InGRID

Supporting expertise in inclusive growth

www.inclusivegrowth.eu

Deliverable 11.7

INTEGRATED DATABASE LINKING INTERNATIONAL EMPLOYER AND EMPLOYEE SURVEYS ON WORK ORGANISATION, JOB QUALITY AND PERFORMANCE

Majda Seghir

July 2021

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 730998

Abstract

There is an increasing consensus regarding the great research potential of linked employee and employer data sources in disentangling a wide range of topics in the fields of economics and, more generally, of social sciences. Such framework is however still missing at the EU level although many European countries have a long standing tradition in developing national linked surveys. This report aims at addressing this data gap by exploring available solutions to combine existing European employee and employer surveys. The objective is an optimal use of existing EU data sources rather than developing an EU linked survey. To this end, this report follows a stepwise approach starting with the scan of available methods for data combination followed by the selection of the best employee and employer data sources before proceeding with data harmonisation and integration. The main outcome is three linked surveys at the sector and the country levels available upon request and for which this report details the methodological characteristics.

This report constitutes Deliverable 11.7, for Work Package 11 of the InGRID-2 project.

July 2021

General contact: inclusive.growth@kuleuven.be
p.a. InGRID
HIVA - Research Institute for Work and Society
Parkstraat 47 box 5300, 3000 LEUVEN, Belgium

For more information majda.seghir@lecnam.net

Please refer to this publication as follows:
Seghir, M. (2021). *Integrated database linking international employer and employee surveys on work organisation, job quality and performance* (Deliverable 11.7), Leuven, InGRID-2 project 730998 – H2020

This publication is part of the InGRID-2 project, this project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 730998.

InGRID
Supporting expertise in inclusive growth

Contents

1. Introduction

Linked employee and employer data sources have a certain advantage in disentangling employer and employee effects when analysing a broad range of topics in the field of labour and industrial relations such as work performances or wage setting. Such data forms are mainly available at the national level but an equivalent is still missing at the European level despite the wealth of interesting policy-relevant questions that may be apprehended from a European perspective. The MEADOW[1] project offers, nonetheless, a previous initiative providing norms for the construction of linked surveys, on organisational changes and work restructuring, that may allow comparability at the European level (Greenan *et al.*, 2010). Taking a step further, this deliverable, as part of the InGRID project, has the objective of exploring the potential of linking existing European employer and employee surveys and therefore fills an important gap in existing data.

One often-raised concern regarding the implementation of linked surveys at the EU level is the high cost of such procedure in the absence of a common administrative EU wide and harmonised register. There are, however a number of surveys, carried out at the EU level targeting employers and employees, though data are collected separately. Investigating data linking solutions to combine these employer and employee surveys opens room for optimal use of existing data sources through ex-post data harmonisation and integration. The search for the best methodological framework to combine employer and employee surveys allows, as well, determining the survey design conditions that may foster data integration while preserving data quality.

To address this objective, a stepwise approach has been undertaken to select eligible European surveys and an adequate methodology to combine these data sources before proceeding with the work of data harmonisation and integration. Each step is detailed in the next section resulting in three combined employer-employee surveys. In each combined data set, the employee information is provided by the European Working Conditions Survey (EWCS) while the employer data is delivered by each one of the following surveys: The European Company Survey (ECS), the Community Innovation Survey (CIS) and Continual Vocational Training Survey (CVTS). The pairs of combined data sets are available upon request while this report summarises the characteristics of each pair of data and lists the country and sectors available for each combined dataset.

Based upon the work in this deliverable, a set of policy recommendations are presented in Section 5, identifying opportunities in the actual data design to facilitate efficient data combination of employer and employee surveys.

1 http://www.meadow-project.eu/.

2. Rationale for combining employee and employer data

2.1 Key topics-of-interest

The benefits of combined information from employee- and employer-level data in advancing labour markets research is well-established (Abowd & Kramarz,1999). The main advantage of linked employer and employee data (LEED) is to provide a better framework to disentangle the consequences of the behaviours of both employees and employers when analysing important issues, such as wage settings, technological and organisational changes, or more generally cross-cutting issues in the fields of labour economics, industrial relations as well as institutional and organisational economics. This linked information usually takes two different forms: longitudinal or cross-sectional. Longitudinal linked data are obtained by matching existing data sources, usually administrative data, while cross-sectional LEED are specifically designed linked surveys.

Linked data sources are mostly available at the national level albeit restricted to a handful set of countries (Greenan & Seghir, 2015). From a European and comparative perspective, the only existing linked survey is the European Structure of Earning Survey (ESES). It provides fully comparable information across European States on earning levels with both employees and employers characteristics. In the absence of linked employer-employee registers at the EU level and with respect to the high costs of conducting an EU wide harmonised linked survey, an alternative solution consists in combining information from distinct employer and employee data sources. Indeed, there is a range of employer and employee surveys conducted at the European level. Their combination hence provides joint information on both employers and employees on a variety of topics. Combining existing data sources is an optimal, timesaving and costless solution to widen research perspectives by using both employer and employee level information. However, this solution is accompanied by a set of challenges and limitations represented by which common information are available to combine existing data sources and to what extent data sources can be efficiently harmonised to allow a successful integration. The purpose of this report is twofold. First, to choose the best candidates among European employer- and employee-level surveys to enrich effectively analyses while filling specific conditions with respect to time and country coverage. Second, to provide guidelines on how to undertake such data integration and to implement it on the chosen surveys.

2.2 2.2. The selection of employee and employer level data sources

The choice of the best employee-employer data combination is constrained by two conditions:

- Country coverage: the country coverage of the employer and employee data sources should be broad and EU-wide.
- Reference period: the employer and the employee surveys should cover, as far as possible, the same period. Table 1 reports the reference periods of existing European data sources. Ideally if the employer survey is in the field a given year, the employee survey should be carried out in the same year. However, and depending on the research purpose of the combined data source, the reference period of the employer survey is allowed to be anterior to the reference period of the employee survey. For instance, some organisational changes initiated by the employer need time to have concrete implications at the employee level.

Table 1. Reference periods for employee and employer data sources

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Employee data source										
EWCS	x					x				
ESS	x		x		x		x		x	
LFS	x	x	x	x	x	x	x	x	x	x
Cedefop ESJ Survey					x					
AES		x					x			
Employer data source										
ECS				x						x
CIS	x		x		x		x			
CVTS	x					x				
ESENER					x					x

Note x denotes the reference period of each survey. Employee data sources acronyms refer to: European Working Condition Survey (EWCS), European Social Survey (ESS), Labour Force Survey (LFS), Cedefop European Skill & Jobs Survey (ESJ). Employer data sources correspond to: European Company Survey(ECS), Community Innovation Survey 5CIS), Continuing Vocational Training Survey (CVTS), European Survey of Enterprises on New and Emerging Risks (ESENER).

The scan of existing employee and employer surveys with respect to the country and time coverage provides a first short list of candidates to data combination among which we choose the best combination with respect to topics' coverage.

2.2.1 Employee data sources

Employee data sources with exhaustive coverage of all the aspects related to workers' life are scarce and there is - to our knowledge - only one unique data source at the European level filling this relevancy condition, namely: The European Working Condition Survey. The EWCS was designed by Eurofound to provide comparative and accurate information on working conditions across European Member States. This survey covers a wide array of issues related to the nature of work performed by workers including the working environment, self-assessment of working life quality, employment quality, health and well-being[2]. Moreover, new topics enrich each survey edition to provide a comprehensive picture of changes in the world of work. A selection of trend themes is nevertheless available throughout the different survey editions allowing a temporal perspective. The natural candidate for European employee data source is therefore and unquestionably the EWCS.

2.2.2 Employer data sources:

From the employer side, three data sources have been selected to be combined separately with the EWCS. This selection is based first and foremost on the concordance of reference periods and secondly on the topics covered.

With respect to the time framework, the reference period of the employer data sources should be analogue (or at least prior) to the reference period of the employee level data source. The last editions of the EWCS considered in this report have been conducted in 2010 and 2015, making the selection of employer data source straightforward. Accordingly, CVTS has the perfect match with the EWCS in terms of reference period as they were in the field the same years (2010 and 2015). The CIS survey is the second best match in terms of reference period as this survey was performed in 2010 and 2014.

2 https://www.eurofound.europa.eu/surveys/european-working-conditions-surveys-ewcs.

Finally, the European Company Survey conducted in 2013 is also considered for data combination with the EWCS though performed 2 years earlier than the last edition of EWCS (2015).

Beyond the reference period concordance with the EWCS survey, we select the employer surveys according to the scope of each survey and their relevance in the current policy debate. Starting with the ECS,[3] this survey covers themes related to work organisation, workplace innovation, human resources practices, employee participation and social dialogue. In short, this survey offers an exhaustive panorama of employers' practices that may have an impact on employees working life. The CIS[4] focuses on the innovative behaviour of enterprises, on the different types of innovation produced and on the sources of innovation. Linking the CIS survey with employee data information may help understanding, how forms of work organisation and task contents relate with company level innovativeness. Finally, the CVTS[5] covers continuing vocational training, skills supply and demand as well as training needs. This survey approaches the incentives/obstacles to continuing vocational training and their associated costs. Linking CVTS with ECWS allows confronting employer's strategies for employee skill development with employees' declaration on their skill needs.

Three data combinations are therefore investigated with the EWCS as the main provider of employee data information while employer data are given by the ECS, CIS and CVTS surveys. The next section reviews the available solutions to integrate and combine datasets with the principal aim of identifying the suitable methodology with respects to the selected employee and employer surveys.

3 https://www.eurofound.europa.eu/surveys/european-company-surveys/european-company-survey-2013.

4 https://ec.europa.eu/eurostat/web/microdata/community-innovation-survey.

5 https://ec.europa.eu/eurostat/web/microdata/continuing-vocational-training-survey.

3. Available methods for data integration

The integration of micro databases depends on their sampling frames, on which variables they have in common and on which variables are recorded in only one database. In the following sections, we present available solutions to combine information from different sources with an emphasis on the advantages and limitations of each method.

3.1 Record linkage

Record linkage brings together information from different records belonging to the same entity based on a set of identifiers or quasi-identifiers. The input is therefore only datasets whose sets of units are (partially) overlapping and the records are assumed to have some common identifying information (matching variables) which determine the type of record linkage (Shlomo, 2019):

- Exact (deterministic) record linkage: generally used when the records share a high-quality identifier such as a social security number, otherwise an ID number can be formed from common variables such as age, gender, address, etc. The identifiers are assumed not to be affected by errors in the way data is recorded, though decision rules may be set to declare a pair as a match if there is an agreement on most of the common variables. Each matching variable in this case has an equal weight, such that an agreement on gender would have the same contribution to the overall decision on a correct match as an agreement on last name. Finally, an exact match implies a one-to-one match with only two possible outcomes: match or not match. An illustration is given in Germany by the record linkage of the linked SOEP-LEE[6] survey with administrative data on establishments provided by the German Institute for Employment Research. The linkage procedure in this case is an exact match where there is a perfect correspondence between identifiers of establishments (establishment names, legal forms and addresses) from the linked survey and the administrative records (Eberle & Weinhardt, 2016).
- Probabilistic record linkage: is used to link records with multiple and not-unique identifiers, reported with errors. Probabilities are used then to decide when a given pair of records refers to the same unit, *i.e.* is a match or not. The linkage yields three different stages starting with data cleaning and standardisation, followed by bringing pairs together for comparison and determining correct matches and ending with post-linkage assessment of accuracy taking into account linkage errors (Gill, 2001). Contrary to deterministic linkage, probabilistic linkage relies on frequency analysis of the number of agreements and disagreements to assign a weight to each matching variable, indicating how likely they refer to the same entity. Abowd *et al.* (2019) provide an application of probabilistic record linkage in the United States between a household-level survey (HRS)[7] and an establishment level data in the absence of unique identifiers. The first data source provides employer identity (name and address of the employer) which will be used in the BR[8] to identify the employer. However, the set of matched pairs is greater than 1[9] for almost every HRS job excluding then the use of deterministic match and using probabilistic linkage instead for data linkage.

6 This data stem from a the SOEP-LEE project which aims at creating a linked employer-employee (LEE) dataset, combining information on employees from the German Socio-economic Panel Study (SOEP) with data on their employers (Weinhardt et al., 2016).

7 The HRS survey corresponds to the American Health and Retirement Study which collects information on individuals to study the labour force participation and health transition toward the end of working life.

8 BR is the American Census Bureau's list of essentially all employers.

9 Specific employers and establishments in the BR may have multiple matches for individuals in the HRS.

The procedure of data preparation represents around 75% of the record linkage procedure followed by the check of match accuracy which allows identifying false matches and missed matches (when the pair is declared as not-match but actually refer to the same units).

The record linkage method outperforms existing solutions of data combination by allowing a micro integration of data with an effectiveness close to 100% in case of exact record linkage and a very low number of mismatches. However, this method imposes overlapping units across the data sets to be matched and requires common identifiers which are barely available in all data sources.

3.2 Statistical matching (or Data Fusion)

Statistical matching (SM) brings together independent samples without any units in common but which are sharing the same sampling unit. It is worth pointing out that contrary to record linkage, the data sources in statistical matching do not observe the same units and the chances to have the same units in the data sources are even close to zero. In the simplest case of statistical matching, there are two data sources A and B sharing a set of common variables X while the variables Y are available only in A and the variables Z only in B. The variables Y and Z are then not jointly observed. An example is given by the PISA[10] and TALIS[11] surveys as each survey is missing an important component to understand the educational system: PISA is missing the teacher level while TALIS is missing student level data. As outlined and tested by Kaplan and McCarty (2013), SM is a robust approach to link PISA and TALIS and hence allowing a joint analysis. Another illustration is given by Eurostat (2013) with the matching of the EU-SILC (European Statistics on Income and Living Conditions) survey and the European Quality of Working Life (EQLS). The first survey provides statistics for monitoring poverty and social exclusion - overall aspects related to economic well-being. The EQLS, on the other hand, provides indicators of subjective well-being such as emotional well-being, social participation and trust in institutions. Linking the two data sources relying on SM results on a synthetic file offering a more comprehensive picture of individual's well-being.

In practice, SM can be regarded as an imputation problem of the target variables from a donor to a recipient survey (D'Orazio, Di Zio & Scanu, 2006). In this context, data in the recipient survey is missing completely at random since the variables of interest Y and Z were collected separately and having them in the same data source would be expensive. The SM solution consists of integrating A and B to jointly analyse Y and Z relying on the strong assumption of conditional independence implying that the association between Y and Z, conditional on X, is assumed to be 0. When this assumption holds, the integrated dataset reflects the true joint distribution of variables. To relax this assumption, which holds very rarely, auxiliary information about the relationship between Y and Z may be used allowing to avoid the conditional independence assumption (Rassler, 2002). SM yields two approaches with different outputs:

- macro approach: the data sources are used to estimate the population parameters, e.g. correlation coefficient of Y and Z, contingency tables, probability of an event, ...;
- micro approach: the data sources are used to create a synthetic data file in which all the variables, X, Y and Z are available.

Both parametric and non-parametric frameworks can be used to perform SM. The non-parametric approach offers more flexibility with mixed types of variables, while parametric methods entail the risk of low reliability of the results if the model is not correctly specified. A mixed alternative with both parametric and non-parametric methods is possible with the micro approach by first using a parametric method to estimate the parameters and then a non-parametric method is used to complete

10 The OECD Program for International Student Assessment.

11 The OECD Teaching and Learning International Survey.

the dataset. An example is provided by the predictive mean matching introduced by Rubin(1986) where the predicted value for the observed data is matched to a predicted value of the missing data using a non-parametric method such as hot deck distance.

SM follows a stepwise procedure in each data file to be matched that starts with the harmonisation of statistical units, reference periods, variables and classification; adjustment for measurement errors, adjustment for missing data and derivation variables. The next step is about selecting the most appropriate matching variables before proceeding with statistical matching.

SM requires a careful validation procedure that aims at least at preserving (i) the marginal and joint distribution of the variables from the donor data source in the synthetic dataset after SM, (ii) the covariance structure of the variables after SM, and (iii) the joint distribution of all the variables in the synthetic dataset. The last level of validity aims at preserving individual values but it is usually unnecessary to achieve this level.

SM presents high potential benefits of efficient use of existing data sources at minimum costs while preserving individual relationships for relevant analysis. On the flip side, this method works only and only if the data sets to be matched share the same sampling unit and presents the inherent limitation of the SM work assumptions, *i.e.* the conditional independence assumption, and model-based imputation.

3.3 Data aggregation

Data aggregation is by far the most flexible approach to combine information from different data sources. The only constraint consists of finding a common aggregating level to which the information can be averaged up and then investigated jointly. This approach entails an evident loss of information by substituting individual (or micro)data with aggregate (or macro)data which also represent the main drawback of this method of data combination. Data obtained by aggregation has no information about the degree of heterogeneity or variability at the individual level which is completely lost by the averaging procedure. Moreover, combing data across individuals nested under a large group (such as the occupation or the sector) may carry the risk of meaningless aggregate output in case of high differences across individuals.

Across social science research areas, aggregation is often simply unavoidable because either individual data are not available or the purpose is to make large-scale comparisons (e.g. cross-country comparisons or longitudinal comparisons) or to protect data-confidentiality. Data aggregation is also a good alternative when data characteristics' do not allow a matching between data sources *via* record linkage or statistical matching. For instance, micro datasets with different sampling units and no common identifiers are not eligible to one of the aforementioned methods. In this particular case, data aggregation may solve the problem of data integration from different sources and allows its subsequent analysis though at higher level than the individual one.

The weakness of this methods is related to the higher risk of ecological fallacy, that is the difference in the relationship between variables relying on individual data and on aggregated data. The inherent issue of ecological fallacy to statistical inference on aggregate data is a long-standing problem with a wide-ranging impact across the social sciences. Each discipline has a vast literature on illustrating the problem and solutions to alleviate it. One classical example refers to infer individual voting decisions by comparing election results across districts (Kramer, 1983) or deriving an individual household demand function from aggregate demand conditions across all households of a country (Theil, 1971). In fact, there is a myriad of applications and disciplines (e.g. epidemiology, medicine, marketing, geography, etc.) where the ecological inference problem is posed.

Although estimates from aggregate data cannot replace relationships analysed at the individual level, the ecological fallacy problem may be mitigated relying on two approaches. The first approach aims to define under which circumstances the aggregate estimates provide unbiased estimates of individual-level analysis. For instance, ecological regression fallacy vanishes when the group variable has

no effect on the dependent variable at the individual level, after controlling for other individual-level explanatory variables (Firebaugh, 2001). The second approach relies on methods-of-bounds to derive individual-level effects. This approach is first described by Duncan and Davis (1953) and consists of using the information about the variables to be aggregated to bound the estimated coefficient of correlation that could be obtained at the individual level.

To sum up, data aggregation presents the advantage of being a simple, time-saving and costless solution for data combination and the drawback of losing individual information and heterogeneity at the aggregate level. Deriving unbiased estimates of individual relationships is not straightforward, nonetheless the estimates derived from aggregate data are valid for the particular system of observational units employed.

3.4 Integrating employee and employer surveys

From the above available solutions for data integration, the only solution to link existing employee and employer surveys at the EU level is data aggregation. Indeed, the presented and discussed surveys are distinct sources of information with different survey designs which make other data integration solutions hardly applicable. With respect to record linkage, employee and employer surveys have no overlapping information that allows, for example, to identify the enterprises of workers from employee surveys within employers' surveys. Statistical matching is not applicable either because the sampling units are different in employee and employer surveys: households in employee surveys and establishment or enterprises in employers' surveys. Although the survey design of employee and employer data is different, both surveys share however some common variables (e.g. sector, country) which allow for data aggregation and thus a joint analysis of both surveys but at a higher level than the individual one.

4. Methodological procedure of employee/employer data combination

The procedure of data combination by aggregation follows a general scheme common to all the pair of datasets to be combined. This procedure begins first with data cleaning and transformation, followed by the important step of data harmonisation and finally the aggregation and combination step.

Step 1: Data cleaning and transformation: the objective herein is to dichotomise all the categorical variables to reach a meaningful summation when aggregating the data set. All the qualitative variables, both binary and categorical, take then a value of 0 if 'NO' and 1 if 'YES'. For the ease of data exploitation, the original variable names are kept in the combined datasets in order to use the questionnaire of each original dataset as main data dictionary.[12]

Step 2: Data harmonisation: the target population in the selected employer surveys are enterprises with 10 or more persons employed while the EWCS population of interest are all individuals aged 15 or over in employment. Different sizes of enterprises are then represented in the EWCS sample ranging from very small enterprises with less than 10 employees to large ones with more than 500 employees. As the interest in combining employer and employee surveys lies in providing information on both employers' practices and employees' working conditions, it seems necessary to keep only employees working in establishments of similar size as in the employer surveys. Employees working in small enterprises with less than 10 employees are then removed from the sample of the EWCS. Self-employed workers are dropped as well from the EWCS survey as they are independent workers - in contrast to employees who are subordinate to and dependent on an employer. As a result, the industrial and employment relations as well as working conditions of this category of workers are differently set in comparison with employee workers (Coletto & Pedersini, 2009).

Next to these adjustments regarding the covered population in employer and employee surveys, the harmonisation procedure consists also of harmonising the common variables that will be used as units of data aggregation and combination, namely the country and the economic sector. The available geographical location in all the used survey is represented by the country level and the list of available countries is varying from one survey to another. Similarly, the classification of economic activities available for each survey change across surveys according to the sector anonymisation criteria applied within each survey. Insofar as each pair of data set requires a specific sector harmonisation, this procedure is described in a dedicated section for each pair of combined data set (Section 4.1 for EWCS/CIS, Section 4.2 for EWCS/ECS and Section 4.3 for EWCS/CVTS).

Step 3: Aggregation and combination: the scan of each data source shows only three common variables that may be used as unit of aggregation, namely country, economic sector defined according to NACE Rev.2 and enterprise size. The last variable is not considered to define the aggregation cell because this variable is often anonymised in employers' surveys leading to a break down in three levels: 10-49, 50-249, and +250 employees. Using this variable in the cell definition and with the available data entails the risk of no observations in many cells. Only the country and the sector of activity are then used to define the aggregation cell which becomes the unit of analysis in the com-

12 The web link to each survey questionnaire used for data combination is provided in Table a1 in the Appendix.

bined dataset. Data aggregation is obtained by averaging[13] each variable over the country-sector cell with the constraint of having sufficient cases in each cell to calculate each aggregate. A minimum of three cases is stated for each country-sector cell to respect the confidentiality rules recommended for employer's anonymised data. Once the aggregates from the employer and employee survey are calculated, they are collapsed in a single database where the country-sector is being the new unit of observation.

4.1 EWCS/CIS

The scientific use file version of the CIS data delivered by Eurostat covers a restricted set of countries with anonymisation rules resulting in varying level of aggregation in NACE codes.[14] Two NACE classifications (NACE_A and NACE_B) are available with one being a subset of the other, requesting thus two adjustments of the EWCS economic sectors. The more detailed NACE classification is provided by NACE_A with 68 two-digits economic sectors whilst NACE_B provide 24 aggregated sectors. As NACE_B is a subset of NACE_A, one option is to recode NACE classification into NACE_B for all the countries which make available 24 aggregate sectors by country. The other option consists of preserving whenever it is possible the NACE details and thus keeping both the NACE_A and the NACE_B coding. This NACE coding holds for both CIS 2010 and 2014. Regarding the country coverage, 14 countries are available from CIS 2010 and 14 countries from CIS 2014. The most detailed NACE classification (NACE_A) is available for 8 countries in both CIS 2010 and CIS 2014.[15] Four linked datasets are therefore created according to NACE classifications: the first dataset combines countries with NACE_A classification from CIS 2010 with EWCS 2010, the second one combines countries with NACE_B classification from CIS2010 with EWCS2010, the third one links countries with NACE_A classification from CIS 2014 with EWCS 2015 and the last one combines countries with NACE_B classification from CIS 2014 with EWCS 2015. The number of available countries as well as the total number of sectors available in each combined dataset are reported in Table 2.[16]

Table 2. Total number of available countries and sectors in the combined CIS and EWCS datasets [1]

Combined dataset	Number of available countries	Total number of available sectors
2010_CIS_EWCS_naceA	8 countries	183 sectors
2010_CIS_EWCS_naceB	5 countries [2]	85 sectors
2014_CIS_EWCS_naceA	8 countries	228 sectors
2014_CIS_EWCS_naceB	6 countries [3]	98 sectors

[1] It is possible to have a longitudinal combined sample of EWCS and CIS by selecting only trend questions from both surveys and common countries across time. In this work, we preferred to provide combined samples with all the available information from the original surveys.
[2] The available countries are: Estonia, Croatia, Cyprus, Lithuania and Slovenia.
[3] The available countries are: Estonia, Greece, Croatia, Cyprus, Latvia and Lithuania.

13 Survey weighting are considered for each variable average calculation.
14 Further anonymisation recoding concerns the number of employees, turnover and expenditures values. Please refer to the User manual of CIS anonymised data for further details. In this work, only the recoding of the NACE categories is detailed as this variable is used in the combination step of CIS and EWCS.
15 The list of countries with NACE_A in CIS 2010 and CIS 2014 is; Bulgaria, Czech Republic, Germany, Spain, Hungary, Portugal, Romania, Slovakia.
16 The full details on available sectors by country as well as by combined dataset is provided in Table a2., a3., a4 and a5. in the Appendix.

4.2 EWCS/ECS

The ECS population target is establishments employing more than 10 employees and operating in all sectors except those in NACE Rev.2 A (Agriculture), T (Households as employers), and U (Extra-territorial Organisations). The EWCS covers, on the other hand, all the workers regardless of their economic sector. The sector harmonisation then consists of removing the sectors excluded from the ECS, namely NACE Rev.2 A, T and U, from the EWCS. For the remaining sectors, the NACE Rev. 2 2-digit classification is used resulting on a combined dataset of 899 country-sector cells representing 28 European countries. The list of available sectors for each country is provided in Table a6. in the Appendix.

4.3 EWCS/CVTS

The CVTS scientific use-file provides two types of datasets available for countries which agreed to deliver anonymised data to researchers. The first dataset, labelled 'Standard dataset', is available for almost all the EU countries but applies a very constraining anonymisation criteria regarding sector information: a 5 classes NACE aggregation. The second dataset, labelled 'Optional dataset B', is made available for a set of countries which allow for a more detailed NACE classification: 20 economic sectors. This optional dataset covers, however, only small and medium enterprises *i.e.* enterprises with 10 to 999 persons employed[17] and some variables[18] are further anonymised. The available economic sectors correspond to the following 20 NACE Rev. 2 categories: B, C10-C12, C13-C15, C17-C18, C19-C23, C24-C25, C26C28+C33, C29-C30, C16+C31-C32, D-E, F, G45, G46, G47, H, I, J, K64-K65, K66, L+M+N+R+S. However, only 11[19] out of 30 countries have made this optional dataset available in 2010 and 8[20] out of 30 in 2015. Germany relies on a different grouping[21] resulting in 18 NACE categories. Only optional datasets are combined with EWCS to preserve to some extent detailed information on sectors. To this end, the EWCS NACE Rev.2 classification is harmonised with respect to the sector grouping of CVTS.

Table 3. Total number of available countries and sectors in the combined CVTS and EWCS datasets

Combined dataset	Number of available countries	Total number of available sectors
2010_CVTS4_EWCS	11 countries	198 sectors
2015_CVTS5_EWCS	8 countries	141 sectors

17 An adjustment of the EWCS establishment size is requested however the workplace size as coded in this survey does not allow the identification of workplaces of 1000 persons employed. The establishment size in the EWCS is a four classes variables corresponding to 1, 2-9, 10-249 and more than 250 employees. The correction of establishment size in the EWCS is made by removing establishment with less than 10 employees and keeping the remaining categories.

18 The cost variables are replaced by new variables for which the details are provided in the survey manual.

19 Bulgaria, Czech-Republic, Denmark, Finland, Germany, Hungary, Italy, Romania, Slovakia, Spain and United-Kingdom.

20 Czech-Republic, Denmark, Estonia, Finland, France, Germany, Italy and Spain.

21 Germany use the following nace grouping: B+C19-C23; C10-C12; C13-C15; C17-C18; C24-C25; C26-C28+C33; C29-C30; C16+C31-C32; D-E; F; G45; G46; G47; H; I; J; K64-K65+K66.

5. Policy recommendations

Make better use of European data that already exist requires a number of ex-ante actions to enhance the potential of data integration and linkage. The fulfilment of this objective requires an integrated process of both - data editing and statistical variable derivation. In an ideal case, a full-micro integration by record linkage is attainable if the data sets observe overlapping groups of units. This is usually possible when data collection starts with registers (e.g. population register, business register, ...) and then linked to data from administrative sources or from surveys. The match of these data is performed at the micro level through a common identifier. Such procedure, whenever possible, is successfully implemented to link employee and employer data sources at the national level. An illustration is given by the linked employer-employee dataset from the Institute for Employment Research (IAB) which links information on establishments from the IAB Establishment Panel, an annual establishment survey, with information on individuals employed at those establishments. The set of identifiers comprise ID and establishment identifiers.[22]

In the absence of European register, linking employee and employer data sources at the micro level may be achieved through statistical matching techniques. As outlined in Section 3, this method has a number of pre-requirements regarding the data generating process that, if integrated in the survey design, will enhance the potential of an efficient matching. They are the following:

- In the context of employee-employer data combination, the sampling unit is the first issue to be addressed. Usually households are the sampling units of employee surveys whilst establishment or companies are the sampling units of employer's surveys. The survey design of European surveys on employees and employers should rely on the same sampling unit to allow the use of matching-based models.
- The design of employee and employer questionnaires should include a set of common variables that may favour statistical matching. Whatever matching procedure is used, the results will indeed be based on the conditional independence assumption (CIA) which is a strong and hardly-testable assumption if no joint data on employee and employer is observable. The common variables should then accumulate as much explanatory power as possible on the variables to match, to approach the fulfilment of the CIA.
- Although existing employee and employer surveys observe separate samples of different populations, it may be useful to incorporate auxiliary information, such as overlaps of samples. The objective is to have a small dataset with common information on the variables of interest from employer and employee surveys. The use of auxiliary information will avoid or relax the CIA.

Challenges related to combining employee and employer data sources through statistical matching may be solved by opting for integrated survey models, such as a nested sampling framework or a split questionnaire design (D'Orazio et al, 2006). Such survey designs have the advantage of: (i) addressing the data harmonisation problem, (ii) designing a common questionnaire with basic information and questionnaires for specific units, and (iii) allowing the collection of auxiliary information.

For the improvement of data aggregation as a means of integrating employee and employer information, the data sets should fulfil some basic requirements to reduce the risk of inconsistencies. First,

22 https://fdz.iab.de/en/Integrated_Establishment_and_Individual_Data/LIAB/LIABQM9317.aspx.

the common variables for aggregation should be identically defined in both data sources or at least their harmonisation process should not introduce any substantial inconsistency. In this regard, the use of international classifications or a clear translation link to them is recommended for the shared variables. For instance, macro aggregation of employee and employer data sources at the sector level requires a homogeneous sector classification in both data sources. Second, the data sources should be large enough to a have minimum number of observations at the aggregated level. When the aggregation level is obtained by the interaction of two variables (e.g. country and sector, country and region, ...), the available observations for data aggregation are more likely to be either small or non-existent. The sample size should be enlarged or at least the response rate improved by, for instance, shortening survey questionnaires.

Combining employee and employer data sources opens up possibilities for reducing costs while increasing the statistical output as well as its quality. Laying the groundwork for data integration in the earlier stages of data generation will enable and enhance the integration process of different data sources.

appendix 1

Table a1. Web links to each survey questionnaire

Survey	Web links
EWCS 2015	https://www.eurofound.europa.eu/sites/default/files/page/field_ef_documents/6th_ewcs_2015_final_source_master_questionnaire.pdf
EWCS2010	https://www.eurofound.europa.eu/sites/default/files/ef_files/surveys/ewcs/2010/documents/masterquestionnaire.pdf
ECS	https://www.eurofound.europa.eu/sites/default/files/ef_files/surveys/ecs/2013/documents/3ecsquestionnairemm.pdf
CIS 2014	https://circabc.europa.eu/ui/group/47133480-29c1-4c23-9199-72a631f4fd96/library/32ab7d19-446e-404c-9ea5-e2524065b2a0/details
CIS 2010	https://circabc.europa.eu/ui/group/47133480-29c1-4c23-9199-72a631f4fd96/library/6f5dc4f5-920e-433a-8576-c97bcea6f863/details
CVTS 2015	https://circabc.europa.eu/sd/a/43cebda2-6b09-4298-b9b9-ba57d1f42301/1_CVTS5manual_V1-2_20161201.pdf
CVTS 2010	https://circabc.europa.eu/sd/a/76f1c351-6b9c-497a-b3b0-876820d91507/1_MANUAL_Version6.pdf

Table a2. List of available sectors by country in the combined dataset 2010_CIS_EWCS_naceA

CIS_EWCS_OP_A_2010	BG	CZ	DE	HU	PT	RO	SK	ES
Accommodation+Food & beverage service activities								x
Activities auxiliary to financial services & insurance activities			x					
Advertising & market research								x
air transport			x					
Architectural & engineering activities; technical testing & analysis								x
Civil engineering					x		x	x
Computer programming, consultancy & related activities+information service activities	x	x	x	x		x	x	x
Construction of buildings							x	x
Education								x
Electricity, gas, steam & air conditioning supply	x	x	x	x		x	x	
Financial service activities, except insurance & pension funding	x	x	x	x				x
Human health activities					x			x
Insurance, reinsurance & pension funding, except compulsory social security		x	x	x				x
Land transport & transport via pipelines	x	x	x	x	x	x	x	x
Legal & accounting activities+Activities of head offices; management consultancy activities			x					x
Manufacture of basic metals		x	x	x	x	x	x	
Manufacture of basic pharmaceutical products		x	x	x				
Manufacture of coke & refined petroleum products+Manufacture of chemicals & chemical products	x		x	x		x		x
Manufacture of electrical equipment	x	x	x	x		x	x	
Manufacture of fabricated metal products, except machinery & equipment	x	x	x	x	x	x	x	x
Manufacture of food products+beverages+tobacco products	x	x	x	x	x	x	x	x
Manufacture of furniture	x	x	x	x	x	x	x	x
Manufacture of machinery & equipment		x	x	x		x	x	
Manufacture of motor vehicles, trailers & semi-trailers		x	x	x	x	x	x	x
Manufacture of other non-metallic mineral products		x	x		x	x	x	x
Manufacture of other transport equipment						x	x	
Manufacture of paper & paper products			x				x	
Manufacture of rubber & plastic products	x	x	x	x	x		x	
Manufacture of textiles		x	x		x	x		
Manufacture of wearing apparel+leather & related products	x	x		x	x	x	x	

CIS_EWCS_OP_A_2010	BG	CZ	DE	HU	PT	RO	SK	ES
Manufacture of wood & of products		x		x	x	x	x	
Motion picture, video & television programme production+Programming & broadcasting activities								x
Office administrative, office support & other business support activities			x					
Other professional, scientific & technical activities								x
Postal & courier activities	x		x	x			x	
Printing & reproduction of recorded media	x	x	x	x	x			x
Repair & installation of machinery & equipment					x		x	
Retail trade, except of motor vehicles & motorcycles					x			x
Section A Agriculture, forestry & fishing								x
Section B mining & quarrying	x	x		x		x		
Security & investigation activities			x					x
Services to buildings & landscape activities			x					x
Sewerage+Waste collection, treatment & disposal activities+Remediation activities & other waste management services			x		x			
Specialised construction activities					x		x	x
Sports activities & amusement & recreation activities								x
Telecommunications	x		x		x	x		x
Warehousing & support activities for transportation			x					x
Water collection, treatment & supply	x			x				
Water transport	x							
Wholesale & retail trade & repair of motor vehicles								x
Wholesale trade, except of motor vehicles & motorcycles	x	x	x	x	x	x	x	x

Table a3. List of available sectors by country in the combined dataset 2010_CIS_EWCS_naceB

CIS_EWCS_OP_B_2010	HR	CY	EE	LT	SI
Accommodation+Food & beverage service activities	x				
Construction of buildings+Civil engineering+Specialised construction activities	x			x	
Electricity, gas, steam & air conditioning supply	x		x	x	x
Financial service activities +insurance, reinsurance & pension funding,+Activities auxiliary to financial services	x	x	x	x	x
Land transport & transport via pipelines+Water transport+ air transport	x	x	x	x	x
Legal & accounting activities+Activities of head offices+Architectural & engineering activities; technical testing & analysis+Scientific research & development&Veterinary activities+Advertising & market research+Other professional, scientific & technical activities	x	x	x	x	x
Manufacture of basic metals+Manufacture of fabricated metal products, except machinery & equipment	x	x	x	x	x
Manufacture of coke & refined petroleum products+ Chemical products+Basic pharmaceutical products	x	x	x	x	x
Manufacture of computer, electronic & optical products+Electrical equipment +Machinery & equipment	x		x	x	x
Manufacture of food products+beverages+tobacco products	x	x	x	x	x
Manufacture of furniture+Other manufacturing	x	x	x	x	x
Manufacture of motor vehicles, trailers & semi-trailers+Manufacture of other transport equipment	x		x		x
Manufacture of rubber & plastic products+Manufacture of other non-metallic mineral products	x	x	x	x	x
Manufacture of textiles+Manufacture of wearing apparel+leather & related products	x		x	x	x
Manufacture of wood & of products+Manufacture of paper & paper products	x		x	x	x
Printing & reproduction of recorded media	x	x	x		x
Publishing activities+Motion picture+Programming & broadcasting activities+Telecommunications+ Computer programming, consultancy & related activities+Information service activities	x	x	x	x	x
Section B mining & quarrying			x		x
Warehousing & support activities for transportation+Postal & courier activities	x	x	x	x	x
Water collection, treatment & supply+Sewerage+Waste collection+Remediation activities	x		x		
Wholesale & retail trade & repair of motor vehicles+Wholesale trade+Retail trade	x	x	x	x	x

Table a4. List of available sectors by country in the combined dataset 2014_CIS_EWCS_naceA

CIS_EWCS_OP_A_2014	BG	CZ	DE	HU	PT	RO	SK	ES
Accommodation+Food & beverage service activities								x
Activities auxiliary to financial services & insurance activities				x				
Activities of membership organisations								x
Advertising & market research			x			x		x
air transport			x					x
Architectural & engineering activities; technical testing & analysis			x				x	x
Civil engineering							x	x
Computer programming, consultancy & related activities+information service activities	x	x	x	x	x			x
Construction of buildings							x	x
Creative, arts & entertainment activities+Libraries, archives, museums & other cultural activities								x
Education								x
Electricity, gas, steam & air conditioning supply	x	x	x	x		x	x	x
Employment activities			x					x
Financial service activities, except insurance & pension funding	x	x	x	x	x		x	x
Gambling & betting activities								x
Human health activities					x			x
Insurance, reinsurance & pension funding, except compulsory social security	x	x	x		x			x
Land transport & transport via pipelines	x	x	x	x	x	x	x	x
Legal & accounting activities+Activities of head offices; management consultancy activities			x		x			x
Manufacture of basic metals		x	x		x		x	x
Manufacture of basic pharmaceutical products			x					x
Manufacture of coke & refined petroleum products+Manufacture of chemicals & chemical products	x		x	x				x
Manufacture of computer, electronic & optical products			x	x			x	x
Manufacture of electrical equipment	x	x	x	x	x	x	x	x
Manufacture of fabricated metal products, except machinery & equipment	x	x	x	x	x	x	x	x
Manufacture of food products+beverages+tobacco products	x	x	x	x	x	x	x	x
Manufacture of furniture	x		x	x			x	
Manufacture of machinery & equipment	x	x	x	x			x	x
Manufacture of motor vehicles, trailers & semi-trailers	x	x	x	x	x	x	x	x
Manufacture of other non-metallic mineral products	x	x	x				x	x
Manufacture of other transport equipment								x
Manufacture of paper & paper products			x					
Manufacture of rubber & plastic products	x	x	x	x	x		x	x
Manufacture of textiles					x		x	x
Manufacture of wearing apparel+leather & related products	x	x	x	x		x	x	x
Manufacture of wood & of products		x	x		x	x		x
Motion picture, video & television programme production+Programming & broadcasting activities	x							x
Office administrative, office support & other business support activities			x					x
Other manufacturing	x	x	x	x				x
Other personal service activities								x
Other professional, scientific & technical activities			x		x			
Postal & courier activities	x	x	x	x	x	x	x	x
Printing & reproduction of recorded media	x		x	x				x
Publishing activities			x					
Real estate activities								x
Repair & installation of machinery & equipment	x		x	x		x	x	x
Residential care activities								x
Retail trade, except of motor vehicles & motorcycles					x			x
Scientific research & development+Veterinary activities	x	x	x				x	x
Section A Agriculture, forestry & fishing								x
Section B mining & quarrying		x	x					
Security & investigation activities			x					x
Services to buildings & landscape activities			x				x	x

CIS_EWCS_OP_A_2014	BG	CZ	DE	HU	PT	RO	SK	ES
Sewerage+Waste collection, treatment & disposal activities+Remediation activities & other waste management services	x		x			x		x
Social work activities without accomodation								x
Specialised construction activities					x		x	x
Sports activities & amusement & recreation activities								x
Telecommunications	x	x	x	x	x			x
Travel agency, tour operator reservation service & related activities			x					x
Warehousing & support activities for transportation	x		x		x	x	x	x
Water collection, treatment & supply		x		x			x	x
Wholesale & retail trade & repair of motor vehicles								x
Wholesale trade, except of motor vehicles & motorcycles	x	x	x	x	x	x	x	x

Table a5. List of available sectors by country in the combined dataset 2014_CIS_EWCS_naceB

CIS_EWCS_OP_B_2014	HR	CY	EE	EL	LV	LT
Accommodation+Food & beverage service activities	x					
Construction of buildings+Civil engineering+Specialised construction activities	x					x
Electricity, gas, steam & air conditioning supply	x		x	x	x	x
Financial service activities +insurance, reinsurance & pension funding,+Activities auxiliary to financial services	x	x	x	x	x	x
Land transport & transport via pipelines+Water transport+ air transport	x	x	x	x	x	x
Legal & accounting activities+Activities of head offices+Architectural & engineering activities; technical testing & analysis+Scientific research & development&Veterinary activities+Advertising & market research+Other professional, scientific & technical activities	x	x	x	x	x	x
Manufacture of basic metals+Manufacture of fabricated metal products, except machinery & equipment	x	x	x	x	x	x
Manufacture of coke & refined petroleum products+ Chemical products+Basic pharmaceutical products	x	x	x	x	x	
Manufacture of computer, electronic & optical products+Electrical equipment +Machinery & equipment	x	x	x	x		
Manufacture of food products+beverages+tobacco products	x	x	x	x	x	x
Manufacture of furniture+Other manufacturing	x		x	x		x
Manufacture of motor vehicles, trailers & semi-trailers+Manufacture of other transport equipment	x					
Manufacture of rubber & plastic products+Manufacture of other non-metallic mineral products	x	x	x	x	x	x
Manufacture of textiles+Manufacture of wearing apparel+leather & related products	x		x	x	x	x
Manufacture of wood & of products+Manufacture of paper & paper products	x	x	x		x	x
Printing & reproduction of recorded media	x		x			
Publishing activities+Motion picture+Programming & broadcasting activities+Telecommunications+ Computer programming, consultancy & related activities+Information service activities	x	x	x	x	x	x
Real estate activities	x					
Rental & leasing activities+Employment activities+Travel agency, tour operator reservation service & related activities+Security & investigation activities+Services to buildings & l&scape activities+Office administrative, office support & other business support activities			x			
Repair & installation of machinery & equipment				x	x	
Section B mining & quarrying	x					
Warehousing & support activities for transportation+Postal & courier activities	x	x	x	x	x	x
Water collection, treatment & supply+Sewerage+Waste collection+Remediation activities	x	x	x	x	x	
Wholesale & retail trade & repair of motor vehicles+Wholesale trade+Retail trade	x	x	x	x	x	x

Table a6. List of available sectors by country in the combined dataset ECS_EWCS

ECS_EWCS	AT	BE	BG	HR	CY	CZ	DK	EE	FI	FR	DE	EL	HU	IE	IT	LV	LT	LU	MT	NL	PL	PT	RO	SK	SI	SP	SE	UK
Accommodation	x	x	x	x	x	x		x		x	x	x		x	x	x	x	x	x	x		x	x	x	x	x	x	x
Activities auxiliary to financial services																												x
Activities of head offices		x			x		x		x									x	x	x						x		x
Activities of membership organisations	x	x					x		x	x	x											x					x	
Advertising & market research							x			x	x					x		x		x			x		x	x	x	x
Air transport																		x		x								
Architectural & engineering activities	x	x					x	x	x	x	x				x		x	x	x	x					x	x	x	x
Civil engineering	x	x	x	x			x	x	x	x	x					x	x			x	x		x	x	x	x	x	x
Computer programing	x	x	x		x	x	x	x	x	x	x	x	x	x			x	x	x	x		x			x	x	x	x
Construction of buildings	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Creative		x					x			x	x						x				x					x	
Electricity	x	x				x	x	x		x	x	x	x		x	x	x	x			x			x	x	x	x	
Employment activities	x	x					x		x	x	x									x					x	x	
Financial service activities	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x		x		x	x	x		x		x	x	x
Food & beverage service activities	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Gambling & betting activities																			x							x		
Information service activities									x		x									x						x		
Insurance	x	x			x	x	x			x	x			x	x			x		x						x	x	x
Land transport & transport via pipeline	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Legal & accounting activities	x	x		x	x		x	x	x	x	x	x	x	x	x		x	x	x	x	x	x		x	x	x		x
Libraries		x		x													x				x				x		x	x
Manufacture of basic metals	x	x				x			x	x	x				x			x			x	x		x	x	x	x	
Manufacture of basic pharmaceutical products	x	x					x			x	x			x											x	x	x	
Manufacture of beverages														x												x		
Manufacture of chemicals & chemical products	x	x			x					x	x		x	x	x					x	x				x	x	x	x
Manufacture of computer		x					x	x		x	x		x	x					x	x				x		x	x	x
Manufacture of electrical equipment	x	x	x	x		x	x	x			x		x		x						x	x	x	x	x	x		x
Manufacture of fabricated metal product	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x		x	x	x	x	x	x	x	x	x
Manufacture of food products	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Manufacture of furniture	x	x	x	x				x			x	x	x		x		x				x			x	x		x	
Manufacture of leather & related products											x				x								x			x		
Manufacture of machinery & equipment	x	x	x			x	x		x	x	x		x	x	x					x	x			x	x	x	x	x
Manufacture of motor vehicles	x	x		x		x				x	x		x		x						x	x	x	x	x	x	x	x
Manufacture of other non-metallic mineral products	x	x	x	x	x	x	x	x	x	x	x			x	x		x	x			x			x	x	x		

ECS_EWCS	AT	BE	BG	HR	CY	CZ	DK	EE	FI	FR	DE	EL	HU	IE	IT	LV	LT	LU	MT	NL	PL	PT	RO	SK	SI	SP	SE	UK
Manufacture of other transport equipment				x																	x					x		x
Manufacture of paper & paper products	x								x	x	x														x			
Manufacture of rubber & plastic products	x	x	x			x		x		x	x	x	x		x		x	x	x		x	x		x	x	x		x
Manufacture of textiles	x	x								x					x		x				x	x			x	x		
Manufacture of wearing apparel			x	x				x				x	x		x	x	x				x		x	x		x		
Manufacture of wood & of products of wood	x	x		x		x		x	x		x				x	x	x				x	x	x		x	x		x
Motion picture																										x		
Office administrative		x			x		x		x	x	x	x	x	x	x			x		x	x	x	x			x	x	x
Other manufacturing	x	x				x	x	x		x	x		x	x	x					x	x						x	x
Other personal service activities		x						x	x	x	x		x		x	x		x			x	x		x		x		
Other professional							x				x											x					x	
Postal & courier activities															x		x	x			x					x		x
Printing & reproduction of recorded media	x	x									x		x						x		x				x	x	x	x
Publishing activities	x	x		x			x		x	x	x									x								
Real estate activities	x	x		x		x	x	x		x	x									x				x		x		x
Rental & leasing activities		x																										x
Repair & installation of machinery		x	x						x	x	x	x	x			x			x	x			x		x	x	x	
Retail trade	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Scientific research & development		x				x	x		x	x	x									x	x			x	x	x		x
Security & investigation activities		x	x	x	x	x				x	x	x	x			x			x	x	x	x	x	x	x	x		x
Services to buildings & landscape activities	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	
Specialised construction activities	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x
Sports activities & amusement		x	x	x	x		x		x	x	x		x		x			x		x						x	x	x
Telecommunications						x	x	x			x		x												x	x	x	x
Travel agency					x		x				x				x				x							x		x
Warehousing & support activities for transportation	x	x	x					x	x	x	x	x		x	x		x			x	x	x	x	x		x	x	x
Waste collection		x	x							x	x	x			x					x			x		x	x		
Water collection		x		x		x				x			x								x					x		
Water transport															x	x			x									
Wholesale & retail trade & repair of motor vehicles	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x
Wholesale trade	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Table a7. List of available sectors by country in the combined dataset 2010_CVTS4_EWCS

	BG	CZ	DK	FI	DE	HU	IT	RO	SK	ES	UK
Accommodation and food service activities	x	x	x	x	x	x	x	x	x	x	x
Activities auxiliary to financial services and insurance activities											x
Construction	x	x	x	x	x	x	x	x	x	x	x
Electricity, gas, steam and air conditioning supply; water supply; sewerage, waste management and remediation activities	x	x	x	x	x	x	x	x	x	x	x
Financial service activities, except insurance and pension funding; insurance, reinsurance and pension funding, except compulsory social security	x	x	x	x		x	x		x	x	x
Financial service activities, except insurance and pension funding; insurance, reinsurance and pension funding, except compulsory social security+Activities auxiliary to financial services and insurance activities					x						
Information and communication	x	x	x	x	x	x	x	x	x	x	x
Manufacture of basic metals; fabricated metal products, except machinery and equipment	x	x	x	x	x	x	x	x	x	x	x
Manufacture of coke and refined petroleum products; chemicals and chemical products; basic pharmaceutical products and pharmaceutical preparations; rubber and plastic products; other non-metallic mineral products	x	x	x	x		x	x	x	x	x	x
Manufacture of computer, electronic and optical products; electrical equipment; machinery and equipment n.e.c.; repair and installation of machinery and equipment	x	x	x	x	x	x	x	x	x	x	x
Manufacture of food products; beverages; tobacco products	x	x	x	x	x	x	x	x	x	x	x
Manufacture of motor vehicles, trailers and semi-trailers; other transport equipment		x	x	x	x	x	x	x	x	x	x
Manufacture of paper and paper products; printing and reproduction of recorded media	x	x	x	x	x	x	x		x	x	x
Manufacture of textiles; wearing apparel; leather and related products	x	x			x	x	x	x	x	x	x
Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials; furniture; other manufacturing	x	x	x	x	x	x	x	x	x	x	
Mining and quarrying	x	x				x		x			x
Mining and quarrying+Manufacture of coke and refined petroleum products; chemicals and chemical products; basic pharmaceutical products and pharmaceutical preparations; rubber and plastic products; other non-metallic mineral products					x						
Real estate activities; professional, scientific and technical activities; administrative and support service activities; arts, entertainment and recreation; other service activities	x	x	x	x	x	x	x	x	x	x	x
Retail trade, except of motor vehicles and motorcycles	x	x	x	x	x	x	x	x	x	x	x
Transportation and storage	x	x	x	x	x	x	x	x	x	x	x
Wholesale and retail trade and repair of motor vehicles and motorcycles	x	x	x	x	x	x	x	x	x	x	x
Wholesale trade, except of motor vehicles and motorcycles	x	x	x	x	x	x	x	x	x	x	x

Table a8. List of available sectors by country in the combined dataset 2015_CVTS5_EWCS

	CZ	DK	EE	FI	FR	DE	IT	ES
Accommodation and food service activities	x	x	x	x	x	x	x	x
Construction	x	x	x	x	x	x	x	x
Electricity, gas, steam and air conditioning supply; water supply; sewerage, waste management and remediation activities	x	x	x	x	x	x	x	x
Financial service activities, except insurance and pension funding; insurance, reinsurance and pension funding, except compulsory social security	x	x	x	x	x		x	x
Financial service activities, except insurance and pension funding; insurance, reinsurance and pension funding, except compulsory social security+Activities auxiliary to financial services and insurance activities						x		
Information and communication	x	x	x	x	x	x	x	x
Manufacture of basic metals; fabricated metal products, except machinery and equipment	x	x	x	x	x	x	x	x
Manufacture of coke and refined petroleum products; chemicals and chemical products; basic pharmaceutical products and pharmaceutical preparations; rubber and plastic products; other non-metallic mineral products	x	x	x	x	x		x	x
Manufacture of computer, electronic and optical products; electrical equipment; machinery and equipment n.e.c.; repair and installation of machinery and equipment	x	x	x	x	x	x	x	x
Manufacture of food products; beverages; tobacco products	x	x	x	x	x	x	x	x
Manufacture of motor vehicles, trailers and semi-trailers; other transport equipment	x			x	x	x	x	x
Manufacture of paper and paper products; printing and reproduction of recorded media			x	x	x	x	x	x
Manufacture of textiles; wearing apparel; leather and related products	x		x	x	x	x	x	x
Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials; furniture; other manufacturing	x	x	x	x	x	x	x	x
Mining and quarrying	x			x				
Mining and quarrying+Manufacture of coke and refined petroleum products; chemicals and chemical products; basic pharmaceutical products and pharmaceutical preparations; rubber and plastic products; other non-metallic mineral products						x		
Real estate activities; professional, scientific and technical activities; administrative and support service activities; arts, entertainment and recreation; other service activities	x	x	x	x	x	x	x	x
Retail trade, except of motor vehicles and motorcycles	x	x	x	x	x	x	x	x
Transportation and storage	x	x	x	x	x	x	x	x
Wholesale and retail trade and repair of motor vehicles and motorcycles	x	x	x	x	x	x	x	x
Wholesale trade, except of motor vehicles and motorcycles	x	x	x	x	x	x	x	x

References

Abowd, J. M., Abramowitz, J., Levenstein, M., McCue, K., Patki, D., Raghunathan, T., ... & Wada, N. (2019). *Optimal Probabilistic Record Linkage: Best Practice for Linking Employers in Survey and Administrative Data.* US Census Bureau, Center for Economic Studies.

Abowd, J. M., & Kramarz, F. (1999). The analysis of labor markets using matched employer-employee data. H*andbook of labor economics, 3*, 2629-2710.

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical matching: Theory and practice.* John Wiley & Sons.

Duncan, O. D., and Davis, B. (1953). An alternative to ecological correlation. A*m. Sociol. Rev. 18*, 665 - 666.

Eberle, J. & Weinhardt, M. (2016). *Record Linkage of the Linked Employer-Employee Survey of the Socio-Economic Panel Study (SOEP-LEE) and the Establishment History Panel (BHP)*. German Record Linkage Center Working Paper Series, No. WP-GRLC-2016-01. http://dx.doi.org/10.2139/ssrn.3530779

Coletto, D & Pedersini, R. (2009). *Self-Employed Workers: Industrial Relations and Working Conditions*. Eurofound Publications.

Eurostat (2013). *Statistical matching: a model based approach for data integration.*

Firebaugh, G. (2001). Ecological Fallacy, Statistics of. Pp. 4023-4026 in Neil J. Smelser and Pual B. Baltes, eds., *International Encyclopedia of the Social and Behavioral Sciences*. Oxford, UK: Elsevier.

Greenan N., Lorenz E., Allan S., Amossé T., Archiburgi D., *et al.* (2010), *The MEADOW Guidelines,* 978-2-7466-1809-1. (halshs-01362486).

Greenan, N. & Seghir, M. (2015). *Inventory of linked employer-employees urveys on working conditions and health and safety issues*. Milestone 21.9 of WP 21. 2015. halshs-01376974

Gill L., (2001). Methods for automatic record matching and linkage and their use in national statistics, *National Statistics Methodological Series No. 25*, London (HMSO).

Kaplan, D., McCarty, A.T. (2013). Data fusion with international large scale assessments: a case study using the OECD PISA and TALIS surveys. *Large-scale Assess Educ 1*, 6.

Kramer, G. H. (1983). The ecological fallacy revisited: Aggregate-versus individual-level findings on economics and elections, and sociotropic voting. *The American Political Science Review*, 92-111.

Leulescu, A., & Agafitei, M. (2013). Statistical matching: a model based approach for data integration. *Eurostat-Methodologies and Working papers, 10*-2.

Rässler, S. (2012). Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches (Vol. 168). Springer Science & Business Media.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiplev imputation. *Journal of Business and Economic Statistics, 4*, 87-95.

Shlomo N. (2019). Overview of Data Linkage Methods for Policy Design and Evaluation. In: Crato N., Paruolo P. (eds) Data-Driven Policy Impact Evaluation. Springer, Cham. https://doi.org/10.1007/978-3-319-78461-8_4.

Theil, H. (1971). *Principles of Econometrics*. New York: John Wiley and Sons.

Weinhardt, M., A. Meyermann, S. Liebig, and J. Schupp (2016). The Linked Employer–Employee

Study of the Socio-Economic Panel (SOEP-LEE): Project Report. SOEPpapers No. 829.

InGRID-2
Integrating Research Infrastructure for European expertise on Inclusive Growth from data to policy

Referring to the increasingly challenging EU2020-ambitions of Inclusive Growth, the objectives of the InGRID-2 project are to advance the integration and innovation of distributed social sciences research infrastructures (RI) on 'poverty, living conditions and social policies' as well as on 'working conditions, vulnerability and labour policies'. InGRID-2 will extend transnational on-site and virtual access, organise mutual learning and discussions of innovations, and improve data services and facilities of comparative research. The focus areas are (a) integrated and harmonised data, (b) links between policy and practice, and (c) indicator-building tools.
Lead users are social scientist involved in comparative research to provide new evidence for European policy innovations. Key science actors and their stakeholders are coupled in the consortium to provide expert services to users of comparative research infrastructures by investing in collaborative efforts to better integrate microdata, identify new ways of collecting data, establish and improve harmonised classification tools, extend available policy databases, optimise statistical quality, and set-up micro-simulation environments and indicator-building tools as important means of valorisation. Helping scientists to enhance their expertise from data to policy is the advanced mission of InGRID-2. A new research portal will be the gateway to this European science infrastructure.

This project is supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 730998.

More detailed information is available on the website: www.inclusivegrowth.eu

Co-ordinator
Monique Ramioul

KU LEUVEN HIVA
RESEARCH INSTITUTE FOR WORK AND SOCIETY

Partners

TÁRKI Social Research Institute Inc. (HU)
Amsterdam Institute for Advanced Labour Studies – AIAS, University of Amsterdam (NL)
Swedish Institute for Social Research - SOFI, Stockholm University (SE)
Economic and Social Statistics Department, Trier University (DE)
Centre for Demographic Studies – CED, University Autonoma of Barcelona (ES)
Luxembourg Institute of Socio-Economic Research – LISER (LU)
Herman Deleeck Centre for Social Policy – CSB, University of Antwerp (BE)
Institute for Social and Economic Research - ISER, University of Essex (UK)
German Institute for Economic Research – DIW (DE)
Centre for Employment and Work Studies – CEET, National Conservatory of Arts and Crafts (FR)
Centre for European Policy Studies – CEPS (BE)
Department of Economics and Management, University of Pisa (IT)
Department of Social Statistics and Demography – SOTON, University of Southampton (UK)
Luxembourg Income Study – LIS, asbl (LU)
School of Social Sciences, University of Manchester (UK)
Central European Labour Studies Institute – CELSI (SK)
Panteion University of Social and Political Sciences (GR)
Central Institute for Labour Protection – CIOP, National Research Institute (PL)

InGRID-2

Integrating Research Infrastructure for European expertise on Inclusive Growth from data to policy Contract N° 730998

For further information about the InGRID-2 project, please contact
inclusive.growth@kuleuven.be
www.inclusivegrowth.eu
p/a HIVA – Research Institute for Work and Society
Parkstraat 47 box 5300
3000 Leuven
Belgium