	78.4	11.4
IE	2 345	61.8	61.8	22.9	22.9	100.0	37.0	37.0
IT	4 577 277	18.0	0.3	0.3	0.2	2.0	2.0	0.9
LU	25 071	85.3	7.0	2.1	1.7	8.2	2.5	2.0
NL	937 362	5.0	0.5	0.7	0.2	10.1	13.7	4.9
NO	398 577	44.5	0.8	1.0	0.4	1.8	2.3	0.8
PL	56 958	100.0	18.0	59.2	11.7	18.0	59.2	11.7
SE	569 478	100.0	0.5	0.7	0.3	0.5	0.7	0.3
SI	136 041	67.7	1.0	0.6	0.1	1.4	0.9	0.1
UK	1 366 044	2.6	0.2	0.2	0.1	6.2	9.0	2.7

Note: Table 1 is based on 2010 figures for all countries except Italy, where 2009 is the latest available year. Italian data refer to 2008. BR for France is constructed, and should be representative, but it is not the universe of firms as in the other countries.

Source: ESSLait Micro Moments Database

Table 2 refers to employment coverage in 2010 and is organised similarly to Table 1. The empty cells represent missing employment data for the PSIS and PSECIS in Germany. For most countries, the coverage rate in the linked datasets is higher for employment than for the number of firms, suggesting that large firms are better represented in the linked samples than smaller firms. As shown in Table 1, employment coverage is higher when the share of firms in each linked sample is compared to the BRPS rather than the BR dataset.

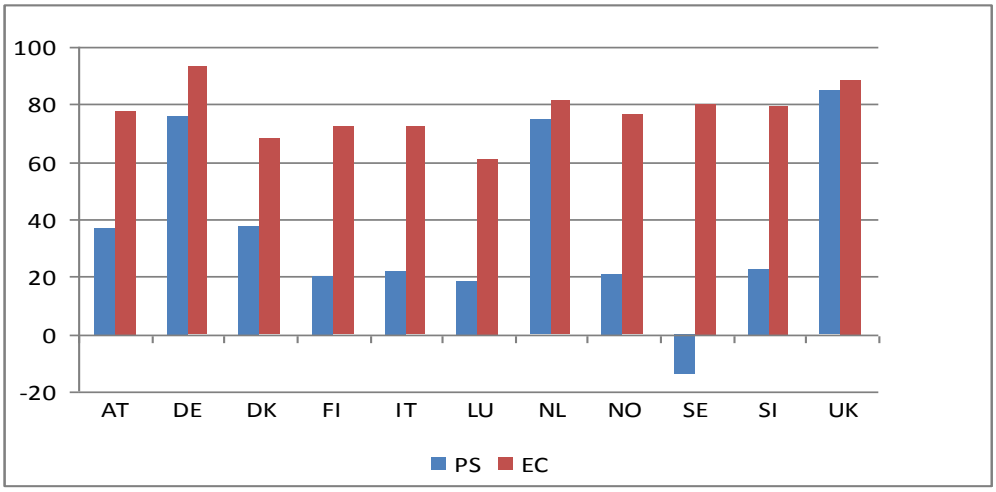
Table 2. Coverage across samples, throughout the merging procedure - Share of employment captured in the linked datasets (population = BR)

	Percentage out of BR					Percentage out of BRPS		
	BR	BRPS	PSEC	PSIS	PSECIS	PSEC	PSIS	PSECIS
DE	20 045 393	55.5	7.8			14.1		
DK	1 130 841	74.3	40.2	40.5	33.0	54.1	54.5	44.4
FI	1 113 088	95.6	43.3	29.2	21.9	45.2	30.5	22.9
FR	4 373 323	100.1	78.1	70.6	48.6	78.0	70.5	48.5
IE	218 781	69.2	69.2	32.5	32.5	100.0	47.0	47.0
IT	17 978 352	53.7	15.6	13.1	10.6	29.0	24.5	19.6
LU	226 068	79.9	50.6	26.8	22.0	63.4	33.6	27.5
NL	4 548 025	50.9	19.5	23.3	13.4	38.3	45.7	26.3
NO	1 218 653	95.4	41.8	34.2	25.0	43.8	35.9	26.2
PL	4 108 381	100.0	55.8	65.2	37.6	55.8	65.2	37.6
SE	1 886 255	100.0	34.1	30.1	21.0	34.1	30.1	21.0
SI	591 644	75.4	32.8	7.8	5.1	43.5	10.4	6.8
UK	15 422 078	52.0	27.8	16.5	9.7	53.6	31.8	18.7

Note: Table 2 is based on 2010 figures for all countries except for Italy, where 2008 is the latest available year. No information about employment coverage in Austria is available.

Source: ESSLait Micro Moments Database

Figure 2. Non-survival rates between 2003 and 2010 in the PS and EC



Note: France, Ireland and Poland are not included due to missing information about turnover rates in the EC and PS, respectively.

Source: ESSLait Micro Moments Database

There is a natural turnover of firms over time, new ones enter and old ones exit. However, survey sampling design also causes attrition – loss of firms during the sample period – as the probability of the same firm being sampled in consecutive survey waves declines with firm size. Turnover rates during the period 2003-2010 for the PS and EC surveys are depicted in

Figure 2. The bars represent the share of firms that disappear from each of the two datasets during this interval.

Non-survival rates in the PS sample are particularly high for Germany, the Netherlands and the United Kingdom, while Sweden is the only country where the share of same firms being sampled in consecutive surveys is increasing. The EC survey has much larger attrition rates for most countries. However, while attrition in the EC survey is strongly affected by sampling bias, non-survival in the PS in most countries is due mainly to firm exits (and entries), since the PS is based more often on administrative data or large samples. Table 3 provides another perspective on attrition – it classifies countries into four categories according to the size of firm turnover with respect to cross-country average values for the PS and EC samples respectively. Countries are included in the category "Low" when the turnover rate is more than one standard deviation below the mean value; in "Low-medium" when the rate is one standard deviation or less below the mean; in "Medium-high" when the rate is one standard deviation or less above the mean; and in "High" when the rate is more than one standard deviation above the mean. One-half of the countries fall within the low-medium intensity category, and only Sweden lies under the low intensity category when it comes to the PS. The countries are more evenly distributed across the four categories with respect to the EC survey.

Table 3. Classification of countries according to attrition intensity

Category / Country	Low	Low-medium	Medium-high	High
AT		PS, EC		
DE				PS, EC
DK	EC	PS		
FI		PS, EC		
FR			PS	
IE				EC
IT		PS, EC		
LU	EC	PS		
NL			EC	PS
NO		PS, EC		
SE	PS		EC	
SI		PS	EC	
UK				PS, EC

Source: ESSLait Micro Moments Database

The kind of analysis performed is affected by the type and severity of the overlap issue – a small overlap in the same survey in time will make longitudinal analyses more difficult, while

a small overlap across surveys in single years will pose problems for cross-sectional analyses. Given that the samples are reasonably representative, high attrition is not so problematic for the micro-aggregated dataset (MMD) as for the country-level micro datasets. The high turnover in the EC survey adds to the natural turnover in the PS (provided by firm entries and exits), making it difficult to use balanced panels and to perform advanced modelling or certain robustness tests.

4. ICT indicators across samples

When ICT indicators are used across various datasets or built by merging them, the differences in coverage may result in issues with sample representativeness, mainly because the linking reduces the datasets, but also because the design of each original survey varies. In this section, we illustrate these effects by following a set of ICT indicators through the linking process. The comparisons are made across industries, time and countries.

We start by showing how the average values of selected ICT usage indicators vary across the linked ESSLait datasets, beginning with the E-Commerce Survey and continuing with the PSEC (e-commerce and production), ECIS (e-commerce and innovations) and PSECIS (e-commerce, production and innovations). Each ICT usage indicator is illustrated as an index, where the base value of 100 refers to the mean of the EC survey value.[5] Looking at a value of 103 for the PSECIS sample, for instance, indicates that the average ICT use is 3 per cent higher than the average value in the single EC sample. Thus, the average ICT use of firms in the smallest of the merged datasets slightly overestimates the ICT usage. The closer the average value is to the base value, the smaller the effect of the selection bias.

Chart 1 brings together the representation of four common ICT usage indicators: the proportion of firms with internet (IACC), with fast internet (BROAD), and with a website (WEB) as well as the proportion of employees with broadband internet access (EMPIUSEPCT) for the manufacturing and services firms where the ICT producing industries have been singled out as a group. The values refer to the year 2010 and represent averages across countries for which the data are available.

[5] Carefully note that this value may still deviate somewhat from official statistics from each country because a data-linking project cannot use standard grossing-up procedures.

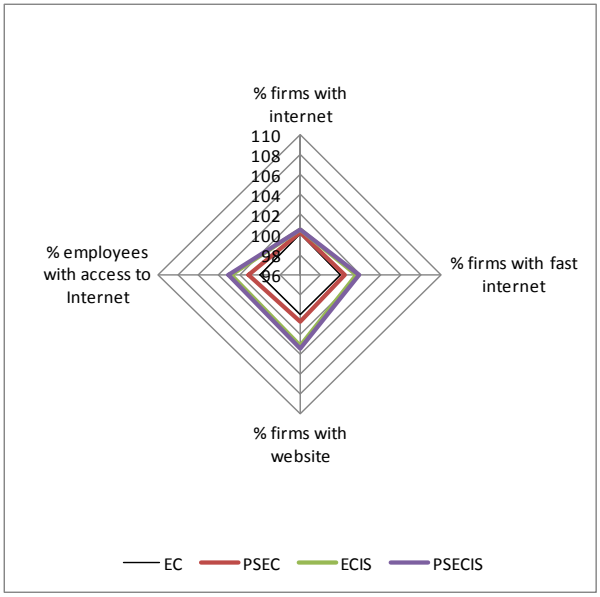
With a high degree of overlap between different surveys, there should ideally not be any slide in variable values when different datasets are merged. However, in reality several things affect this: the cost of surveys, response burden of firms, and last but not least important and emphasised by Denisova [6], the underlying structure of national statistics with their overall primary targets to produce representative macro estimates. In the merged dataset including innovation data, the average values for two of the ICT indicators (proportion of firms with website and proportion of employees with access to internet) are clearly higher than the base value of the EC sample, especially in the case of services. Moreover, in the case of the services firms, the variation among countries is also larger, which is illustrated in Table 4a, where the standard deviation is presented for the index pertaining to the PSECIS sample. We can conclude that in the case of the services sector, the sample coordination between innovation survey and other surveys is less pronounced and the sample coordination practices are more divergent across countries. Some specifics can also be discerned when it comes to ICT producing industries. Average ICT use seems to be quite similar across the samples and even underestimated in the case of the indicator capturing the proportion of employees with access to internet (the value of the index for the PSECIS sample is 96.2).

Another finding worth emphasising is that the less saturated the ICT indicator, the greater the difference in average ICT use across samples. Saturation is here defined as close to the full usage, that is 100 per cent. The first two ICT indicators exhibit very high levels of saturation (nearly all firms have internet and most of them with high speed), as can be seen in Table 4a. Typically, this leads to only minor differences across samples for the two ICT indicators. However, the deviation is apparently larger when the particular ICT tool in question is not yet fully widespread, as in the case of the proportion of employees with broadband access to internet. This becomes even more pronounced when we consider another set of ICT usage indicators presented in Table 4b: proportion of employees with access to fast internet (BROADpct), proportion of firms with mobile internet access (MOB), and proportion of firms with e-purchases (AEBUY) and e-sales (AESELL).

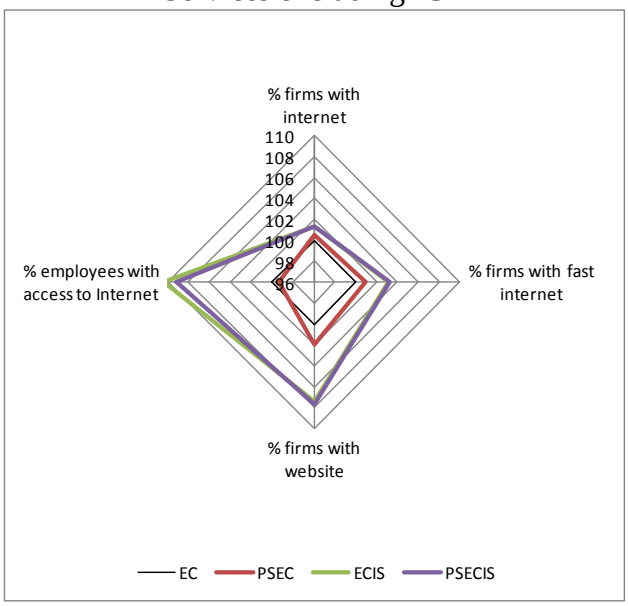
Chart 1. Average ICT intensities in merged datasets, by industry across countries

EC 2010=100

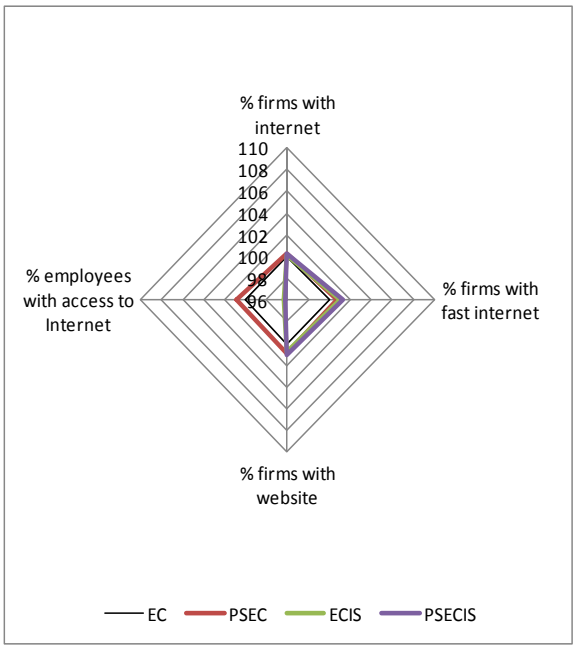
Manufacturing excluding ICT



Services excluding ICT



ICT producing industries



Source: ESSLait Micro Moments Database

Table 4a. Average ICT intensities by industry

EC 2010=100

		% firms with internet	% firms with fast internet	% firms with website	% empl. with access to Internet
Manufacturing, excl. ICT	Mean (absolute value)	0.99	0.94	0.88	0.44
	PSECIS_Index (100=EC)	100.5	101.8	103.5	103.0
	Std (PSECIS_Index)	0.4	1.4	3.6	2.8
Services, excl. ICT	Mean (absolute value)	0.98	0.93	0.85	0.61
	PSECIS_Index (100=EC)	101.4	103.3	107.7	109.2
	Std (PSECIS_Index)	1.6	2.5	7.5	13.4
ICT producing industries	Mean (absolute value)	0.99	0.96	0.92	0.63
	PSECIS_Index (100=EC)	100.3	101.3	101.0	96.2
	Std (PSECIS_Index)	0.8	1.8	3.2	9.6

Source: ESSLait Micro Moments Database

Table 4b. Average ICT intensities by industry

EC 2010=100

		% emp. with access to fast internet	% firms with mobile internet access	% firms with e-purchases	% firms with e-sales
Manufacturing, excl. ICT	Mean (absolute value)	0.43	0.34	0.51	0.33
	PSECIS_Index (100=EC)	103.9	110.6	112.8	127.7
	Std (PSECIS_Index)	2.4	5.6	20.7	35.2
Services, excl. ICT	Mean (absolute value)	0.59	0.46	0.55	0.31
	PSECIS_Index (100=EC)	109.9	124.2	112.7	132.7
	Std (PSECIS_Index)	13.5	26.9	6.9	24.2
ICT producing industries	Mean (absolute value)	0.61	0.51	0.64	0.34
	PSECIS_Index (100=EC)	96.6	99.7	98.7	120.6
	Std (PSECIS_Index)	9.5	10.7	12.5	27.3

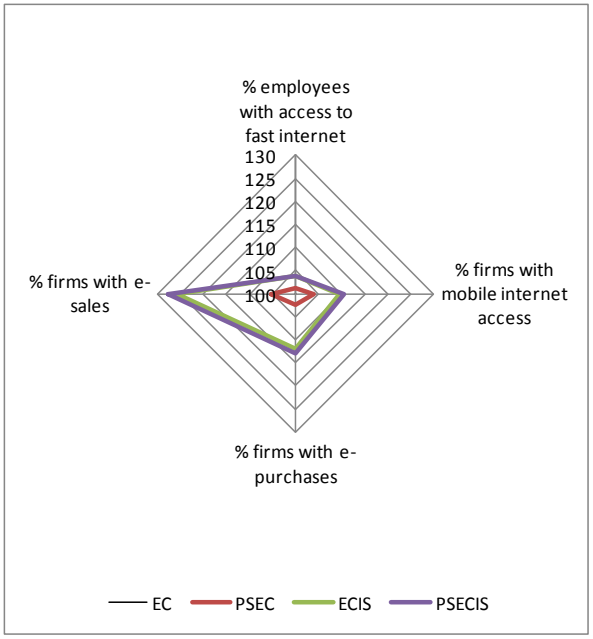
Source: ESSLait Micro Moments Database

The average ICT usage measured by this second batch of variables is much lower and ranges from 31 to a maximum of 64 per cent of firms, confirming a level far below saturation. For these variables, we can observe much larger differences in average values across the datasets and higher standard deviation of the index (see also Chart 2 for a graphical representation). The ICT producing industries again exhibit a specific situation where the average ICT use turns out to be underestimated in the PSECIS sample (with the exception of the proportion of firms with e-sales). One possible explanation could be that additional efforts are made by the statistical offices to target relevant ICT producing firms for their EC samples, but no similar efforts are made when devising a sample for innovation surveys.

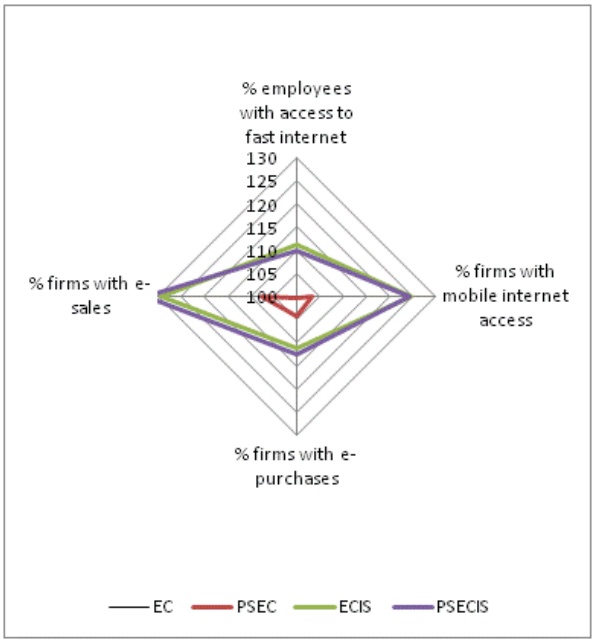
Chart 2. Average ICT intensities in merged datasets, by industry across countries

EC 2010=100

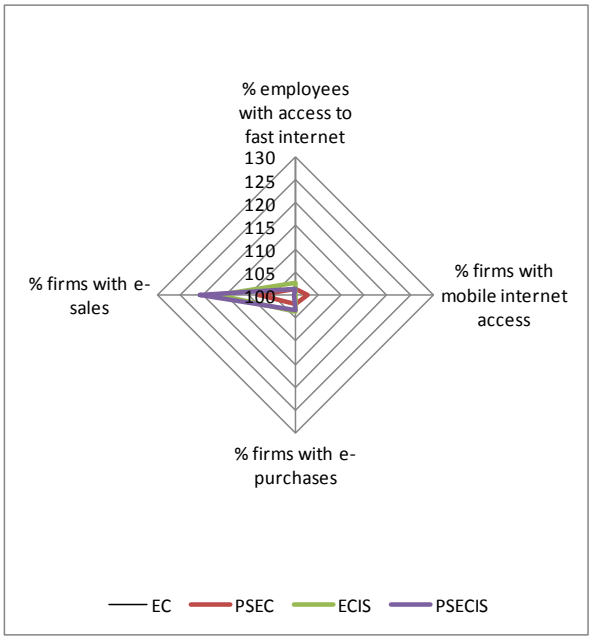
Manufacturing excluding ICT



Services excluding ICT



ICT producing industries

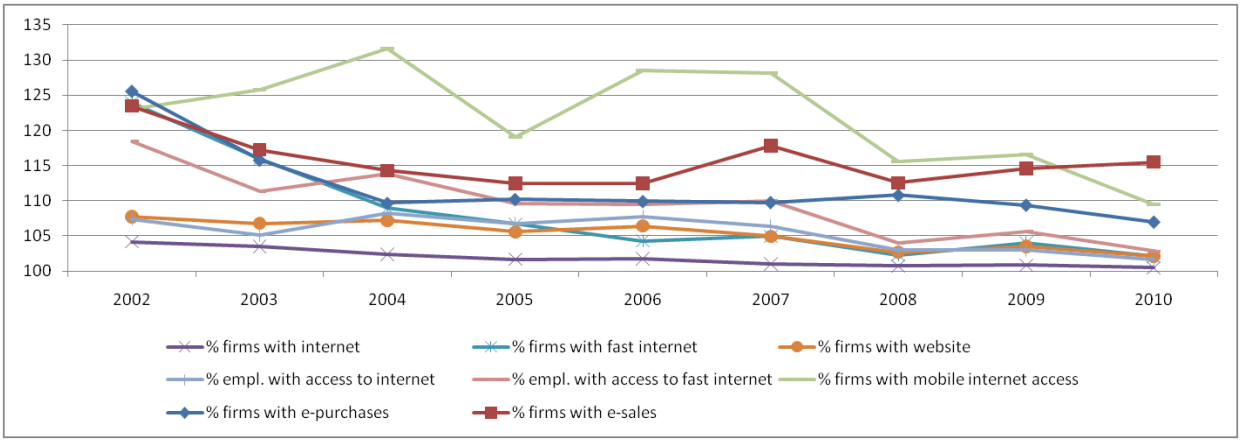


Source: ESSLait Micro Moments Database

In Figures 3 and 4, we show the changes in the index for the PSECIS sample over time for manufacturing and services firms. The index is slightly falling for most variables, which is consistent with the conclusion that the saturation of the ICT indicator plays a role (the saturation is increasing over time), and that in some countries the awareness of sample co-

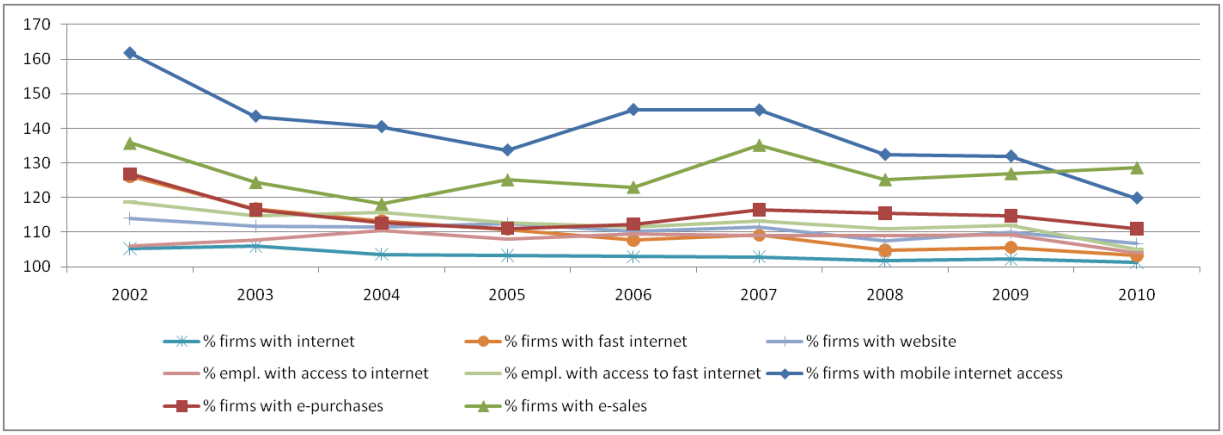
ordination has increased. The figures also reveal that this is a more marked matter for services firms.

Figure 3. Average ICT intensities in the PSECIS sample, Manufacturing excluding ICT



Note: Averages are calculated for a set of 11 countries for which the data for all years in the 2002-2010 period were available. Slovenia, Poland and Italy are excluded.
Source: ESSLait Micro Moments Database

Figure 4. Average ICT intensities in the PSECIS sample, Services excluding ICT



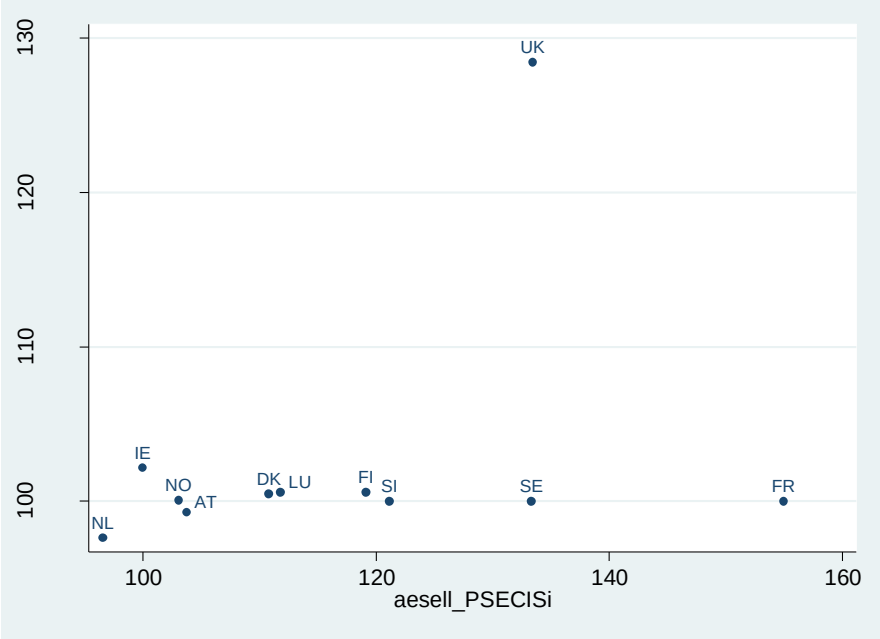
Note: Averages are calculated for a set of 11 countries for which the data for all years in the 2002-2010 period were available. Slovenia, Poland and Italy are excluded.
Source: ESSLait Micro Moments Database

By averaging the indexes in this way, the large variation across countries following sampling strategies and levels of saturation remains hidden. Figure 5a shows each country according to its average ICT use in manufacturing in 2006, measured by the proportion of firms with e-sales in the respective PSEC and PSECIS samples (EC 2006=100). The United Kingdom value apparently slides away much more than for the other countries, already in the early linking process. This pattern may appear when the production statistics are sample-based and not fully co-ordinated with other surveys. For the remaining countries, the differences between the samples become evident only when the smallest of the merged datasets is

investigated (PSECIS). Now the value of the index ranges from less than 100 to more than 150. Comparing this with the data for 2010 (Figure 5b), we can see that the positions of some countries have changed. The most notable examples are Sweden, where an increased positive sample coordination between the ICT usage and innovation surveys from 2008 onwards has resulted in vastly reduced biases, with a value of the index that is close to 100; and Slovenia, where the reverse process has taken place in line with efforts to reduce the response burden, which has increased the value of the index for manufacturing firms well above 200.

Figure 5a. Average ICT use in manufacturing (excluding ICT) across samples, 2006

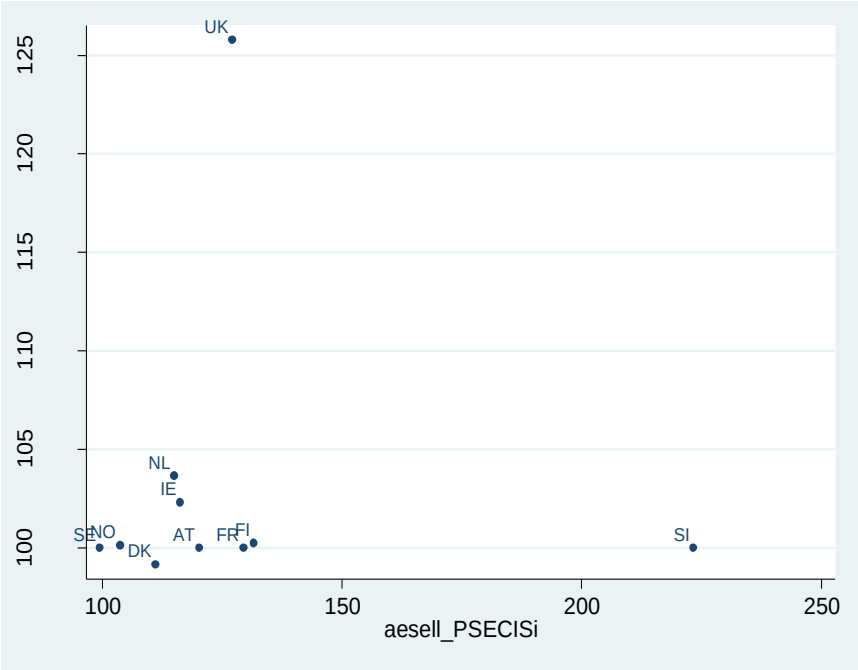
Index, EC 2006=100



Source: ESSLait Micro Moments Database

Figure 5b. Average ICT use in manufacturing (excluding ICT) across samples, 2010

Index, EC 2010=100



Source: ESSLait Micro Moments Database

To conclude, differences in sample overlaps tend to cause an overestimation of ICT use values in the linked datasets, especially if non-saturated indicators are considered. Acknowledging the underlying sampling strategies across countries, the question arises how this could be dealt with in international comparisons. One way to address this problem is to use reweighted values of the indicators, to which we turn in the next section.

5. Ex-post control of selection bias

We address the issue of sample bias in the linked datasets by reweighting the variables to make the linked samples more representative of the underlying universe of firms. Statistics offices apply similar practices when dealing with a single dataset, but original weights become inappropriate after linking several data sources with different sampling designs. Therefore, each descriptive Micro Moments dataset includes not only the aggregated value of each variable but also weighted aggregates. The first set of weights does sample reweighting based on the ratio of the number of firms in the population (from the Business Register) relative to the number of firms in the sample, in each size group and industry. This method inflates the sample, so that the weighted sum of firms in the sample equals the number of firms in the population. The second set of weights is used for variables whose aggregate value

is an average.[6] These are weighted using firm size (measured as number of employees so that weighted aggregate firm size equals the total industry employment divided by total number of firms). Finally, the third set of weights applies both the business register sample reweights and the firm size weights. A more detailed description of the reweighting method used in the ICT Impacts-ESSLimit-ESSLait projects is provided by Bartelsman [2], who builds on a method developed by Renssen and Nieuwenbroek [11].

The effectiveness of the three reweighting schemes is assessed by comparing register-reweighted values of several variables from the PS and EC samples with the same variables in the PSEC and PSECIS datasets. Sample reweighted variables from the linked business register – PS dataset are considered to reflect closest the population of firms. Therefore, we expect reweighting to bring aggregate values of PSEC and PSECIS variables closer to their reweighted PS (or EC) values.

Table 5. Comparison of total employment for the PS and PSEC samples, by different reweighting approaches (in 1000s)

Country	PS, BR reweighting	PSEC, no reweighting	PSEC, BR reweighting	PSECIS, no reweighting	PSECIS, BR reweighting
AT	1 700	586	1780	249	1 030
DE	14 600	1 840	12 800		
DK	1 010	509	1 070	207	413
FI	1 390	440	949	231	649
FR	5 290	1 920	3 030	1 250	2 520
IE	766	237	741	69	308
IT	8 901	2 276	10 467	1 284	7 089
LU	115	68	110	31	67
NL	3 680	929	3 830	497	2 810
NO	1 110	468	1 140	221	635
SE	1 620	628	1 660	341	908
SI	315	156	307	99	232
UK	13 400	4 430	9 790	1 850	7 930

Note: The figures in the table represent averages over all the available years, for each of the countries.
Source: ESSLait Micro Moments Database

Table 5 shows the total employment in each country, for the PS sample (with business register-based weights), and for the PSEC and PSECIS samples (without reweighting and with business register-based weights). Without reweighting, employment is much lower in

[6] These include ratios, and Boolean variables, which are aggregated by taking the percentage of firms in the industry that have a 'true' value.

both the PSEC and PSECIS datasets, but once the business register-based reweighting is taken into account, employment is inflated to values that are closer to the PS population of firms. The effect of reweighting is lower for the PSECIS, as this is the smallest created dataset and therefore the one where the selection bias problem risks becoming the most severe. Employment figures for Germany are missing for the PSECIS because the IS could not be linked with other datasets.

Table 6 similarly provides a comparison of mean values of the share of ICT intensive human capital[7] (measured as the percentage of employees with post-upper secondary education in the fields of information technology, engineering, mathematics or physics) in the PS and PSEC samples by different reweighting approaches. In the absence of reweighting, average values are slightly higher in the PSEC compared with the register-weighted PS sample, and reweighting generally produces similar results to the unweighted PSEC.

Table 6. Comparison of mean values for HKITpct for the PS and PSEC samples, by different reweighting approaches (per cent)

Country	PS, BR reweighting	PSEC, no reweighting	PSEC, BR reweighting	PSEC, empl. reweighting	PSEC, BR & empl. rewg.
DK	5.08	5.81	4.66	6.62	6.01
FI	9.03	10.28	10.03	11.62	11.01
FR	2.14	2.29	2.27	3.07	2.93
NO	5.01	5.61	4.82	5.72	5.48
SE	5.15	6.19	5.07	7.91	6.72
UK	5.60	5.90	5.71	4.88	4.87

Note: The figures in the table represent averages over all the available years, for each of the countries.
Source: ESSLait Micro Moments Database

We now investigate two variables from the EC survey, the proportion of broadband internet-enabled employees, BROADpct, and a variable showing whether the firm had e-sales during the year (AESELL), as illustrated in Tables 7 and 8, respectively. As far as BROADpct is concerned, the change in average values when merging the PS and EC surveys is again very small or non-existent in most cases. Reweighting the PSEC variable with the help of business-register weights lowers the average share of broadband-enabled employees in all countries except Germany, the Netherlands and the United Kingdom and gives the closest values to the EC sample for most countries.

[7] The ESSLait datasets include three variables describing human capital: HKpct – share of workers with post-upper secondary education, HKITpct – share of workers with ICT intensive post-upper secondary education, and HKNITpct – percentage of workers with post-upper secondary education in fields not related to ICT.

Table 7. Comparison of mean values for BROADpct 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.
AT	0.33	0.36	0.33	0.36	0.36
DE	0.33	0.37	0.34	0.40	0.38
DK	0.33	0.34	0.32	0.34	0.34
FI	0.52	0.54	0.52	0.58	0.55
FR	0.38	0.39	0.38	0.40	0.39
IE	0.23	0.24	0.22	0.28	0.26
IT	0.36	0.27	0.35	0.27	0.27
LU	0.43	0.42	0.42	0.40	0.40
NL	0.36	0.41	0.40	0.41	0.40
NO	0.48	0.51	0.48	0.55	0.52
SE	0.49	0.53	0.49	0.57	0.54
SI	0.44	0.41	0.42	0.36	0.38
UK	0.35	0.41	0.38	0.41	0.40

Note: The figures in the table represent averages over all the available years, for each of the countries.
Source: ESSLait Micro Moments Database

Table 8. 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.
AT	0.16	0.27	0.16	0.47	0.36
DE	0.19	0.29	0.23	0.42	0.36
DK	0.27	0.32	0.28	0.45	0.40
FI	0.20	0.31	0.20	0.55	0.44
FR	0.17	0.25	0.17	0.47	0.43
IE	0.26	0.32	0.28	0.45	0.41
IT	0.07	0.11	0.07	0.28	0.17
LU	0.14	0.19	0.15	0.34	0.31
NL	0.20	0.27	0.22	0.37	0.31
NO	0.27	0.34	0.27	0.46	0.40
SE	0.25	0.35	0.25	0.60	0.47
SI	0.15	0.23	0.15	0.46	0.36
UK	0.17	0.33	0.24	0.48	0.42

Note: The figures in the table represent averages over all the available years, for each of the countries.
Source: ESSLait Micro Moments Database

Unlike BROADpct, which is a composite variable, AESELL is more straightforward to interpret and exhibits consistent patterns across countries. Register-weighted PSEC values are closest to the register-weighted EC mean values for all countries. In contrast, the set of

weights based on firm size leads to the highest estimates, irrespective of the country. This was to be expected because e-sales propensity sharply increases with firm size.

We also assess the sensitivity of results to the different reweighting schemes by performing a series of Mann-Whitney (two-sample Wilcoxon rank-sum) tests on the HKITpct variable (the share of ICT intensive human capital) for the six countries where the variable is available: Denmark, Finland, France, Norway, Sweden and the United Kingdom. The Mann-Whitney test is similar to a t-test, but it is more efficient on non-normal distributions. It is used here to examine whether there are significant differences between the PS population and the PSEC sample, and whether reweighting reduces these discrepancies. We find no significant difference between the underlying distributions of the register-weighted HKITpct values in the PS and weighted PSEC values, irrespective of the reweighting method considered. However, when the register-weighted PS HKITpct is compared with unweighted PSEC HKITpct, the test reveals significant differences for Denmark, Finland and France, with the PSEC variable having a higher rank in the first two countries and a lower rank in the latter. These results suggest that the distribution of the weighted HKITpct variable in the PSEC is the most similar to the PS population, as reweighting seems to bring the HKITpct variable back toward its original (reweighted PS) distribution. However, it is possible that conducting this test on other variables might produce different conclusions.

Based on the results presented in this section, we conclude that the use of reweighted variables is optimal for descriptive statistics. Using average versus weighted average variables depends on the purpose of the analysis. The business register-based weights produce results that are most representative of the universe. Using employment-weighted averages depends on whether one is interested in average values of firms in an industry, or aggregate values. For instance, if an industry only included a small firm with 10 employees and 10 per cent e-sales and a large firm with 90 employees and 90 per cent e-sales, the average e-sales for the industry would be 50 per cent, while the size-weighted average e-sales would be 82 per cent in the industry.

6. Industry-level analyses and reweights

Bartelsman [2] gives an example of how regression estimates vary across micro-aggregated samples as shown in Table 9. In a simple regression of labour productivity (LPV) on highly

skilled human capital (HKITpct) with time and industry held fixed, the results show that the linking does not vastly disturb the parameter estimates – at least not in this case where only one sample survey is linked to the production statistics. PS refers to the merged production survey and business register; PSEC also includes the survey on ICT usages in firms; and HKITpct is the proportion of formally schooled employees with ICT skills.

When deciding whether to use weights to deal with sample bias it is important to consider what firm characteristics are of interest for the analysis. According to Fazio et al. [10], for a sample with many small firms but with a high turnover volume concentrated in a few large firms, as is the case of the PSEC and PSECIS samples, unweighted estimates will provide a reliable picture of production drivers, while weighted estimates may reflect the presence of small firms better.

Table 9. Industry-level regressions with highly ICT skilled human capital across samples

2001-05	**Dependent variable**	**(log) Labour productivity**		**Growth labour productivity**	
Country	Sample	PS	PSEC	PS	PSEC
Finland	HKITpct	1.51	1.82	0.64	0.60
	t-stat	(3.0)	(4.9)	(3.5)	(6.1)
	R-squared	0.09	0.12	0.26	0.23
	Observations	276	276	128	128
Norway	HKITpct	0.76	0.57	0.36	0.27
	t-stat	(3.5)	(2.7)	(1.7)	(2.3)
	R-squared	0.04	0.03	0.02	0.02
	Observations	258	258	210	210
Sweden	HKITpct	1.32	0.79	-0.04	-0.02
	t-stat	(2.5)	(1.9)	(1.7)	(0.8)
	R-squared	0.03	0.01	0.01	0.00
	Observations	260	260	201	201

Note: HKITpct refers to the proportion of ICT schooled employees in accordance with international ISCED classifications. PS means merged production survey and business register, PSEC also includes the survey on ICT usage in firms.
Source: Bartelsman [2]

In order to explore the effect of reweighting on marginal analysis, it would be interesting to perform a series of regressions at the micro level with and without weights in the regression procedure. This approach would make it possible to compare the impact of the various weighting schemes on the estimates. However, we cannot perform such an analysis as the ESSLait project provides micro-aggregated datasets. A solution would be the inclusion of test

regressions in the common code run by NSIs on the firm-level data before aggregating the data.

At the micro aggregated level, a comparison of regressions with and without reweighting is related to scaling the dependent and independent variables to achieve an improved representativeness for the country or industry in question. When both the dependent and the explanatory variables have the same distribution across the sample, they will be affected proportionally by selection bias. However, since the Micro Moments Database includes a set of reweights, we find it necessary to investigate what happens if these reweights are used in marginal analysis.

Table 10. Industry-level regressions

		PS	PS_EC			
Dependent variable		**log(LPV)**	**log(LPV)**	**log(rLPV)**	**log(uLPV)**	**log(ruLPV)**
Country	Sample	No reweighting	No reweighting	BR reweighting	Empl. reweighting	BR & empl. reweighting
DK	log(HKITpct)	*** 0.41	0.09	-0.06	*** 0.46	0.06
	(t)	2.67	(0.97)	(-0.67)	(4.45)	(0.72)
	R-sq	0.09	0.01	0.01	0.22	0.01
	N. obs.	77	77	77	77	77
FI	log(HKITpct)	*** 3.04	* 0.12	0.02	*** 0.41	*** 0.34
	(t)	3.24	(1.85)	(0.17)	(4.11)	(3.69)
	R-sq	0.12	0.05	0.00	0.21	0.18
	N. obs.	84	70	70	70	70
FR	log(HKITpct)	-0.01	0.00	* 0.21	-0.01	-0.03
	(t)	-0.14	(0.00)	(1.92)	(-0.18)	(-0.83)
	R-sq	0.00	0.00	0.10	0.00	0.02
	N. obs.	42	42	42	42	42
NO	log(HKITpct)	0.22	*** 0.28	0.10	*** 0.51	*** 0.34
	(t)	1.67	(3.01)	(1.46)	(5.12)	(3.75)
	R-sq	0.04	0.12	0.03	0.28	0.17
	N. obs.	77	77	77	77	77
SE	log(HKITpct)	*** 0.64	*** 0.40	0.04	*** 0.41	*** 0.39
	(t)	5.65	(2.88)	(0.22)	(3.37)	(3.70)
	R-sq	0.34	0.12	0.00	0.15	0.18
	N. obs.	70	70	69	70	70
UK	log(HKITpct)	*** 0.59	0.02	0.05	-0.16	-0.18
	(t)	3.12	(0.15)	(0.24)	(-1.37)	(-1.30)
	R-sq	0.14	0.00	0.00	0.03	0.03
	N. obs.	70	70	70	70	70

Note: For each regression, we apply the same reweighting treatment to the independent variable (HKITpct) as we do to labour productivity (LPV). LPV refers to the unweighted labour productivity; rLPV refers to the business register reweighted labour productivity; uLPV to the sample size reweighted productivity and ruLPV to

the combined weights variable. Three asterisks indicate that the coefficient is significant at a 1% confidence level, two asterisks – at a 5% confidence level, and one asterisk – at a 10% level.
Source: ESSLait Micro Moments Database and own calculations

Table 10 displays the results of a set of test regressions of labour productivity on ICT intensive human capital in the PS and PSEC samples, using the alternative EUKLEMS industry classification provided in the datasets and including industry and year fixed effects. Although the regression with reweighed variables tends to increase the coefficients, it is difficult to identify clear patterns across countries. An important factor when drawing a conclusion as to which set of weights should be used – if any – is the type of relationship examined. When firm-level relationships are of interest, unweighted variables can be used. In contrast, for macroeconomic relationships, weights are necessary to emphasize the relevance of larger firms, and the employment-based weights achieve this purpose best.

Table 11 summarises results from a series of regressions of labour productivity (LPV) on highly skilled human capital (HKpct), with country, industry and time fixed effects. The regressions are run on several samples for the eight countries where the human capital variable is available – Denmark, Finland, France, the Netherlands, Norway, Sweden, Slovenia and the United Kingdom. The magnitude of the coefficients generally increases as sample size diminishes, most likely due to the fact that smaller, linked samples tend to be dominated by larger firms. Similarly, coefficients are more stable and the relationships are stronger when more weight is assigned to larger firms by applying the employment-based reweighting scheme.

Table 11. Comparison of reweighting schemes in pooled regressions

Dependent variable: Labour productivity (LPV, 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)

Source: ESSLait Micro Moments Database

Fazio et al. [10] propose a method to assess whether the linked sample and population database have similar characteristics. In our case, the procedure would consist of comparing the PSEC sample with a series of random samples extracted from the BR and having the same number of observations as the PSEC. If the mean values from the PSEC are within the range of the random samples, then the PSEC may be treated as a random subset of the BR in regression analysis. Such a test could also be inserted in the common code before the aggregation stage and used to assess the representativeness of the linked dataset.

We recommend users of the ESSLait project datasets to consider carefully the choice of reweighted versus non-reweighted variables, especially for marginal analysis. The take-away point of this chapter is that reweighted values may be used when comparing descriptive statistics between countries, while the use of weights for micro-aggregated regressions depends on the question asked.

7. Firm-level analyses and selection bias

Throughout the ICT Impacts projects, awareness has been high of the possible biases that could appear when different firm-level datasets are merged. Fazio et al. [10] found that marginal analysis of linked firm-level production and ICT usage data were not particularly sensitive to selection bias. Moreover, they experienced that industry and size dummy variables are hugely beneficial for the robustness of firm level regressions.

To investigate if the same conclusion can be drawn for the ICT Impacts-ESSLimit-ESSLait project datasets, we start by looking at some initial results where production function estimations are performed on the full production (PS) as well as the merged production-ICT-usage (PSEC) datasets. These estimations explore the marginal effect on labour productivity from employees highly schooled in ICT (post-upper secondary education in physics, mathematics, engineering or information technology in accordance with the international ISCED classification). Controls are added to this for generally highly skilled employees, firm age, size, international experience, affiliation, as well as fixed industry and time effects. The caveat is that the estimation on linked datasets also includes a couple of ICT intensity variables, and some of these datasets are no longer available for re-estimations. Thus these variables describe the proportion of broadband internet-enabled employees (BROADpct) and

the sum of the degrees of online purchases and sales (ECpct) and might affect the fit of the model (R-squared), although with generally tiny impacts on productivity.

Table 12. 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)

Note: Variables BROADpct and ECpct mean proportion of broadband internet-enabled employees and sum of proportion of e-sales and e-purchases, respectively. The table refers to the years 2001-2009.
Source: ESSLimit PS and PSEC datasets

All estimates remain significant with the same sign when the dataset changes from full production to the merged PSEC sample. In Finland there is hardly even a stir, while in Sweden and Norway the effects of ICT human capital on firm productivity seem to become stronger; in Sweden this is the case despite the BROADpct variable, which renders a not completely negligible impact.

In conclusion, merging one smaller sample survey with a larger dataset or census does not seem to distort regression estimates qualitatively, but may affect the magnitude of the estimated coefficients. An intricate question is what happens when a linked dataset includes more than one small sample survey. As opposed to the example above, identical regressions can be compared for the PSEC and PSECIS datasets, although a smaller deviation exists for the panel of firms. Information is available in the PSEC from 2001 to 2009, while the PSECIS dataset is available only up to the innovation survey wave of 2008.

Firstly, it is important to note that the number of observations fall drastically during the final steps of the linking procedure. The biennial character of the innovation survey is part of the explanation. This also depends on the relatively small sample surveys in most countries and how they are drawn over time. Strategies used for reducing the response burden may vary

greatly, as highlighted in the section above on attrition, which implies that firms will not appear regularly over time in a survey unless they are large. Thus the PSECIS is a subsample of the PSEC only to a certain degree.

Table 13. Number of observations in merged panels

Country	PSEC	PSECIS
AT	25483	4794
DE	10172	
DK	13769	2889
FI	27774	5959
FR	45844	8532
IE	16764	2673
IT	164834	47523
LU	9631	1113
NL	20700	8990
NO	30523	8496
SE	24796	4284
SI	1970	875
UK	25228	8698

The PSEC relates to the years 2001-2009 and the PSECIS to 2002-2008.
Source: ESSLait PSEC and PSECIS datasets

Table 14. Average variable values across samples in 2008

	W, euro thousand		BROADpct	
Country	PSECIS	PSEC	PSECIS	PSEC
AT	55	48	44	48
DK	46	46	42	44
FI	42	37	63	62
FR	53	51	42	42
IE	46	39	24	23
IT	43	41	37	34
LU	53	45	48	54
NL	46	47	54	52
NO	46	39	62	67
SE	60	55	59	59
SI	23	21	50	44
UK	33	34	51	47

Note: W is calculated as total wage bill per employee.
Source: ESSLait PSEC and PSECIS dataset

Before continuing with comparing the regression results, some descriptives of a couple of important explanatory variables are presented. The example will focus on the BROADpct and Wages (W) variables. The former was found by Eurostat [8, 9] to be a good measure of ICT

intensity, and the latter is used as a proxy for human capital when information on formal education is not available.

As can be seen in Table 14, there is a particularly clear pattern for the Wages variable, which generally becomes upward biased in the smaller dataset. However, the change in magnitude is not substantial for most countries. The largest difference is found for firms in Luxembourg. The BROADpct variable is affected somewhat differently, since a group of countries actually exhibit lower or similar values for the ICT intensity of firms in the smaller sample as in the PSEC dataset.

Next, we return to the exploration of regression coefficients with a specification including wages and BROADpct (Table 15). Since information on educational achievement is available only in a few countries, we continue to use the wage variable as a proxy for education.

Table 15. Firm level regressions with ICT intensity variable across samples

	BROADpct Coefficient (*t-stat*)		LnW Coefficient (*t-stat*)		R-squared	
Sample	PSEC	PSECIS	PSEC	PSECIS	PSEC	PSECIS
AT	0.068	0.026	0.964	1.049	0.92	0.94
	(5.67)	*(1.08)*	*(96.4)*	*(47.68)*		
DK	-0.001	0.003	0.969	0.960	0.93	0.92
	(-0.13)	*(0.18)*	*(69.21)*	*(30.00)*		
FI	0.011	0.017	0.915	0.942	0.93	0.91
	(1.22)	*(0.77)*	*(91.50)*	*(34.89)*		
FR	0.049	0.061	0.974	0.989	0.95	0.95
	(8.17)	*(4.36)*	*(162.33)*	*(76.08)*		
IE	0.209	0.308	0.862	1.021	0.83	0.84
	(9.95)	*(5.92)*	*(66.31)*	*(24.90)*		
IT	0.117	0.077	1.024	1.095	0.89	0.91
	(23.40)	*(9.63)*	*(256.00)*	*(156.43)*		
LU	0.113	0.139	0.830	0.646	0.79	0.84
	(5.14)	*(2.04)*	*(39.52)*	*(9.10)*		
NL	0.058	0.083	0.872	0.885	0.92	0.90
	(5.27)	*(5.19)*	*(109.00)*	*(68.08)*		
NO	0.015	0.029	0.978	0.990	0.94	0.92
	(2.14)	*(1.81)*	*(163.00)*	*(66.00)*		
SE	0.028	0.063	0.973	1.013	0.94	0.94
	(3.11)	*(2.52)*	*(97.30)*	*(36.18)*		
SI	-0.004	-0.004	1.185	1.253	0.91	0.93
	(-0.08)	*(-0.06)*	*(30.38)*	*(26.10)*		
UK	0.165	0.175	0.979	1.006	0.87	0.85
	(12.69)	*(8.33)*	*(139.86)*	*(83.83)*		

Note: T-statistics are shown in *italic* within brackets. Grey means insignificant value. Refers to years 2001-2009 for PSEC and 2002-2008 for PSECIS.

Source: ESSLait PSEC and PSECIS datasets

The estimates of the ICT variable appear with certain robustness. In most countries except Austria, where the effect on labour productivity becomes non-significant for the smaller sample, the direction is the same although it becomes slightly larger in the smaller sample.

A pattern even more consistent is shown by the wages variable, where all estimates turn out significant, even if firms in Denmark and Luxembourg reveal a stronger link to productivity in the larger dataset. The fit of the model does not follow any particular pattern and varies only a little from an already high level. In some countries, there is a slight decrease and in others the opposite occurs. These results are well in line with the findings of Fazio et al. [10], who showed that selection bias does not pose a major threat to the robustness of marginal analysis, at least not when conditioning variables are included. Reassuringly, this conclusion seems to hold even for the linking of more than one sample survey to the production data.

8. Conclusions

In this paper we have disentangled and provided examples of how the appearance of selection bias may affect analyses of multi-survey linked firm-level and micro-aggregated datasets, with particular focus on the ESSLait national datasets and the Micro Moments Database.

A general finding is that indicators from descriptive statistics and marginal analysis become somewhat upward biased as more surveys are linked, particularly when exploring our smallest and most unique dataset linking production, ICT and innovation data. ICT indicators seem to be affected from several directions. Naturally, if the production survey in a country is large, or a census, the slide tends to be smaller. This is also the case if a sample co-ordination system is in use, or if the variable is already close to saturation. Finally, the ICT and manufacturing firms seem less sensitive to selection bias than the services firms.

In an ideal world, selection bias can be mitigated by larger surveys or by improved sample co-ordination. These measures are simple enough in theory but in practice they come down due to such issues as costs and the response burden of firms, or are at least not possible to introduce in the short run. Another alternative is to use reweighting, the effects of which have also been analysed here. Comparisons of a set of reweighted values across samples show that applying such a method could shift variable values closer to those observed in the larger, unlinked dataset. A business register-based reweighting scheme seems to be what works best

to bring average values from the smaller linked dataset closer to the larger, supposedly less biased, dataset, while an employment-based reweighting scheme is more effective when dealing with aggregate values.

While Fazio et al. [10] discuss the possible benefits of using reweights in firm-level marginal analysis, we have investigated whether a set of reweights could also be used to improve the representativeness of the industry-level regressions. Bartelsman [2] showed that there was a bias in estimates across samples, although this bias was not large enough to affect interpretations or general conclusions. However, our attempt shows quite inconclusive results and the deviation across samples may not be properly corrected by using micro-aggregated reweights.

Reassuringly, and as concluded by Fazio et al. [10] and Ritchie [12], firm-level estimations seem to be robust against selection bias, even in the smallest multi-linked datasets. The latter show slightly higher estimates of impact, but would not change any qualitative conclusion or interpretation of results.

References

[1] Bartelsman, Eric J., Eva Hagsten and Michael Polder (2013), Cross-Country Analysis of ICT Impact Using Firm-level Data: Micro Moments Database and Research Infrastructure, Eurostat.

[2] Bartelsman, Eric J. (2008), Properties of Linked Data Evidence from the ICT Impacts Project, Final Report, Information Society: ICT Impact Assessment by Linking Data from different Sources.

[3] Bartelsman, Eric J. and Mark Doms (2000), Understanding Productivity: Lessons from Longitudinal Microdata, Journal of Economic Literature, 38(3): 569-594.

[4] Chesher, Andrew and Lars Nesheim (2006), Review of the Literature on the Statistical Properties of Linked Datasets, DTI Occasional Paper 3:
http://www.dti.gov.uk/files/file24832.pdf.

[5] Denisova, Ekaterina (2013), ESSLait Metadata Report, Eurostat.

[6] Denisova, Ekaterina (2012), Final Report on Works Stream C, Survey Strategy Feasibility, ESSNet on Linking of Microdata on ICT Usage, Eurostat, November.

[7] Eurostat (2013), Community Survey on ICT Usage and E-Commerce in Enterprises, https://circabc.europa.eu/sd/d/c98b7697-c8dc-4b3e-b864-d562c2a03788/ICT-Entr%202013%20-%20Model%20Questionnaire%20V%201.1.pdf.

[8] Eurostat (2008), Final Report, Information Society: ICT Impacts Assessment by Linking Data from Different Sources.

[9] Eurostat (2012), Final Report, ESSNet on Linking of Microdata on ICT Usage, November.

[10] Fazio, Gian, Katherine H. Lam and Felix Ritchie (2006), Sample Bias in Microeconometric Analyses of Official Microdata, Report to the Department of Trade and Industry URN 06/737.

[11] Renssen, Robbert H. and Nico J. Nieuwenbroek (1997), Aligning Estimates for Common Variables in Two or More Sample Surveys, Journal of the American Statistical Association, 92(437): 368-374.

[12] Ritchie, Felix (2004), Business Data Linking – Recent UK Experience, Austrian Journal of Statistics, 33(1/2): 89-97.