WORKING PAPER

Combining data from different surveys in analysis: Compatibility of the 2013 European Company Survey and the 2015 European Working Conditions Survey

Contents

Author: Gijs van Houten (Eurofound)

Acknowledgements: The author would like to thank Tsjangis van Oostrom for the preparatory work on the datasets; Franz Eiffe, Alex Lehr, Eszter Sandor, Mathijn Wilkens and Irene Mandl for their advice and feedback; and Cécile Deneys for the editorial support.

Eurofound reference number: WPEF17036

Related projects: European Working Conditions Survey and European Company Survey

The European Foundation for the Improvement of Living and Working Conditions (Eurofound) is a tripartite European Union Agency, whose role is to provide knowledge in the area of social, employment and work-related policies. Eurofound was established in 1975 by Council Regulation (EEC) No. 1365/75 to contribute to the planning and design of better living and working conditions in Europe.

European Foundation for the Improvement of Living and Working Conditions

Telephone: (+353 1) 204 31 00

Email: information@eurofound.europa.eu

Web: www.eurofound.europa.eu

1 Introduction

Eurofound has a keen interest in making optimal use of the data that is collected through the three surveys it regularly carries out. One approach to making more extensive and more efficient use of survey data is to use data from different sources in combined analysis.

This paper focuses on the extent to which data collected in the 2013 European Company Survey (3rd ECS) and the 2015 European Working Conditions Survey (6th EWCS) can be combined in analysis. The 3rd ECS captures workplace practices with regard to work organisation, human resource management, direct employee participation and social dialogue.[1] The 6th EWCS captures a broad range of indicators of the working conditions of both employees and the self-employed.[2] Both are cross-national surveys, based on probability samples, but the ECS targets establishments with at least 10 employees, whereas the EWCS targets workers.

The most ambitious approach to combined analysis is to carry out statistical matching. This aims to integrate two or more datasets characterised by the fact that the different data sets contain information on a set of common variables as well as variables that are not jointly observed (Rodgers, 1984; D'Orazio et al, 2006; Eurostat, 2013). However, a requirement for statistical matching is that the units of analysis are drawn from the same population. Given that the ECS and the EWCS target different populations, the ambition level of this paper is more limited.

This paper proposes an approach in which statistical matches are made on an aggregate level, rather than an individual level. More precisely, the approach involves the identification of key variables.[3] These are variables that are present in both surveys, and can be used to generate aggregate level estimates of variables of interest for the categories of the key variables that are distinguished in both surveys. The aggregate level estimates based on one survey can then be integrated in the micro-level dataset of the other survey. When referring to 'combined analysis' in the remainder of this paper it is this approach – of analysing a micro-dataset into which aggregate level estimates from another survey have been integrated – that is meant.

There are some principle and practical challenges when carrying out analysis on aggregate level data – some easier to overcome than others – that will be discussed throughout this paper.

1.1 Avoiding the ecological fallacy

The ecological fallacy occurs when associations between group level characteristics are assumed to also hold for the individuals that make up those groups (Robinson, 1950; Freedman, 1999). For instance, the (imaginary) finding that sectors with a high proportion of female employees have higher average income levels does not imply that – within those sectors – women earn more than men. The 'ecological correlation' of gender and income on the sectoral level can occur even if within each sector men earn more than women. The implication is that one cannot infer individual level associations from ecological associations. However, it does not imply that aggregate level information is not of interest, but that 'ecological associations' are just that: associations between aggregate level characteristics,

[1] https://www.eurofound.europa.eu/surveys/european-company-surveys/european-company-survey-2013

[2] https://www.eurofound.europa.eu/surveys/european-working-conditions-surveys/sixth-european-working-conditions-survey-2015

[3] 'Key variables' are defined as variables that are present in both datasets, and which can therefore be used for linking records between them. See for instance: https://stats.oecd.org/glossary/detail.asp?ID=6936

which can be interesting in their own right. Schwartz (1999) argues that the ecological fallacy is an issue of validity, and that ecological variables can be very useful, as long as they do not serve as substitutes for individual level variables.

This latter point requires particular attention when operationalising analytical models for combined analysis. Care needs to be taken that aggregate level variables are only used to indicate aggregate level, or 'contextual', effects. If, for example, the research question asks about the effect of gender on income, the gender balance in sectors cannot substitute looking at whether respondents are men or women; however, it can be meaningful because the individual level effect of gender on income might be affected by the sectoral context.

Applying this to the ECS and the EWCS, it can be argued that for many research questions integrating aggregated results from one survey in the other is not particularly useful, because many claims would require an explicit link between the establishment and the worker. There are, however, research questions that refer to contextual effects, which implies that the effects of the aggregate level characteristics are of interest. For instance, workers in sectors where a high proportion of establishments indicate that they have vacancies can be expected to be more confident about their career prospects, regardless of whether or not the establishment they work for has vacancies.

Combined analysis of ECS and EWCS is therefore only useful if there are explicit theoretical expectations about the effect of the context on the dependent variable of interest.

1.2 Using multilevel analysis

The proposed approach to integrate aggregate level estimates from one survey in the micro-level dataset of the other survey ensures that associations are only made between aggregate level 'ecological' characteristics from one survey and individual characteristics from the other survey. This approach has implications for the types of analysis that can be carried out with the resulting data.

The most appropriate way to analyse the effect of macro-level characteristics on micro-level characteristics is multilevel analysis (Snijders and Bosker, 1999; Hox, 2002). The term multilevel refers to the fact that there is a multi-layered, nested structure: for instance, workers work in companies which operate within sectors and within countries. Multilevel analysis allows for the estimation of variance at all of the difference levels, as well as for the estimation of the effects of characteristics of each these levels on an individual level dependent variable.

In this paper multilevel analysis will be used to assess the relevance of different levels of aggregation in terms of the extent to which variance is assigned to them, as well as to estimate the effect of aggregate level and individual characteristics on a dependent variable.

1.3 Research questions

To further illustrate the issues outlined above, as well as the proposed solutions, this paper aims to answer the following questions:

1. *To what extent does the structure of the ECS and the EWCS allow for combined analysis of the data from the two surveys?*
   a) *What aggregate level units can be used?*
   b) *How do the data need to be manipulated to enable this?*
2. *How can the topics that are most suited for combined analysis be identified?*
3. *What could such a combined analysis look like?*

The first question will be addressed in section 2. It requires looking at the design, sample size and sample distribution of each of the surveys and the key variables that are available. Combined analysis is only possible if one or more appropriate units of comparison can be identified, and if the survey data can be aggregated to the level of the selected unit of comparison in a sufficiently reliable way. The second question will be addressed in section 3 by looking at the variation of core variables from the ECS and EWCS across sectors and countries. In the same section the third question will be answered in a discussion of the results of an example of a combined analysis. The example entails integrating aggregate level results from the ECS in a multilevel analysis of the EWCS micro-data, to show the analytical possibility as well as to highlight its short-comings and the questions it raises going forward.

2 Enabling comparison

2.1 Survey design

The ECS and the EWCS are two very different surveys. The ECS is a survey of establishments, looking at the practices organisations put in place in terms of work organisation, human resource management, and direct and indirect employee participation.[4] The EWCS is a survey of workers, which aims to gain in-depth insight into their working conditions. Whereas the differences in design give rise to a range of challenges regarding the feasibility of comparing and integrating results, the substantive focus of the surveys does show a considerable overlap. The workplace practices covered in the ECS can be expected to shape the working conditions that are the focus of the EWCS.

Table 1 shows that the two surveys not only differ in terms of the unit of analysis and the respondents they target, but also in terms of the section of the economy they cover. Because the EWCS targets workers and identifies workers using a very broad definition, it comprehensively covers the economy in terms of establishment sizes (including one-person enterprises) and sectors of activity. The ECS on the other hand applies a size threshold of at least 10 employees, and excludes NACE sectors A (Agriculture), T (Households as employers), and U (Extraterritorial Organisations). Due to challenges with compiling representative and comparable sampling frames for NACE sectors O (Public Administration), P (Education) and Q (Health and social work), the data for these sectors cannot straightforwardly be included in the same analysis as the data for the other sectors.[5] Furthermore, the 6th EWCS covers a few countries that are not covered in the 3rd ECS: Albania, Serbia, Norway and Switzerland. Reversely, the 3rd ECS covers Iceland, which is not covered in the 6th EWCS.

A discrepancy of a different nature is the time of data collection. Data collection for the 3rd ECS took place in the spring of 2013, whereas data for the 6th EWCS were collected in the spring and summer of 2015. This means there is a time lag of over two years which needs to be taken into consideration in the variable selection and design of any combined analysis.

To enable combined analysis, sub-selections of the cases of each of the surveys will need to be used. In terms of countries, only the 28 EU Member States, which are covered in both surveys, are included in the analysis.

4 Establishments are the local units of companies. Most companies only consist of a single unit (single-establishment companies), but larger companies do have multiple local units. The ECS targets single-establishment companies and the local units of multi-establishment companies.

5 The Statistical Classification of Economic Activities in the European Community, or NACE (for the French "nomenclature statistique des activités économiques dans la Communauté européenne"), is the standard classification system for sector of activity used in the European Union. Also see: http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=DSP_GEN_DESC_VIEW_NOHDR&StrNom=NACE_REV2&StrLanguageCode=EN

Table 1: Comparison of the survey design of the 3rd ECS and the 6th EWCS

	3rd ECS	**6th EWCS**
Fieldwork	2013	2015
Target population	Establishments	Workers
Eligibility criteria	At least 10 employees NACE sectors B-S	Aged 15 and over Having worked at least one hour for compensation in the week preceding the survey Living independently (non-institutionalised)
Respondent(s)	Management respondent (most senior person in charge of personnel) Employee representative respondent (most senior member of the employee representative body representing the largest group of workers in the establishment)	Workers as defined above
Survey mode	Telephone	Face-to-face
Interview duration	27 minutes on average for the management respondent 18 minutes on average for the employee representative respondent	45 minutes on average
Country coverage	EU28, MK, ME, IS, TR	EU28, AL, MK, ME, RS, TR, NO, CH
Sampling	Stratified random, oversampling large establishments	Multi-stage, stratified random
Sampling frames	Registers of establishments where available Elsewhere registers of companies	Register based sampling of households or individuals in countries where registers are available Enumeration through random walk elsewhere
Target sample size per country	300-1,650	1,000-3,300
Stratification	Size by sector	Region by urbanisation level
Weighting	Design weight Establishment proportional calibration weight based on size and broad sector compared to distribution of establishments Employee proportional calibration weight based on size and broad sector compared to distribution of employees	Design weight Post-stratification weight, based on distribution of gender, age, region, occupation and sector of economic activity Supranational weight based on relative sampling fraction

Source: 3rd ECS and 6th EWCS, compiled by author

For the EWCS, only those employees that work in establishments with at least 10 employees that carry out so-called 'market activities' (NACE sectors B-N, R, S) are included.[6] Consequently, the total sample size that is available for combined analysis is 24,251 for the 3rd ECS and 24,494 for the 6th EWCS.

2.2 Sample distribution: ensuring sufficiently populated cells

Combined analysis requires that data from one survey is integrated into the dataset of the other survey. Because the unit of analysis is not the same across the two considered surveys, data can only be integrated at an aggregate level, using one or more key variables. This aggregate level information can then be included in the analysis of the micro-level data from the other survey.

Table 2 provides an overview of the shared demographic variables that are included, and of the granularity at which these variables correspond. It shows that there are only few shared variables available.

Table 2: Shared demographic variables in the 3rd ECS and the 6th EWCS

Shared demographic variables		**3rd ECS**	**6th EWCS**	**Correspondence**
Sector of activity[7]	NACE Rev 1	4-digits	4-digits	Up to 514[a] classes
	NACE Rev 2	4-digits	4-digits	Up to 529[b] classes
Workplace size		1 - 0-9 (screened out) 2 - 10-19 3 - 20-49 4 - 50-249 5 - 250-499 6 - 500+	1 - 1 (interviewee works alone) 2 - 2-9 3 - 10-249 4 - 250+	Two matching categories
Geographic location		Country	Country	28 Countries

[a] Total number of classes in the classification[8]

[b] Total number of classes in both datasets

Source: 3rd ECS and 6th EWCS, compiled by author

For sector of activity, both surveys hold very detailed information.[9] In terms of geographic location, the lowest shared unit is country, which is not particularly precise. For workplace

[6] It must be noted that whereas in the ECS the sector of activity is explicitly captured on the level of the establishment, this is not the case for the EWCS. In some cases where employees work in an establishment that has a different NACE classification as the company it belongs to, the NACE classification of the company might be reported.

[7] For more information on the NACE classification see http://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF

[8] The datasets need cleaning up to establish the total number of classes at the NACE Rev 1.1 4-digit level in both datasets. This was not carried out for this exercise, because we will use NACE Rev 2 which is the most recent revision of the classification.

[9] In both surveys the sector of activity was asked for in a very similar open-ended question. The answers to this open-ended questions were coded using very similar coding strategies (for the ECS see: https://www.eurofound.europa.eu/sites/default/files/ef_files/surveys/ecs/2013/documents/ecs2013docs/3rdECS%202013Coding_2.pdf; for the EWCS see:

size, more detail is available in the ECS than for the EWCS. The categories can be collapsed to match in such a way that only two categories remain: 10-249 employees, and 250 employees or more.

When calculating aggregate level estimates, it needs to be ensured that

- the number of cells for which an aggregate level estimate is calculated is sufficiently large to reliably include the variable in a multilevel analysis of micro-level data, and
- the number of cases in each cell is sufficiently large to reliably calculate an aggregate level estimate.

Table 3: Sample size per country, for each survey

	3rd ECS	6th EWCS
Austria	972	530
Belgium	952	1,293
Bulgaria	504	588
Croatia	445	540
Cyprus	451	598
Czech Republic	986	590
Denmark	969	469
Estonia	495	551
Finland	991	562
France	1438	826
Germany	1,458	1,246
Greece	993	497
Hungary	1,015	549
Ireland	487	557
Italy	1,514	765
Latvia	505	491
Lithuania	492	512
Luxembourg	499	597
Malta	280	588
Netherlands	987	507
Poland	1,431	650
Portugal	1,002	526
Romania	502	557
Slovakia	490	515
Slovenia	490	951
Spain	1,428	1,862
Sweden	996	501
United Kingdom	1,479	821

Source: 3rd ECS and 6th EWCS, calculations by author

https://www.eurofound.europa.eu/sites/default/files/ef_survey/field_ef_documents/6th_ewcs_coding_report_for_web_publication.pdf)

There are no hard criteria available for either of these conditions. In their paper on appropriate sample sizes for multilevel analysis Maas and Hox (2005) show that regression coefficients and the variance components are estimated without bias even if the number of groups (higher-level units) is as low as 30 and the number of cases in each group is as low as five, and they argue that this still holds true if the number of groups drops down to 10. However, the more the number of groups drops below 100, the greater the underestimation of the standard errors of the variance components of the higher-level units. For this a low number of micro-level cases that are nested in the higher-level units is not of great concern.

Cell size is, however, a concern when calculating aggregate level estimates. The confidence interval around an estimated percentage can already be as large as ±5 percentage points for a sub-sample of 100 cases, increasing to as much as ±9 percentage points for a sample of 30 cases. For this paper simple estimation will be used to arrive at aggregate level estimates.[10] The aim is therefore that each category of the key variables contains at least 100 cases.

The characteristics of the key variables in Table 2 show that, as long as the underestimation of standard errors is kept in mind, and although more is better, there is no objection to introducing 28 countries or a similar number of sectors of activity as higher-level units in a multilevel analysis.

The key variable workplace size is problematic because it only distinguishes two size classes. The variable can therefore only be used in the construction of aggregate level categories if it was to be combined with one or both of the other two variables. Depending on the research question at hand, this could be something to consider. This paper will focus on the other two key variables: country and sector.

For the key variable country, even though the analyses are limited to a sub-sample (see section 2.1), the net sample in each country is comfortably large (Table 3).

For the key variable sector, the datasets contain information with the highest level of granularity of the NACE classification (4-digits). However, as most of these detailed categories contain (far) fewer than 100 cases, combining these categories into sufficiently large categories would constitute an elaborate and complex exercise.

Therefore, for this paper the NACE Rev. 2 2-digit classification is used as a starting point. The NACE Rev. 2 2-digit classification distinguishes between 88 divisions, 77 of which are available in the sub-sample. Out of those 77 divisions, 22 divisions contained fewer than 100 cases in the ECS, and 24 divisions contained fewer than 100 cases in the EWCS. To address this, these divisions are combined with other similar divisions. In some cases, it was not obvious to find a meaningful combination. In one case (Manufacture of tobacco products) it was decided to exclude the division. In other cases, divisions were retained despite the relatively small cell size (most often in the EWCS). There is also a case where combining divisions to the section level (Mining and quarrying) still did not result in sufficiently large cell sizes in both surveys.

Here as well, the decision was taken not to combine further. Altogether, 63 categories were created, five of which contain fewer than 100 cases in the ECS and nine contain fewer than 100 cases in the EWCS. The resulting categories and sample distributions are displayed in Table 4.

[10] It might, however, be worthwhile to explore more sophisticated estimation techniques in the future. For instance, by using multilevel models to arrive at the aggregate level estimates, it might be possible to achieve sufficiently reliable estimates for much smaller subsamples (Skrondal and Rabe-Hesketh, 2009).

Table 4: Sample size for 63 sectors of activity, for each survey

	3rd ECS	6th EWCS		3rd ECS	6th EWCS
Mining and quarrying	135	*87*	Water and air transport	115	184
Manufacture of food products	1,038	853	Warehousing, transportation and postal activities	564	649
Manufacture of beverages	118	75	Accommodation	785	400
Manufacture of textiles	220	177	Food and beverage service activities	707	1,493
Manufacture of wearing apparel	266	249	Publishing activities	200	130
Manufacture of leather and related products	*95*	*93*	Motion pictures, television, programming and broadcasting activities	126	*95*
Manufacture of wood & of products of wood & cork, except furniture, manufacture of articles of straw & plaiting material	263	269	Telecommunications	*99*	226
Manufacture of paper and paper products	199	*81*	Information technology service activities	516	372
Printing and reproduction of recorded media	206	128	Information service activities	*91*	*94*
Manufacture of chemicals and refined petroleum products	305	199	Financial intermediation, except insurance & pension funding	434	620
Manufacture of basic pharmaceutical products and pharmaceutical preparations	116	154	Insurance, reinsurance & pension funding, except compulsory social security	167	286
Manufacture of rubber and plastic products	417	187	Other financial activities	123	*77*
Manufacture of other non-metallic mineral products	396	178	Real estate activities	360	232
Manufacture of basic metals	288	167	Legal and accounting activities	373	610
Manufacture of fabricated metal products, except machinery and equipment	844	490	Activities of head offices, management consultancy activities	379	136
Manufacture of computer, electronic and optical products	241	147	Architectural & engineering activities, technical testing & analysis	413	253
Manufacture of electrical equipment	336	196	Scientific research & development	182	141
Manufacture of machinery and equipment	678	271	Advertising & market research	210	158

	3rd ECS	6th EWCS		3rd ECS	6th EWCS
Manufacture of motor vehicles and other transport equipment	389	410	Other professional, scientific & technical activities (including veterinary activities)	147	221
Manufacture of furniture	269	198	Rental and leasing activities	*96*	*54*
Other manufacturing	239	219	Employment activities	233	167
Repair and installation of machinery and equipment	213	184	Travel agency, tour operator & other reservation service & related activities	133	114
Electricity, gas, steam and air conditioning supply	259	226	Security & investigation activities	241	362
Water collection, treatment and supply and sewerage	230	104	Services to buildings & landscape activities	532	991
Waste collection and other waste management	302	153	Office administrative, office support & other business support activities	387	350
Construction of buildings	683	908	Creative, arts and entertainment activities	133	208
Civil engineering	427	234	Libraries, archives, museums and other cultural activities	201	147
construction activities	904	1168	Gambling and betting activities	*55*	*59*
Wholesale & retail trade & repair of motor vehicles & motorcycles	747	833	Sports activities and amusement and recreation activities	258	236
Wholesale trade, except of motor vehicles & motorcycles	1,924	865	Activities of membership organizations	230	193
Retail trade, except of motor vehicles & motorcycles	1,782	3,557	Other personal service activities (including computer and household good repair)	187	936
Land transport and transport via pipelines	1,033	1,175			

Italics: fewer than 100 cases

Source: 3rd ECS and 6th EWCS, calculations by author

The decisions to combine categories and/or to retain categories of sub-optimal size are to some extent arbitrary and could benefit from a more thorough review. Moreover, some of the categories are still rather large, for instance, 'Manufacture of food products', Wholesale trade, except of motor vehicles & motorcycles', 'Retail trade, except of motor vehicles & motorcycles', or 'Land transport and transport via pipelines'. Here, greater granularity, and

consequently a closer integration of the datasets, could be achieved by considering the four-digit level.[11]

3 Carrying out combined analysis

The previous section focused on the structural characteristics of the two survey datasets that enable or prohibit combined analysis. This section will look at the explanatory potential of combined analysis and will illustrate what combined analysis can look like.

3.1 Decomposing variance: the relevance of different levels of aggregation

In the introduction it was pointed out that a clear conceptual framework, with resulting expectations with regard to the mechanisms at play, is crucial for meaningful combined analysis. There is, however, an empirical way to identify those variables for which it is worthwhile to think about the aggregate level characteristics that might affect them, which is variance decomposition. This section will show an assessment of the extent to which variance in some of the core variables included in the ECS and the EWCS can be assigned to the country level and to the sector level. If this proportion of variance is limited, there is little promise in exploring country or sector characteristics that could explain this variance.

To decompose the variance, cross-classified multilevel models are applied (Hox, 2002). The models assume that each micro-level unit – establishments in the ECS and workers in the EWCS – is nested in a sector in a country (country*sector) and that these country-sector combinations are nested in both a country and a sector. The analyses will show to what extent variation in the dependent variables is observed at individual level (establishments or workers), at the level of country-sector combinations, at the level of countries and at the level of sectors. The models are 'empty', in the sense that no independent variables are included and only the intercept is estimated, which is the estimated mean of the dependent variable. They therefore only estimate the covariance parameters, which indicate the variance that is allocated to the individual level, the country level, the sector level and the country*sector level.

Table 5: Descriptive statistics of the core variables from the ECS

	N	Minimum	Maximum	Mean	SD
Workplace well-being	24,218	0	100	73.921	17.444
Establishment performance	23,289	0	100	60.322	26.220
Interactive and involving	18,601	0	1	0.125	0.244
Systematic, direct involvement only	18,601	0	1	0.211	0.298
Internally oriented	18,601	0	1	0.162	0.268
Passive management	18,601	0	1	0.205	0.326
Systematic and involving	18,601	0	1	0.157	0.294
Moderate structure and investment in staff	18,601	0	1	0.140	0.257

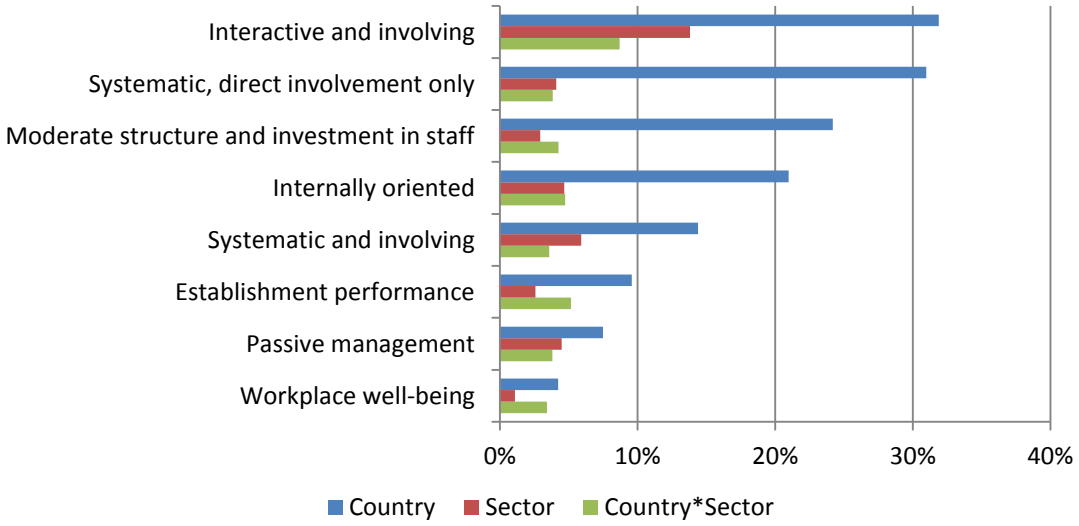
Source: 3rd ECS and 6th EWCS, calculations by author

For the ECS eight dependent variables are selected (see Table 5). The first two are composite indicators; the other six are the posterior probabilities of membership of each of six classes, resulting from a latent class cluster analysis of a wide range of indicators. [12]

[11] Files showing the exact way in which categories were combined can be provided on request.

To allow for comparison between the variables, the variance components are expressed as a percentage of the total variance.

*Figure 1: Proportion of variance assigned to country, sector and country*sector for a selection of core variables in the ECS*[13]



Source: 3rd ECS and 6th EWCS, calculations by author

Figure 1 shows considerable variability between the eight variables in the extent to which variance is assigned to either of the higher levels. With regard to workplace well-being, establishment performance and the posterior probability of being a member of the 'passive management' class around 15% of the total variance is assigned to the higher levels, implying that almost all of the variability should be explained by looking at characteristics of the individual establishments. For the posterior probabilities for membership of most of the other classes a sizeable variance component is assigned to the country level. This suggests that country characteristics are a relevant determinant of the likelihood that establishments have these bundles of workplace practices in place. Sector appears less relevant in this regard, although for the posterior probabilities for membership of the 'interactive and involving' class the variance component of sector is certainly large enough to warrant investigation of the effect of sector characteristics. The variance assigned to the country*sector level is very similar for all variables. It is less than 5% for all variables, apart from the posterior probability of being in the 'Interactive and involving' class, for which it accounts for 9% of the variance.

For the EWCS eight variables are selected as well (see Table 6). These variables are the seven dimensions of job quality that have been developed in recent years, as well as the indicator of mental well-being that is based on items suggested by the World Health Organization.[14]

[12] For more information on the construction and distribution of these variables see the overview report of the 3rd ECS (Eurofound, 2013) and the working paper on associations between employment levels, performance, innovation and workplace practices (Eurofound, forthcoming 2018).

[13] All variance components differ significantly from 0 at $\alpha = .001$, apart from the sector component for Workplace well-being which differs significantly from 0 at $\alpha = .01$, so even though standard errors might be underestimated, the probability of type I errors is extremely low.

Table 6: Descriptive statistics of the core variables from the 6th EWCS

	N	Minimum	Maximum	Mean	SD
Mental well-being (WHO-5)	24,389	0	100	68.770	19.890
Monthly earnings (log)	20,273	-0.07	9.69	7.021	0.668
Skills and discretion	24,427	2	98	53.937	21.824
Social environment	22,417	0	100	78.420	23.091
Physical environment	24,420	0	100	82.760	15.097
Work intensity	24,374	0	100	33.660	18.501
Prospects	24,360	0	100	62.310	19.887
Working time quality	24,429	6	100	70.510	14.358

Source: 3rd ECS and 6th EWCS, calculations by author

Monthly earnings were captured in the EWCS by asking respondents for their net monthly earnings from their main paid job. The answers were first converted into Euros and subsequently into Purchasing Power Parity (PPP). Because income variables tend to be right-skewed, a log transformation was carried out in order to create a variable that approximates a normal distribution. All other indices are composite variables that are constructed such that they run between 0 and 100. Any deviations from this range in Table 6 are due to the fact that a sub-sample is used for this paper. Like in Figure 1, in Figure 2 variance components are expressed as a percentage of the total variance.

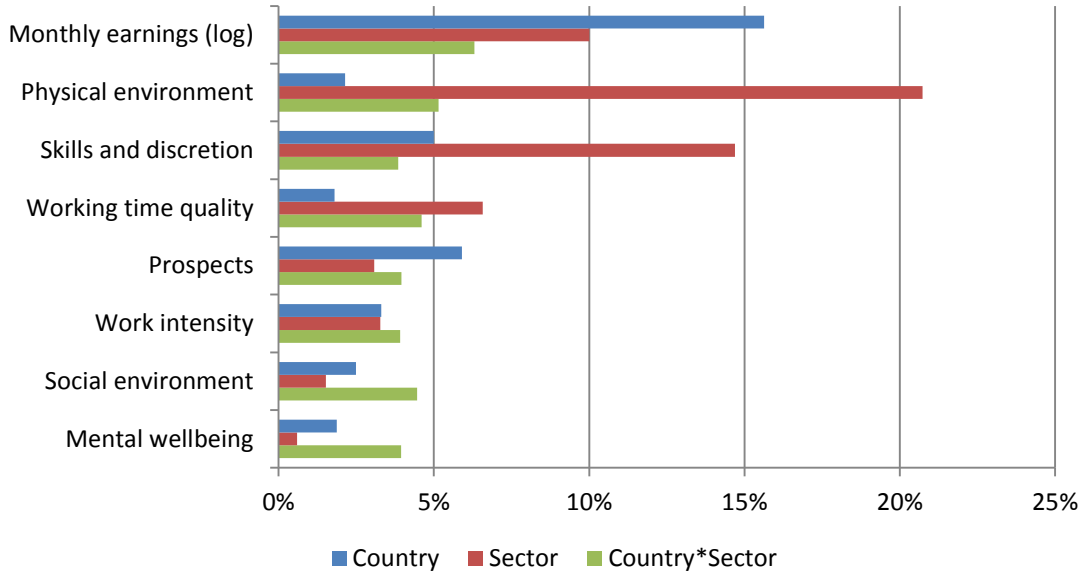
The analysis of the core variables from the EWCS shows a similar pattern as was found for the ECS variables, although overall the variance components of country and sector are somewhat smaller. For work intensity, prospects, working time quality, social environment, and mental well-being less than 15% of the total variance is assigned to the higher levels, implying that almost all of the variability is explained by characteristics of individual workers.

Here it is important to note that there is a potentially relevant intermediate level that – due to lack of data – is not included in the analysis, which is the workplace. For the physical environment and for (the use of) skills and discretion, a sizeable part of the variance is assigned to the level of sectors. For monthly earnings, both the sector and the country are of considerable relevance.

The variance assigned to the country*sector level is very similar for all variables. It is not a very large proportion (around 5%), but for work intensity, social environment and mental well-being, this level is more important than the country and sector levels.

[14] For more information on the construction and distribution of these variables see the overview report of the 6th EWCS (Eurofound, 2017). More specific information on the WHO-5 can be found here: http://www.euro.who.int/__data/assets/pdf_file/0016/130750/E60246.pdf.

*Figure 2: Proportion of variance assigned to country, sector and country*sector for a selection of key variables in the EWCS*[15]



Source: 3rd ECS and 6th EWCS, calculations by author

The results in Figures 2 and 3 indicate that the usefulness of integrating information on country or sector characteristics greatly depends on the dependent variable that is being analysed. For some of the variables analysed, country or sector characteristics do not seem particularly relevant, while for others characteristics of sectors, countries, or both have the potential to significantly improve our explanatory models. Although the country*sector level is moderately relevant for all variables, it might not be possible to integrate aggregate level information from another survey at this level of detail. This most certainly is not possible with the simple approach to estimation that is used for this paper. It might be worth exploring if, and, if so, how, this might be possible using more advance approaches to estimation.

3.2 An example: effects of employee representation on earnings

A simple example has been elaborated to show the different steps that need to be taken for a combined analysis of ECS and EWCS, as well as some indicative results. The example is based on the integration of aggregate level data on employee representation, collected in the ECS, in the micro-data of the EWCS.

The 3rd ECS holds quite a lot of information on social dialogue, whereas the 6th EWCS only asks whether a 'trade union, works council or a similar committee representing employees' exists at the company or organisation the respondent works for. Also, many characteristics of social dialogue do not change very quickly, limiting the extent to which the time lag between the ECS and EWCS is a problem. The use of aggregate level ECS data on social dialogue in multivariate analysis of the EWCS data is therefore promising.

[15] All variance components for sector and for country*sector differ significantly from 0 at $\alpha = .001$ (except for Mental well-being where the sector component differs from 0 at $\alpha = .05$), and all variance components for country differ significantly from 0 at $\alpha = .01$, so, here as well, even though standard errors might be underestimated, the probability of type I errors is extremely low.

The variance decomposition in section 3.1 showed that a substantial amount of the variance in monthly earnings was assigned to the sectoral and country level. It is therefore promising to include sector and country characteristics in a model that predicts monthly earnings.

Conceptually, it can be argued that collective bargaining, and by extension, the presence of bodies for employee representation, have an impact on working conditions beyond the establishment in which bargaining takes place. It could be hypothesised that if bodies for employee representation push earnings up in those organisations where they are present, other organisations in the same sector or in the same country follow suit, because otherwise they would not be able to recruit enough appropriately qualified staff. Obviously, reality is a lot more complicated than this, but for the sake of this example, the following hypotheses could be formulated:

1. *The presence of a body for employee representation at a company has a positive effect on earnings.*
2. *The higher the proportion of establishments with an employee representation in a sector, the higher the earnings in that sector.*
3. *The higher the proportion of establishments with an employee representation in a country, the higher the earnings in that country.*

A combined analysis of the ECS and the EWCS would provide the best available data to test these hypotheses. It would be possible to aggregate the data from the EWCS on the presence of employee representation bodies reported by workers to the level of sectors or countries. However, this would then refer to the proportion of workers working in an establishment with an employee representation, rather than the proportion of establishments. Also, arguably general managers and HR managers are a more reliable source when it comes to the awareness of the presence of employee representation bodies.

As discussed in section 2.2, for this paper simple estimation is used to arrive at aggregate results. Proportions of establishments where an employee representation body was present were calculated for each sector (using the 63 category variable discussed in section 2.2) and for each country using the 3rd ECS dataset. Subsequently these variables were merged into the 6th EWCS dataset.

Table 7: Descriptive statistics of the variables in the example model and of the standard errors of the aggregate level estimates

	N	Minimum	Maximum	Mean	SD
Monthly Earnings (log)	16,676	1.44	9.54	7.022	0.637
Employee representation present at company	16,676	0	1	0.422	0.494
Estimated proportion of establishments with employee representation in the sector	63	0.19	0.65	0.370	0.106
Standard error of sector estimates	63	0.01	0.07	0.030	0.011
Proportion of establishments with employee representation in the country	28	0.08	0.80	0.359	0.199
Standard error of country estimates	28	0.01	0.02	0.016	0.005

Source: 3rd ECS and 6th EWCS, calculations by author

To get an indication of the precision of the estimates, the standard errors were estimated (see Table 7). These show that the estimation of the aggregate scores for country are quite precise, the average standard error is 1.6 percentage points and the maximum standard error only 2

percentage points. This implies that the population mean on average would be no further than ±3.1 percentage points away from the sample mean. Unsurprisingly, the estimates for sector are less precise. The average standard error is 3 percentage points, and the maximum standard error as much as 7 percentage points. So although on average the population mean does not deviate more than ±6 percentage points from the sample mean, for the smallest and most heterogeneous sectors this deviation can be as much as ±14 percentage points.[16] Here the trade-off between creating as precise as possible an aggregate unit and having precise estimates becomes apparent.

Table 7 also shows the distribution of the dependent variable and the independent variable at the micro-level. The N indicates the number of cases on the level at which the distribution was estimated. The dependent variable in this example are earnings, for which, again, the log of monthly earning is used. As mentioned, the EWCS asks whether a trade union, works council or a similar committee representing employees exists at the company or organisation the respondent works for. This is a dichotomous variable which will be used to estimate the fixed effect of the presence of an employee representation body at the establishment.

The model is a cross-classified multilevel model, in which fixed effects are estimated for the presence of an employee representation at the establishment, the proportion of establishments with employee representation in the sector and the proportion of establishments with employee representation in the country. Random effects (covariance parameters) are estimated for country, sector and country*sector. Like in the variance decomposition in section 3.1, it is assumed that workers are nested in country-sector combinations which are nested in both a country and a sector.

Table 8 shows a statistically significant positive effect of the presence of an employee representation at the establishment as well as of the proportion of establishments with employee representation in the country on earnings.

Table 8: Cross-classified multilevel model of employee representation presence on monthly earnings (log)

		Estimate	SE	p	
Fixed effects	Intercept	6.618	0.129	0.000	
	Employee representation present at the establishment	0.215	0.009	0.000	
	Proportion of establishments with employee representation in the sector	0.282	0.234	0.232	
	Proportion of establishments with employee representation in the country	0.679	0.227	0.006	
					Residual variance
Covariance parameters	Residual	0.256	0.003	0.000	70%
	Country	0.052	0.015	0.000	14%
	Sector	0.034	0.007	0.000	9%
	Country*Sector	0.026	0.002	0.000	7%

Source: 3rd ECS and 6th EWCS, calculations by author

The proportion of establishments with employee representation in the sector is not significantly associated with earnings. This implies that, regardless of whether or not there is an employee representation at the establishment, earnings are positively affected by the

[16] Note that these are outliers: 95% of estimates have a standard error of less than 4.91 percentage points and 75% of the estimates have standard error of less than 3.6 percentage points.

proportion of establishments with employee representation in the country. These results support hypotheses 1 and 3.

There are still significant remaining variance components at country and sector level. The proportion of variance assigned to the country level decreased from 16% to 14% after the introduction of the fixed effects. The proportion of variance assigned to the sector level decreased from 10% to 9%. The proportion of residual variance assigned to the country-sector level increased slightly from 6% to 7%. The reduced variance components for sector and country, and the difference between them, imply that the proportion of establishments with employee representation does a better job at explaining country variation in earnings, than it does at explaining sectoral variation in this regard. More generally, it implies that other indicators are needed in order to satisfactorily explain country and sector variation.

This is not the place to reflect on the substantive implications of these findings. The model is extremely simplified and just serves as a case that combined analysis of the ECS allows us to answer questions that could not be answered with each of the individual datasets.

4 Conclusions and discussion

This paper set out to answer three research questions. The first research questions asked to what extent the structure of the ECS and the EWCS allow for combined analysis of the data from the two surveys. More specifically, it asked what aggregate level units can be used and how data need to be manipulated to facilitate this. Sections 2.1 and 2.2 illustrated that there are very important differences between the two surveys, but that there are a small number of key variables that are present in both surveys (country, sector of activity, and workplace size), and that there are substantive areas for which combining data could be of interest. Section 2.2 subsequently showed that the key variables that are present in both surveys, although currently limited, do allow for the identification of meaningful categories (aggregate level units). In this paper data manipulation was geared towards enabling a simple approach to aggregation, which implied that cell sizes of the key variables were kept relatively large. In section 2 it was also shown that there is a loss of cases, due to the fact that the coverage of countries, sectors of activity, and workplace size classes is not the same across the two surveys.

The second research questions asked how the topics that are most suited for combined analysis can be identified. Section 3.1 showed that variance decomposition can be used to assess the relative importance of different levels for the explanation of variation in variables of interest. The proportion of variance that is assigned to levels above the individual can be seen as an indicator of the promise combined analysis holds. The variance decomposition showed that countries, sectors and country-sector combinations account for a substantial amount of the variability of some of the variables that are at the core of the ECS and the EWCS.

The third research question asked what combined analysis could look like. In section 3.2 the results of an analytical model (cross-classified multilevel model) were presented, which illustrated how integrating aggregate level data from the ECS in the analysis of the EWCS can generate meaningful results.

Based on the above it can be concluded that there is value in exploring and exploiting opportunities for combined analysis of the ECS and the EWCS. It has been shown that meaningful matches can be made using the available key variables, and that there is indicative evidence that shows that combined analysis using these key variables can generate meaningful results. However, this paper had an exploratory nature, and many of the analytical decisions may have been fit-for-purpose, but will not fit all purposes.

First of all, this paper was about ‘combined analysis’, which was defined as the analysis of a micro-dataset into which aggregate level estimates from another survey have been integrated.

This is a rather narrow and specific variety of statistical matching. Although some of the lessons learned in this exercise can be useful to further the thinking on the statistical matching of sample surveys, the assumptions that need to be met to allow for the more mainstream approaches to statistical matching are much more restrictive and more work is needed to explore the feasibility and usefulness of matching Eurofound surveys with each other or with other data sources.

Secondly, a simple approach to generating estimates was taken. For this paper means were calculated for all the categories of the key variables. Consequently, restrictions were placed on the size of these categories. It might be possible to relax these restrictions if more advanced approaches for generating estimates are considered. For the collapsing of sectors discussed in section 2.2 a rationale was followed that maximised the number of categories, while still retaining a relative high number of cases in each of these categories in both the ECS and EWCS. Other models can be conceived in which it would be more appropriate to have a slightly lower number of sectors, and an even higher number of cases in each category of the sector variable, as this might enable generating aggregate level estimates on the country*sector level. Reversely, models can be conceived in which the threshold for the number of cases is considerably lower, creating a much larger number of sectors, for which aggregate level estimates are generated using the posterior distributions that can be derived from multilevel models. It must be kept in mind that the way in which categories are collapsed and combined impact on the results from the models that were used for variance decomposition in section 3.1. Although it is unlikely for the broad pattern to change dramatically, the level at which the sectoral (or potentially the regional units) are collapsed will affect the results.

Thirdly, although in section 3.1 the standard errors of the aggregate level estimates are briefly discussed, this paper did not include an assessment of the impact of the fact that the higher level characteristics included in the multilevel analyses in section 3.2 are based on estimates. It might be worth carrying out a simulation study, to assess the robustness of the results of the multilevel models if estimates are allowed to vary across the range as defined by their reliability interval. Such a study could be combined with a study that examines the potential of using multilevel model to generate the aggregate level estimates.

Fourthly, the example in section 3.2 is just that: an example. The dependent variable 'monthly earning' is not the most obvious choice among the topics covered in the EWCS, as many other topics are measured more reliably. Also, the circumstance of availing of information at the individual level from the EWCS and at higher levels from the ECS, while still having a clear rationale for using the ECS for the higher level information, is fairly unique. In most cases, no information is available at individual level. In those cases it would still be worthwhile to run similar models, but the pressure on ensuring conceptual clarity is even greater, ensuring that the contextual and individual level are clearly distinguished. Nevertheless, other examples of possible analysis can be conceived of. Integrating characteristics of social dialogue from the ECS into the EWCS seems to be a promising avenue, but other topics could be explored as well, for instance, the example raised in the introduction of looking at the job security of workers in sectors with different proportions of establishments indicating that they have difficulty finding staff. The reverse is less obvious: it is not easy to think of questions that require the integration of aggregate information from the EWCS into the ECS. Particularly, given that the variance decomposition indicated that workplace practices seem to be determined by country rather than sector characteristics. Here it is likely that the political and business environment in the country is a key factor, rather than the worker characteristics that can be derived from the EWCS.

Fifthly, combined analysis can never substitute a true linking of the ECS and EWCS. As was mentioned, to answer most research questions an explicit link between the worker and the workplace needs to be made. Although contextual information can be derived from the ECS, and although for some variables a considerable amount of variance might be explained by this

contextual information, the often crucial workplace level would still be missing from the analysis.

Finally, and more positively, this paper has revealed the promise of variance decomposition as a tool to gain insight in the structure of the data that are included in the Eurofound surveys. In this paper these findings were used as a mere illustration, and informed decisions on the example analysis. However, the results in Figures 1 and 2 might warrant further attention, as the proportion of variance assigned to each of the levels does not only indicate what level is relevant to include in research efforts, this relevance might extend to policy efforts as well.

References

All Eurofound publications are available at www.eurofound.europa.eu

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

Eurofound (2013), *Third European Company Survey – Overview report: Workplace practices – Patterns, performance and wellbeing*, Publications Office of the European Union, Luxembourg.

Eurofound (2017), *Sixth European Working Conditions Survey – Overview report (2017 update)*, Publications Office of the European Union, Luxembourg.

Eurofound (forthcoming 2018), *Associations of workplace practices, innovation and performance with changes in employment (working title).* Eurofound, Dublin.

Eurostat (2013), *Statistical matching: a model based approach for data integration*, Publications Office of the European Union, Luxembourg.

Freedman, D.A. (1999), 'Ecological Inference and the Ecological Fallacy', *International Encyclopedia of the Social & Behavioral Sciences*, Technical Report No. 549.

Hox, J.J. (2002), *Multilevel Analysis. Techniques and Applications*, Lawrence Erlbaum Associates Publishers, New Jersey.

Maas, C.J.M. & Hox, J.J. (2005), 'Sufficient Sample Sizes for Multilevel Modelling', *Methodology*, 1(3), pp. 86–92.

Robinson, W.S. (1950), 'Ecological correlations and the behaviour of individuals', *American Sociological Review*, 15, pp. 351-357.

Schwarz, S. (1994), 'The Fallacy of the Ecological Fallacy: The Potential Misuse of a Concept and the Consequences', *American Journal of Public Health*, 84(5), pp. 819-824.

Skrondal, A. & Rabe-Hesketh, S. (2009), 'Prediction in multilevel generalized linear models', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172, pp. 659–687.

Snijders, T.A.B., & Bosker, R.J. (1999), *Multilevel Analysis. An introduction to basic and advanced multilevel modelling*, Sage, London.

WPEF17036

The European Foundation for the Improvement of Living and Working Conditions (Eurofound) is a tripartite European Union Agency, whose role is to provide knowledge in the area of social, employment and work-related policies. Eurofound was established in 1975 by Council Regulation (EEC) No. 1365/75, to contribute to the planning and design of better living and working conditions in Europe.