Deliverable Number: D8.11

Deliverable Title: **Occupation > industry predictions for measuring industry in surveys**

Work Package: WP8

Deliverable type: Other

Dissemination status: Public

Submitted by: University of Amsterdam and SHARE

Authors:

Michele Belloni (SHARE, University Ca' Foscari of Venice)

Kea Tijdens (AIAS, University of Amsterdam)

Date Submitted: October 2017

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 654221.

www.seriss.eu @SERISS_EU

SERISS (Synergies for Europe's Research Infrastructures in the Social Sciences) aims to exploit synergies, foster collaboration and develop shared standards between Europe's social science infrastructures in order to better equip these infrastructures to play a major role in addressing Europe's grand societal challenges and ensure that European policymaking is built on a solid base of the highest-quality socio-economic evidence.

The four year project (2015-19) is a collaboration between the three leading European Research Infrastructures in the social sciences – the European Social Survey (ESS ERIC), the Survey of Health Ageing and Retirement in Europe (SHARE ERIC) and the Consortium of European Social Science Data Archives (CESSDA AS) – and organisations representing the Generations and Gender Programme (GGP), European Values Study (EVS) and the WageIndicator Survey.

Work focuses on three key areas: Addressing key challenges for cross-national data collection, breaking down barriers between social science infrastructures and embracing the future of the social sciences.

Please cite this deliverable as: Belloni, M., Tijdens, K.G. (2017) Occupation > industry predictions for measuring industry in surveys. Deliverable 8.11 of the SERISS project funded under the *European Union's Horizon 2020 research and innovation programme* GA No: 654221. Available at: www.seriss.eu/resources/deliverables

Table of content

Executive summary

Many questionnaires have a question “Please write the main business activity of the organisation where you work”. The answer is commonly asked as an open text field, challenging the survey holder to code the response into an industry classification. Alternatively, in web surveys respondents can self-identify their industry from a database. Task 8.4 in SERISS includes two deliverables, D8.10 and D8.11. For D8.10 a database of 321 industry names was developed and translated for use in 99 countries, all coded in 3- or 4-digits according to the classification NACE Rev. 2. The database facilitates survey respondents to self-identify their industry from this lookup table by either an autosuggest box or a two-level search tree. Concerning D8.11, the WageIndicator web survey shows that respondents tend to skip the question about industry relatively more often compared to other questions, presumably because they judge answering the question as cognitively too demanding. Therefore, for D8.11 an occupation>industry prediction has been developed, providing survey respondents with a limited set of industries, most likely for their occupation. Of course, the limited list of industries, shown to the respondent, always includes an option ‘other’, with the full look-up table shown in the next step.

A multi-national occupation>industry prediction for 4-digit ISCO-08 occupations requires a dataset large enough to include as many countries as possible from among those covered by WP8. Such multi-country datasets do not exist and therefore we decided to merge datasets from several sources. We relied on the most recent waves of ESS and EWCS which use classification structures which are homogeneous and currently in place. In addition to CAPI surveys, we exploited some web surveys, mostly the WageIndicator database. The initial idea to include controls in the predicting equations using auxiliary variables was dropped in favour of a pooled dataset NACE Rev. 2 with valid observations for only two variables: a 4-digit ISCO-08 code and a 2-digit NACE Rev. 2 code. We explored country-differences, but it turned out that the estimated most likely industries were very similar across country groups.

We then estimate a set of linear probability models (LPM) – one for each ISCO code. An LPM is a multiple linear regression model with a binary dependent variable (Wooldridge, 2010) – equal to one if the observation reported that specific ISCO unit group and 0 otherwise; the explanatory variables are given by a full set of dummy variables for the 88 divisions (i.e. 2 digits groups) included in the NACE Rev. 2 code. Estimated coefficients represent marginal effects and can be directly interpreted as a probability that each NACE division is associated with that specific ISCO group. In the deliverable’s **accompanying database *RESULTS_DELIVERABLE_D8.11_V2*** the results of the occupation>industry predictions are included for 4-digit occupational units. From February 2018 the occupation>industry predictions are available as an extension of the occupation tool, see http://surveycodings.org/.

1. Introducing SERISS

Synergies for Europe's Research Infrastructures in the Social Sciences (SERISS) is a four-year project that aims to strengthen and harmonise social science research across Europe (2015-19). Work Package 8 (WP8) of SERISS aims to provide cross-country harmonised, fast, high-quality and cost-effective coding of open ended questions on respondents' occupations, industries and education into international standardized classification systems, and to develop a tool to collect standardized social network information. Occupation, industry, employment status, educational attainment and field of education are core variables in many socio- economic and health surveys. Moreover, the size and intensity of social networks are key variables in social surveys. However, their measurement, especially in a cross-cultural, cross-national and longitudinal context, is cumbersome, not sufficiently standardized and often expensive. This work package takes recent scientific and technological developments as an opportunity to improve this situation in order to improve survey measurement quality and provide cost-effective solutions to Research Infrastructures (SERISS Annex 1, European Commission, 2015).

Building on the current technology and the partners' experiences, WP8 develops a cross-country harmonised, fast, high-quality and cost-effective coding module for the core variables. The module uses a large multi-lingual dictionary with tens of thousands of entries about job titles, industry names, fields of education and training, and employment status categories. Additionally, the module will include country-specific, structured lists of educational qualifications. The module will provide up-to-date codes to classify the variables, using international standardized classification systems. It will facilitate surveys in the ESS, EVS, GGP, SHARE and WageIndicator countries and their associated networks to serve infrastructures reaching out to a global audience, including the most spoken languages outside the EU28 area, notably Russian, Mandarin, Arabic, Hindi and Bahasa. Note that WP 8 covers in total 47 languages servicing 99 countries. For the choice of countries and languages, see Deliverable D8.14 (Tijdens, 2016b).

This report concerns deliverable D8.11 of WP8, part of Task 8.4: "Compile the API-database of industries" (SERISS Annex 1, European Commission, 2015). The responsible partners are the University Ca' Foscari of Venice and the University of Amsterdam (UvA).[1]

2. Introducing SERISS Task 8.4

Task 8.4 consists of two deliverables. D8.10 details the development of an industry database to facilitate the survey question "Please write the main business activity of the organisation where you work". The answer is commonly asked as an open text field, challenging the survey holder to code the response into an industry classification after the fieldwork. Alternatively, in web surveys respondents can self-identify their industry from a database (or lookup table). Departing from a database coded NACE Rev. 2[2] and used in the WageIndicator web survey, for D8.10 the multilingual database was extended from 32 to 47 languages, for use in 99 countries (Tijdens, 2016a). The database provides 321 entries, all coded in 3- or on 4-digits NACE Rev. 2. In the web survey respondents can self-identify their industry from this lookup table by either an autosuggest box or a two-level search tree.

[1] Three people made up a team to prepare and conduct the predictions for this Deliverable, notably Michele Belloni from University Ca' Foscari of Venice and Kea Tijdens and Alejandro Zerain from the University of Amsterdam/AIAS (SEP 2016 till JUN 2017). Alejandro Zerain was a student assistant recruited specifically for this deliverable.

[2] This is the European Community NACE Statistical Classification of Economic Activities, commonly referred to as NACE.

The original aim of Deliverable D8.11 was “to facilitate the coding of industries, a database of company names will be compiled, thereby extending existing databases (FORBES, FORTUNE, etc.), to be used as synonyms in the semantic matching techniques in the API” (see SERISS ANNEX 1). The reasoning behind this deliverable was that some survey respondents tend to answer the name of the company where they work instead of the industry. They rather tell that they work for Walmart than that they work in a department store or a supermarket. Hence, coding of industries requires a database of company names and their industry classification codes. At the time of writing the SERISS proposal a database of company names for the 99 countries seemed to be a solution to this problem, because this approach was already used in the Dutch web survey of WageIndicator. However, extending this idea to 99 countries encountered a couple of difficulties. First, quite a number of the 99 countries have no databases with company names at all or only available at very high costs. If databases were available, in some cases the NACE Rev. 2 coding or the company size turned out to be absent. Second, any country has tens of thousands companies, if not more. This seems to be a too large number for respondent’s self-identification. In the Dutch web survey this problem was tackled by reducing the company database to the largest companies, but a company size variable was often absent in the available databases with company names. Third, collecting data about respondents’ company names may raise privacy concerns and therefore it is not wise to include them in a freely available survey tool, where the designers of the database have no say over its use. Finally, the data of the WageIndicator web survey revealed another problem. Survey respondents tend to skip the question about industry relatively more often than other questions, presumably because they judge answering the question as cognitively too demanding.

For these reasons, other solutions to facilitate the answering of the survey question on industry had to be explored. Two possibilities have been considered. The first was an occupation>industry prediction thereby reducing the number of entries for the self-selection, the second was the use of an unlikely occupation-industry combinations database providing respondents with an alert if they had selected an unlikely combination. The second option was not viable, given the very large number of unlikely combinations. In web surveys his ran the risk of causing respondents’ anger and thus drop-out of the survey. The choice was made in favour of the occupation>industry prediction.

To prepare such an occupation>industry prediction, a huge dataset is required, given that at 4-digit level the ISCO-08 classification has 436 so-called unit groups and our 3- or 4-digit industry database holds 321 entries. No single multi-country survey provides sufficient observations for such a prediction at this level of detail. Hence, datasets from several surveys had to be merged. Luckily, an increasing number of surveys records occupation at a detailed 4-digit ISCO-08 level. In some datasets this variable is included, for other datasets permission needs to be asked. When coding industry, most surveys do so at NACE Rev. 2 2-digit, which has 88 entries. Merging datasets proved to be a viable option.

An occupation>industry prediction for 4-digit occupations allows the survey holder to reduce substantially the lookup list of industries, thus easing the self-identification of industry from 321 entries to a maximum of preferably less than 10 entries, which are most relevant for the respondent. Of course, the last entry of this list of limited entries should always be an option ‘Other’, routing to the full industry lookup table. Taken all considerations in this section in mind, the content of deliverable D8.11 was changed. This paper describes how the occupation>industry prediction was prepared. The results of the predictions are included in the **accompanying database *RESULTS_DELIVERABLE_D8.11_V2***.

3. The database for the occupation>industry prediction

Preparatory work

Initially the team followed a line of reasoning that the occupation>industry prediction would improve when core socio-economic variables were taken into account, such as gender, age or education, and this section discusses the subsequent dataset selection. In a later stage this approach was considered to be "too predictive", i.e. industry prediction based on other observable variables in the survey could lead to lack of an independent source of heterogeneity for this variable in the survey.

Dataset selection

As said, the occupation>industry prediction for 4-digit ISCO-08 occupations requires a dataset large enough to include as many countries as possible among those included in WP8. Such multi-country datasets do not exist and therefore we decided to merge datasets from several sources. Datasets were selected based on the availability of all of the following variables:

- Occupation: both ISCO 88 and ISCO 08 have been considered (up to 4-digits)
- Industry: NACE v1, v1.1 and v.2 are included (2-digits classification).
- Educational level

Once these requirements were fulfilled, other variables were also considered, albeit with less priority:

- Public Sector: dummy question on whether respondent works in public sector
- Supervisory position: dummy question on whether respondent has a supervisory position
- Number of supervisees
- Firm size: organised in three categories
- Paid work: dummy question on whether respondent has a paid job
- Self-employed: dummy question on whether respondent is self-employed
- Gender
- Year of birth

The datasets were retrieved from the GESIS Data Archive, the UK Data Archive and the Data Archive of the ESS; and they include the following:

- European Social Survey (Rounds 1 to 7)
- European Quality of Life Survey (3rd Wave)
- European Working Conditions Survey (4th and 5th Wave)

Data from the voluntary surveys of the WageIndicator Foundation have also been included in the deliverable dataset. These are available from the IZA data archive in Bonn.

See Appendix A1 for the list of datasets, and Appendix A2 for the countries covered.

None of these surveys include panel data. Therefore, the different waves/rounds from each survey have been treated as different samples.

Two variables containing meta-data have been included for administrative purposes:

- Year of interview (INTYEAR)
- Survey and wave identifier

Harmonisation table of variables and values

The variables aforementioned have been recorded and harmonised for consistency across the included surveys. In some of the datasets collected, variables had to be recoded to define some of the dummy variables (e.g. self-employed, public sector, work, among others). Appendix A3 includes a mapping table showing which variables were used from each survey to define harmonised variables.

Due to the lack of heterogeneous categorisation of educational levels, it was unviable to harmonise the educational level of respondents using ISCED. Instead, two categorical variables have been created for educational level of respondents: one with three categories and one with six. The harmonisation tables for these two variables can be found in Appendix A4.

The final pooled dataset[3]

As stated at the beginning of this section, the initial idea to include controls in the predicting equations using auxiliary socio-demographic variables was dropped in favour of a pooled dataset with valid observations for only two variables: 4-digit ISCO-08 code and a 2-digit NACE Rev. 2 code. Among the CAPI surveys in Appendix A3, these two variables are jointly present in only ESS Wave 6-7 and EWCS Wave 5 (year 2010)[4]. Note that it would have been very useful to additionally use previous waves of these sources of data, in order to increase the number of observations. However, this was not possible because of the change in the classification structure of ISCO and NACE which occurred across these waves. For instance, in Wave 5 of ESS occupations were coded in ISCO 88. There are conversion tables available both for ISCO and NACE, which allow conversion of codes over time. However, these tables do not guarantee a one-to-one correspondence. Thus, we relied on the most recent Waves of ESS and EWCS which use classification structures which are homogeneous and currently in place. In addition to CAPI surveys, we exploited some web surveys, mostly the WageIndicator database (see Appendix A1 for more details).

The final dataset used in the estimation contains 1,114,240 observations. Appendix A1 shows that about 10 percent of the observations come from CAPI surveys while the residual 90 percent consists of voluntary data collected through the WageIndicator web surveys. Appendix A2 reports the frequency of observations by country and by dataset (CAPI vs Web surveys). It shows that our final pooled dataset covers almost all the 99 countries (38 are covered by the CAPI surveys; only 12 are not covered at all).

With this data, we then estimate a set of linear probability models (LPM) – one for each ISCO code. An LPM is a multiple linear regression model with a binary dependent variable (Wooldridge, 2010) –equal to one if the observation reported that specific ISCO unit group and 0 otherwise; the explanatory variables are given by a full set of dummy variables[5] for the 88 divisions (i.e. 2 digits groups) included in the NACE Rev. 2 code. Estimated coefficients represent marginal effects and can be directly interpreted as a probability that each NACE division is associated with that specific ISCO group.

We estimated one model for each unit ISCO 08 group (i.e. 436 models). Moreover, we were able to estimate additional models for a substantial number of 3 and 2 digits ISCO codes as

[3] We cannot include the merged dataset as an accompanying database to Deliverable D8.11. The observations retrieved from the European Quality of Life Survey (EQLS) and the European Working Conditions Survey (EWCS) cannot be delivered due to access restrictions. For both of these surveys, sharing data to third parties is only allowed as long as these are registered users with authorization. The observations retrieved from the European Social Survey and the WageIndicator surveys can only be shared for non-commercial use. As there is no control over downloads, we decided not to prepare an accompanying database, but instead include our mapping tables.

[4] We exploited a restricted version of EWCS Wave 5, which provides ISCO unit groups. The public release of the data only reports ISCO at 2 digit level (sub-major groups).

[5] We excluded the constant for convenience.

well as for all 1 digit ISCO major groups. Details on the estimated groups are presented in Appendix A5.

Before turning to the overall results, the next section discusses the question of whether occupation>industry predictions are country-specific, or whether they point to global patterns.

4. The occupation>industry prediction

Do occupation>industry predictions point to global patterns?

As said earlier, not all countries are covered by the data at our disposal. Moreover, those that were covered do not always have a high frequency of observations. Therefore, a country specific estimation for each ISCO08 4 digit unit group would not be possible. It could be possible to group similar countries. However, the question arises of how to cluster the countries into groups characterized by a similar occupation>industry relationship. We made a simple exercise: we split the 38 countries included in the CAPI surveys into two groups: the European Union countries and the other countries which supposedly have a different labour market and possibly are characterised by different associations between occupation and industry. We compared the results picking up some ISCO unit groups at random. It turned out that the estimated most likely industries were very similar (although in some case in a different priority order, see the example shown in Appendix A6) among the two groups of countries. In this exercise, we leave out the web surveys because their inclusion would have mixed up group of county differences with the different type of sampling of the two types of datasets (random vs volunteer). Result of this exercise provide some evidence in favour of a global pattern in the relationship between occupation and the most likely industries.

Results of the occupation>industry predictions

In this section we report our findings for one specific ISCO code, similarly to what we did in Appendix A6 – but this time the model has been estimated on the whole dataset (CAPI+WEB surveys). As in Appendix A6, we proceed in the following way: once the model is estimated, we sort the NACE Rev. 2 dummy variables from the highest to the lowest estimated marginal effects. We then only keep statistically significant variables (up to 10 percent significance level, denoted by a single *). This shortlist of NACE Rev. 2 variables represents the most likely industries associated with that specific ISCO unit group. We do not impose any limit on the number of associated industries for each ISCO unit group at this stage. This limit is likely to be imposed later, when these findings will be filled in into the online coding tool for respondents' self-selection of their occupations, called Jobcoder.

In the example given in Table 1 (unit group 2310 - University and higher education teachers), two NACE industries clearly emerge: There is a 32 percent probability that university and higher education teachers work in the industry "Scientific research and development"; there is also a high probability, equal to 12 percent, for the industry "Education". The other industries, although statistically significant, have a much lower coefficient: equivalent to a probability of less than three percent. The long list of significant dummy variables is justified by the high frequency of observations for this unit group (15,813). In this specific case, we could have imposed a one percent significant level to select the most likely industries and still obtained quite a long list. The **accompanying database *RESULTS_DELIVERABLE_D8.11_V2*** includes predictions for all occupations.

Table 1: Estimated most likely industries for the unit ISCO group 2310 (*University and higher education teachers)*

ISCO 08: LABELS	2310 University and higher education teachers	
LABELS	coef	aster
Industry, NACE Rev.2 2-digit = 72, Scientific research and development	0,322	***
Industry, NACE Rev.2 2-digit = 85, Education	0,127	***
Industry, NACE Rev.2 2-digit = 99, Activities of extraterritorial organisations and bodies	0,0244	***
Industry, NACE Rev.2 2-digit = 94, Activities of membership organisations	0,0203	***
Industry, NACE Rev.2 2-digit = 75, Veterinary activities	0,0166	***
Industry, NACE Rev.2 2-digit = 3, Fishing and aquaculture	0,0157	***
Industry, NACE Rev.2 2-digit = 21, Manufacture of basic pharmaceutical products and pharmaceutical preparations	0,0153	***
Industry, NACE Rev.2 2-digit = 91, Libraries, archives, museums and other cultural activities	0,0149	***
Industry, NACE Rev.2 2-digit = 2, Forestry and logging	0,0128	***
Industry, NACE Rev.2 2-digit = 70, Activities of head offices; management consultancy activities	0,0122	***
Industry, NACE Rev.2 2-digit = 90, Creative, arts and entertainment activities	0,0122	***
Industry, NACE Rev.2 2-digit = 73, Advertising and market research	0,0118	***
Industry, NACE Rev.2 2-digit = 84, Public administration and defence; compulsory social security	0,0112	***
Industry, NACE Rev.2 2-digit = 74, Other professional, scientific and technical activities	0,0100	***
Industry, NACE Rev.2 2-digit = 86, Human health activities	0,00838	***
Industry, NACE Rev.2 2-digit = 96, Other personal service activities	0,0072	***
Industry, NACE Rev.2 2-digit = 20, Manufacture of chemicals and chemical products	0,00577	***
Industry, NACE Rev.2 2-digit = 77, Rental and leasing activities	0,00576	**
Industry, NACE Rev.2 2-digit = 79, Travel agency, tour operator and other reservation service a	0,00537	***
Industry, NACE Rev.2 2-digit = 93, Sports activities and amusement and recreation activities	0,00518	***
Industry, NACE Rev.2 2-digit = 8, Other mining and quarrying	0,00486	***
Industry, NACE Rev.2 2-digit = 36, Water collection, treatment and supply	0,00475	**
Industry, NACE Rev.2 2-digit = 7, Mining of metal ores	0,00466	**
Industry, NACE Rev.2 2-digit = 59, Motion picture, video and television programme production, s	0,00439	**
Industry, NACE Rev.2 2-digit = 6, Extraction of crude petroleum and natural gas	0,0043	**
Industry, NACE Rev.2 2-digit = 38, Waste collection, treatment and disposal activities; materials recovery	0,00392	**
Industry, NACE Rev.2 2-digit = 58, Publishing activities	0,00381	***
Industry, NACE Rev.2 2-digit = 9, Mining support service activities	0,00347	*
Industry, NACE Rev.2 2-digit = 69, Legal and accounting activities	0,00337	***
Industry, NACE Rev.2 2-digit = 26, Manufacture of computer, electronic and optical products	0,00335	*
Industry, NACE Rev.2 2-digit = 63, Information service activities	0,00329	***
Industry, NACE Rev.2 2-digit = 88, Social work activities without accommodation	0,00326	***
Industry, NACE Rev.2 2-digit = 71, Architectural and engineering activities; technical testing	0,00307	***
Industry, NACE Rev.2 2-digit = 78, Employment activities	0,00282	**
Industry, NACE Rev.2 2-digit = 32, Other manufacturing	0,0028	***
Industry, NACE Rev.2 2-digit = 87, Residential care activities	0,00275	***
Industry, NACE Rev.2 2-digit = 62, Computer programming, consultancy and related activities	0,00264	***
Industry, NACE Rev.2 2-digit = 66, Activities auxiliary to financial services and insurance act	0,00254	**
Industry, NACE Rev.2 2-digit = 68, Real estate activities	0,0024	**
Industry, NACE Rev.2 2-digit = 10, Manufacture of food products	0,00212	**
Industry, NACE Rev.2 2-digit = 82, Office administrative, office support and other business sup	0,0021	**
Industry, NACE Rev.2 2-digit = 28, Manufacture of machinery and equipment n.e.c.	0,00206	***
Industry, NACE Rev.2 2-digit = 42, Civil engineering	0,00206	**
Industry, NACE Rev.2 2-digit = 18, Printing and reproduction of recorded media	0,00198	*
Industry, NACE Rev.2 2-digit = 27, Manufacture of electrical equipment	0,00197	*
Industry, NACE Rev.2 2-digit = 64, Financial service activities, except insurance and pension f	0,0016	***
Industry, NACE Rev.2 2-digit = 41, Construction of buildings	0,00131	**
Industry, NACE Rev.2 2-digit = 46, Wholesale trade, except of motor vehicles and motorcycles	0,00117	**
Industry, NACE Rev.2 2-digit = 47, Retail trade, except of motor vehicles and motorcycles	0,000863	**
Nobs	1,114,240	
Number obs in group 2310	15,813	

5. How can survey holders use the industry_API?

In the deliverable's **accompanying database *RESULTS_DELIVERABLE_D8.11_V2*** the results of the occupation>industry predictions are included for 4-digit occupational units. From February 2018 the occupation>industry predictions are available as an extension of the occupation tool, see http://surveycodings.org/.

****

6. References

EUROSTAT (2008) *NACE Rev. 2 Statistical classification of economic activities in the European Community.* Luxembourg: Office for Official Publications of the European Communities, ISBN 978-92-79-04741-1

European Commission, Directorate-General For Research & Innovation, Research infrastructure (2015) ANNEX 1 (part A) Research and Innovation action NUMBER — 654221 — SERISS, Brussels

Tijdens, K.G. (2016a) Database of industries + explanatory note. Deliverable D8.10 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221. Available at: www.seriss.eu/resources/deliverables

Tijdens, K.G. (2016b) Survey Q&A + explanatory note. Deliverable D8.14 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221. Available at: www.seriss.eu/resources/deliverables .

Wooldridge, J.M. 2010. Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press.

Appendix

A1: Retrieved surveys and their respective sample size: estimation sample

Questionnaire	Freq.	Percent	Cum.
CAPI SURVEYS			
ESS 6	46,012	4.13	4.13
ESS 7	25,536	2.29	6.42
EWCS 2010 RES	43,589	3.91	10.33
WEB SURVEYS			
monster-empl	4,571	0.41	10.74
t-online	27,346	2.45	13.20
wageindicator	458,926	41.19	54.39
wageindicator-empl	72,136	6.47	60.86
wageindicator-journalists	12	0.00	60.86
wageindicator-self	3,872	0.35	61.21
wageindicator-unem	514	0.05	61.25
wageindicator-wicare	66	0.01	61.26
websurvey2015	3,193	0.29	61.55
wilite	230,931	20.73	82.27
wilite-empl	188,337	16.90	99.17
wilite-self	4,447	0.40	99.57
wilite-unem	3,39	0.30	99.88
wilite201301-empl	7	0.00	99.88
wilite201301-self	2	0.00	99.88
wipaper	1,353	0.12	100.00
Total	1,114,240	100.00	

A2: List of SERISS countries included in the merged dataset and number of observations per country (legend at the bottom of the table)

COUNTRY	ESS-EWCS	WEBSURVEYS	LIST OF COUNTRIES IN WP	N
Albania	1000		Albania	1
Angola		3640	Angola	2
Argentina		43909	Argentina	3
Armenia		7		
Australia		2382	Australia	4
Austria	2644	710	Austria	5
Azerbaijan		8094		
Bangladesh		5		
Belarus		95780	Belarus	6
Belgium	7161	32167	Belgium	7
Benin		168	Benin	8
			Bolivia	9
Botswana		255	Bosnia and Herzegovina	10
Brazil		67604	Brazil	11
Bulgaria	3118	652	Bulgaria	12
Burundi		67	Burundi	13
Cambodia		329	Cambodia	14
Cameroon		7	Cameroon	15
Canada		80	Canada	16
Chile		9217	Chile	17
China		9550	China	18
Colombia		11246	Colombia	19
Costa Rica		1955	Costa Rica	20
Croatia	1098	29	Croatia	21
Cyprus	1932	9	Cyprus	22
Czech Republic	4567	10218	Czech Republic	23
Denmark	4013	1981	Denmark	24
DR Congo		5	Ecuador	25
Egypt		4963	Egypt	26
El Salvador		1055	El Salvador	27
Estonia	5018	258	Estonia	28
Ethiopia		239	Ethiopia	29
Finland	5110	4629	Finland	30
France	6475	2262	France	31
Georgia		14		
Germany	7627	133378	Germany	32
Ghana		1898	Ghana	33

COUNTRY	ESS-EWCS	WEBSURVEYS
Greece	1037	44
Guatemala		1838
Guinea		66
Honduras		990
Hungary	2654	3792
Iceland	732	
India		50514
Indonesia		41257
Ireland	5397	288
Israel	2200	
Italy	2223	7246
Kazakhstan		49405
Kenya		3359
Kyrgyztan		72
Latvia	989	67
Lebanon		1
Lithuania	2776	74
Luxembourg	998	1131
Macao		1
Macedonia	1075	
Madagascar		895
Malawi		359
Malta	1000	169
Mexico		20554
Moldova		286
Montenegro	1041	3
Mozambique		5277
Namibia		388
Netherlands	4564	139160
Nicaragua		1028
Niger		119
Norway	3805	
Pakistan		2447
Paraguay		2790
Peru		1748
Philippines		1
Poland	4575	427

LIST OF COUNTRIES IN WP	N
Greece	34
Guatemala	35
Guinea	36
Honduras	37
Hungary	38
Iceland	39
India	40
Indonesia	41
Iraq	42
Ireland	43
Israel	44
Italy	45
Japan	46
Kazakhstan	47
Kenya	48
Kosovo	49
Latvia	50
Lithuania	51
Luxembourg	52
Macedonia	53
Madagascar	54
Malawi	55
Malaysia	56
Malta	57
Mexico	58
Moldova	59
Montenegro	60
Mozambique	61
Netherlands	62
New Zealand	63
Nicaragua	64
Nigeria	65
Norway	66
Pakistan	67
Paraguay	68
Peru	69
Philippines	70
Poland	71

COUNTRY	ESS-EWCS	WEBSURVEYS	LIST OF COUNTRIES IN WP	N
Portugal	2854	5820	Portugal	72
Puerto Rico		1		
Romania	1013	484	Romania	73
Russia	2056	23638	Russian Federation	74
Rwanda		224	Rwanda	75
Saudi Arabia		1		
Senegal		1976	Senegal	76
			Serbia	77
Slovakia	2647	6609	Slovak Republic	78
Slovenia	3409	170	Slovenia	79
South Africa		39827	South Africa	80
South Georgia		1		
South Korea		474	South Korea	81
South Sudan		81	South Sudan	82
Spain	2521	18303	Spain	83
Sri Lanka		4108	Suriname	84
Sweden	4435	5460	Sweden	85
Switzerland	2800		Switzerland	86
Tajikistan		4		
Tanzania		1438	Tanzania	87
			Thailand	88
Togo		170	Togo	89
Turkey	2097	74	Turkey	90
Turkmenistan		2		
Uganda		850	Uganda	91
Ukraine	1844	73344	Ukraine	92
United Kingdom	3615	10805	United Kingdom	93
United States		9339	United States of America	94
Uzbekistan		30	Uzbekistan	95
			Venezuela	96
Vietnam		7209	Viet Nam	97
Virgin Islands		35		
Zambia		1608	Zambia	98
Zimbabwe		2456	Zimbabwe	99

LEGEND:

	THE COUNTRY IS IN THE DATA, AND IN THE WP LIST WITH SUFFICIENT OBS (>100)
	THE COUNTRY IS IN THE DATA, BUT NOT IN THE WP LIST
	THE COUNTRY IS IN THE WP LIST, BUT NOT IN THE DATA
	THE COUNTRY IS IN THE WP LIST AND DATA, BUT INSUFF OBS (<100)

A3: Mapping of variables across surveys for harmonisation

	ESS - Wave 1	ESS - Wave 2	ESS - Wave 3	ESS - Wave 4	ESS - Wave 5	ESS - Wave 6	ESS - Wave 7	EQLS - Wave 3 (2012)	EWCS (2005) - Waves 4 &5	WageIndicator
Country of residence	cntry	cntry	cntry	cntry	cntry	cntry	cntry	Y11_Country	countid	COUNTRY
Year of interview	inwyr	inwyr	inwyys	inwyys	inwyys	inwyys	inwyys	Wave	year	SURVEYY
Gender	gndr	gndr	gndr	gndr	gndr	gndr	gndr	Y11_HH2a	y10_hh2_	GENDER
Year of birth	yrbrn	yrbrn	yrbrn	yrbrn	yrbrn	yrbrn	yrbrn	Y11_HH2b (age)	y10_hh_2	yybirth
Paid work	crpdwk	crpdwk	crpdwk	crpdwk	crpdwk	crpdwk	crpdwk	Y11_HH2d	y10_hh_3	CONTST
Self-employed if paid work	emplrel	emplrel	emplrel	emplrel	emplrel	emplrel	emplrel	Y11_Q2	y10_q6	CONTST
Education level	eisced	eisced	eisced	eisced	eisced	eisced	eisced	Y11_ISCEDsimple		EDUISCED
ISCO88 (4-digit)	iscoco	iscoco	iscoco	iscoco	iscoco	.		.		
ISCO88 (3-digit)	.	.	.	.	.	.	.	.		
ISCO88 (2-digit)	.	.	.	.	.	.	.	.	y10_is_1	
ISCO08	.	.	.	.	.	isco08	isco08	Y11_Q4		ISCO0804
Code industry NACE 1.1	.	nacer11	nacer11	nacer11	.	.		.	y10_nace	
Code industry NACE 2.0	.	.	.	.	nacer2	nacer2	nacer2	.	y10_na_1	NACE2004
Code industry classification other	nacer1	.	.	.	.	.		.		
Public sector	.	.	.	tporgwk	tporgwk	tporgwk	tporgwk	Y11_Q6	y10_q10	
Supervisory position (dummy)	jbspv	jbspv	jbspv	jbspv	jbspv	jbspv	jbspv	Y03_Q6	y10_q17	supv0
Number of Supervisees	njbspv	njbspv	njbspv	njbspv	njbspv	njbspv	njbspv	.	y10_q17	supv0
Firm Size	estsz	estsz	estsz	estsz	estsz	estsz	estsz	Y03_Q5	y10_q11	firmsize

A4: Harmonisation of education variables: three and six categories

EDU_6C		EDU_3C		eisced		Y11_ISCEDsimple & EDUISCED[6]	
ISCED 1: Primary Education	1	Primary education	1	ES-ISCED I, less than lower secondary	1	Primary education (ISCED 1)	1
ISCED 2: Lower secondary education	2	Secondary education	2	ES-ISCED II, lower secondary	2	Lower secondary education (ISCED 2)	2
ISCED 3: Upper secondary education	3			ES-ISCED IIIb, lower tier upper secondary	3	Upper secondary education (ISCED 3)	3
				ES-ISCED IIIa, upper tier upper secondary	4		
ISCED 4: Post-secondary including pre-vocational or vocational education	4	Tertiary education	3	ES-ISCED IV, advanced vocational, sub-degree	5	Post-secondary including pre-vocational or vocational education but not tertiary (ISCED 4)	4
ISCED 5: Tertiary education – first level	5			ES-ISCED V1, lower tertiary education, BA level	6	Tertiary education – first level (ISCED 5)	5
ISCED 6: Tertiary education – advanced level	6			ES-ISCED V2, higher tertiary education, >= MA level	7	Tertiary education – advanced level (ISCED 6)	6

[6] EDUISCED is specified up two-digits but only the first level categorization is relevant in our dataset.

A5: List of ISCO 08 codes for which a model has been estimated

ISCO 08 GROUPS	4 digit	3 digit	2 digit	1 digit
number of estimated groups:	**436**	**92**	**34**	**10**
number of groups in the ISCO structure:	436	130	43	10
specific group:				0
		11		
	110			
	210			
	310			
				1
			11	
		111		
	1111			
	1112			
	1113			
	1114			
	1120			
			12	
		121		
	1211			
	1212			
	1213			
	1219			
		122		
	1221			
	1222			
	1223			
			13	
	1311			
	1312			
		132		
	1321			
	1322			
	1323			
	1324			
	1330			
		134		
	1341			
	1342			

ISCO 08 GROUPS	4 digit	3 digit	2 digit	1 digit
	1343			
	1344			
	1345			
	1346			
	1349			
			14	
		141		
	1411			
	1412			
	1420			
		143		
	1431			
	1439			
				2
			21	
	2111			
	2112			
	2113			
	2114			
	2120			
		213		
	2131			
	2132			
	2133			
		214		
	2141			
	2142			
	2143			
	2144			
	2145			
	2146			
	2149			
		215		
	2151			
	2152			
	2153			
		216		
	2161			
	2162			
	2163			
	2164			
	2165			

ISCO 08 GROUPS	4 digit	3 digit	2 digit	1 digit
	2166			
			22	
		221		
	2211			
	2212			
		222		
	2221			
	2222			
	2230			
	2240			
	2250			
		226		
	2261			
	2262			
	2263			
	2264			
	2265			
	2266			
	2267			
	2269			
			23	
	2310			
	2320			
	2330			
		234		
	2341			
	2342			
		235		
	2351			
	2352			
	2353			
	2354			
	2355			
	2356			
	2359			
			24	
		241		
	2411			
	2412			
	2413			
		242		
	2421			

ISCO 08 GROUPS	4 digit	3 digit	2 digit	1 digit
	2422			
	2423			
	2424			
		243		
	2431			
	2432			
	2433			
	2434			
			25	
		251		
	2511			
	2512			
	2513			
	2514			
	2519			
		252		
	2521			
	2522			
	2523			
	2529			
			26	
		261		
	2611			
	2612			
	2619			
		262		
	2621			
	2622			
		263		
	2631			
	2632			
	2633			
	2634			
	2635			
	2636			
	2641			
	2642			
	2643			
		265		
	2651			
	2652			
	2653			

ISCO 08 GROUPS	4 digit	3 digit	2 digit	1 digit
	2654			
	2655			
	2656			
	2659			
				3
			31	
		311		
	3111			
	3112			
	3113			
	3114			
	3115			
	3116			
	3117			
	3118			
	3119			
		312		
	3121			
	2122			
	3123			
		313		
	3131			
	3132			
	3133			
	3134			
	3135			
	3139			
	3141			
	3142			
	3143			
		315		
	3151			
	3152			
	3153			
	3154			
	3155			
			32	
		321		
	3211			
	3212			
	3213			
	3214			

ISCO 08 GROUPS	4 digit	3 digit	2 digit	1 digit
		322		
	3221			
	3222			
	3230			
	3240			
		325		
	3251			
	3252			
	3253			
	3254			
	3255			
	3256			
	3257			
	3258			
	3259			
			33	
		331		
	3311			
	3312			
	3313			
	3314			
	3315			
		332		
	3321			
	3322			
	3323			
	3324			
		333		
	3331			
	3332			
	3333			
	3334			
	3339			
		334		
	3341			
	3342			
	3343			
	3344			
		335		
	3351			
	3352			
	3353			

ISCO 08 GROUPS	4 digit	3 digit	2 digit	1 digit
	3354			
	3355			
	3359			
			34	
		341		
	3411			
	3412			
	3413			
		342		
	3421			
	3422			
	3423			
		343		
	3431			
	3432			
	3433			
	3434			
	3435			
			35	
		351		
	3511			
	3512			
	3513			
	3514			
		352		
	3521			
	3522			
				4
			41	
	4110			
	4120			
		413		
	4131			
	4132			
			42	
		421		
	4211			
	4212			
	4213			
	4214			
		422		
	4221			

ISCO 08 GROUPS	4 digit	3 digit	2 digit	1 digit
	4222			
	4223			
	4224			
	4225			
	4226			
	4227			
	4229			
			43	
		431		
	4311			
	4312			
	4313			
		432		
	4321			
	4322			
	4323			
			44	
		441		
	4411			
	4412			
	4413			
	4414			
	4415			
	4416			
	4419			
				5
			51	
		511		
	5111			
	5112			
	5113			
	5120			
		513		
	5131			
	5132			
		514		
	5141			
	5142			
		515		
	5151			
	5152			
	5153			

ISCO 08 GROUPS	4 digit	3 digit	2 digit	1 digit
	5161			
	5162			
	5163			
	5164			
	5165			
	5169			
			52	
		521		
	5211			
	5212			
		522		
	5221			
	5222			
	5223			
	5230			
		524		
	5241			
	5242			
	5243			
	5244			
	5245			
	5246			
	5249			
			53	
		531		
	5311			
	5312			
		532		
	5321			
	5322			
	5329			
		541		
	5411			
	5412			
	5413			
	5414			
	5419			
				6
			61	
		611		
	6111			
	6112			

ISCO 08 GROUPS	4 digit	3 digit	2 digit	1 digit
	6113			
	6114			
		612		
	6121			
	6122			
	6123			
	6129			
	6130			
	6210			
		622		
	6221			
	6222			
	6223			
	6224			
	6310			
	6320			
	6330			
	6340			
				7
			71	
		711		
	7111			
	7112			
	7113			
	7114			
	7115			
	7119			
		712		
	7121			
	7122			
	7123			
	7124			
	7125			
	7126			
	7127			
		713		
	7131			
	7132			
	7133			
			72	
		721		
	7211			

ISCO 08 GROUPS	4 digit	3 digit	2 digit	1 digit
	7212			
	7213			
	7214			
	7215			
		722		
	7221			
	7222			
	7223			
	7224			
		723		
	7231			
	7232			
	7233			
	7234			
		731		
	7311			
	7312			
	7313			
	7314			
	7315			
	7316			
	7317			
	7318			
	7319			
		732		
	7321			
	7322			
	7323			
			74	
		741		
	7411			
	7412			
	7413			
		742		
	7421			
	7422			
			75	
		751		
	7511			
	7512			
	7513			
	7514			

ISCO 08 GROUPS	4 digit	3 digit	2 digit	1 digit
	7515			
	7516			
		752		
	7521			
	7522			
	7523			
		753		
	7531			
	7532			
	7533			
	7534			
	7535			
	7536			
	7541			
	7542			
	7543			
	7544			
	7549			
				8
			81	
		811		
	8111			
	8112			
	8113			
	8114			
		812		
	8121			
	8122			
	8131			
	8132			
		814		
	8141			
	8142			
	8143			
		815		
	8151			
	8152			
	8153			
	8154			
	8155			
	8156			
	8157			

ISCO 08 GROUPS	4 digit	3 digit	2 digit	1 digit
	8159			
	8160			
	8171			
	8172			
		818		
	8181			
	8182			
	8183			
	8189			
			82	
		821		
	8211			
	8212			
	8219			
			83	
		831		
	8311			
	8312			
		832		
	8321			
	8322			
		833		
	8331			
	8332			
		834		
	8341			
	8342			
	8343			
	8344			
	8350			
				9
			91	
		911		
	9111			
	9112			
	9121			
	9122			
	9123			
	9129			
		921		
	9211			
	9212			

ISCO 08 GROUPS	4 digit	3 digit	2 digit	1 digit
	9213			
	9214			
	9215			
	9216			
			93	
		931		
	9311			
	9312			
	9313			
		932		
	9321			
	9329			
		933		
	9331			
	9332			
	9333			
	9334			
			94	
		941		
	9411			
	9412			
	9510			
	9520			
			96	
		961		
	9611			
	9612			
	9613			
	9621			
	9622			
	9623			
	9624			
	9629			

A6: Most likely industries for the unit ISCO group 3111 (*Chemical and physical science technicians)* - A comparison of results obtained using EU versus NON-EU survey data

ISCO group: Chemical and physical science technicians

EU data			**NON EU data**	coef	aster
Industry, NACE Rev.2 2-digit = 72, Scientific research and development	0.0380	***	Industry, NACE Rev.2 2-digit = 38, Waste collection, treatment and disposal activities; materials recovery	0.0455	***
Industry, NACE Rev.2 2-digit = 20, Manufacture of chemicals and chemical products	0.0343	***	Industry, NACE Rev.2 2-digit = 20, Manufacture of chemicals and chemical products	0.0407	***
Industry, NACE Rev.2 2-digit = 39, Remediation activities and other waste management services	0.0222	***	Industry, NACE Rev.2 2-digit = 72, Scientific research and development	0.0204	***
Industry, NACE Rev.2 2-digit = 9, Mining support service activities	0.0154	***	Industry, NACE Rev.2 2-digit = 21, Manufacture of basic pharmaceutical products and pharmaceutical preparations	0.0149	***
Industry, NACE Rev.2 2-digit = 21, Manufacture of basic pharmaceutical products and pharmaceutical preparations	0.0143	***	Industry, NACE Rev.2 2-digit = 10, Manufacture of food products	0.00806	***
Industry, NACE Rev.2 2-digit = 37, Sewerage	0.0123	***	Industry, NACE Rev.2 2-digit = 35, Electricity, gas, steam and air conditioning supply	0.00645	**
Industry, NACE Rev.2 2-digit = 7, Mining of metal ores	0.00917	***	Industry, NACE Rev.2 2-digit = 86, Human health activities	0.00398	***
Industry, NACE Rev.2 2-digit = 19, Manufacture of coke and refined petroleum products	0.00909	***	Industry, NACE Rev.2 2-digit = 28, Manufacture of machinery and equipment n.e.c.	0.00222	**
Industry, NACE Rev.2 2-digit = 71, Architectural and engineering activities; technical testing	0.00830	***			
Industry, NACE Rev.2 2-digit = 38, Waste collection, treatment and disposal activities; materials recovery	0.00631	***			
Industry, NACE Rev.2 2-digit = 17, Manufacture of paper and paper products	0.00557	***			
Industry, NACE Rev.2 2-digit = 35, Electricity, gas, steam and air conditioning supply	0.00360	***			
Industry, NACE Rev.2 2-digit = 11, Manufacture of beverages	0.00345	*			
Industry, NACE Rev.2 2-digit = 61, Telecommunications	0.00257	**			
Industry, NACE Rev.2 2-digit = 86, Human health activities	0.00235	***			
Industry, NACE Rev.2 2-digit = 10, Manufacture of food products	0.00156	**			
Industry, NACE Rev.2 2-digit = 85, Education	0.00111	***			
Industry, NACE Rev.2 2-digit = 84, Public administration and defence; compulsory social security	0.000742	*			

Note: green represents most likely NACE divisions obtained in both estimation. Orange represents group only obtained using EU data. It must be pointed out that the number of observations is much higher for the sample of EU countries than for the sample of non-EU countries. Therefore, a longer list of NACE divisions is expected for the former.