Pertti Taskinen

Register or/and survey – formation of EU variables in Finnish LFS

NSO's, like Statistics Finland, typically have many administrative data sources and data reserves. Some of them can also serve for data collection of the LFS.

The overall goal of this presentation is to show how EU variables are collected in the Finnish LFS, especially, using the existing data reserves, calling them also “register data”. As an objective, the presentation will show how EU variables can be classified according to the way of data collection: (1) by asking questions, (2) from register, or (3) both of them, and ultimately, by (4) reasoning.

In a specific manner, we will introduce how ISCO and NACE information are collected, using both the register information and direct questions. The way of questionnaire design in these specific variables will be shown, as well as results, regarding the comparison between results of the register data and the interview data.

The two main feature of using register data can be found: using them direct as a variable or using them at the questionnaire as pre-filled data. The latter is more complicated way. On the other hand, using the easier way (applying register data directly), only certain variables can be applied direct from register.

In the Finnish LFS, it seems that not much more data from administrative sources can be exploited, speaking of EU variables, although it is obvious that register data is going to be more accurate constantly. There are some obstacles: many of EU-LFS variables are only possible to get direct from the interviewees; or the data from register should be checked, resulting, that it is easier to only ask them shortly.

However, the most interesting task in future is becoming ICSE-18 variable and also existing STAPRO: how can we enhance the use of the existing administrative data.

Recommendations will the most probably be that there is no common rule for exploiting the register data: Advantages and disadvantages must be weighed case by case (variable by variable). A clear advantage for direct use of existing register data is diminishing respondent burden.

As a conclusion, we also see that the LFS data will mostly remain to be a direct survey. It is possible that labour market phenomenon can be interpreted ever better and better by using register information such as Income register. However, that does not fulfill the users need, or, are not in the line with the EU-LFS regulation.

Categorizing variables by source

Microdata containing 116 EU variables are provided to Eurostat on a regular basis, from quarterly to biannual variables (module variables are not included in this count).

The table below indicates the source of data of EU variables in the FI-LFS. Eurostat also uses classification of variables (*variable type*) calling them as Technical, Derived or Collected. Eurostat's breakdown partly coincides with that presented in this paper but is partly different.

Table 1: Number and share of EU variables in FI-LFS by type of data source.

Type of source	**Number of EU var**	**Percent**
Question (Q)	67	58 %
Register (R)	8	7 %
Both (Q and R)	14	12 %
Deductived	27	22 %
Total	116	100 %

Note that low number of register variables indicates that some variables come partly from register (group “Both Q and R”), and deducted variables often are based on register information.

In this paper, we will have examples from each category. The detailed distribution of variables by source is shown in the Annex.

Variables via answers (questionnaire)

Some of data must be collected from the respondents themselves. The number of forced questions (variables WKSTAT, ABSREAS, JATTACH, SEEKWORK, ACTMETNE and AVAILBL) is not high. However, the number of questions is much higher than the amount of these variable alone.

In addition to that, the rest of the questions at the questionnaire form about half of the EU variables. They are acquired by asking because registers or administrative data do not contain them. It is obvious e.g., that REAS-variables are not possible to get any other way than asking.

Also, working time variables are defined so that any administrative data cannot produce such data, for example actual worked hours, absences or working in unsocial hours.

On the other hand, few more variables possibly *could* be derived from administrative sources but are not at least yet. An example is the variable Registration at a public employment service, or REGISTER.

Difficulties to capture REGISTER from existing data reserve include both the reference time and the conceptual difficulty related to the time of payment of the unemployment benefit.

The reference time issue is not a minor challenge regarding REGISTER and some other variables in the LFS.

E.g., in NUMJOB it is the basically the reference week, but Explanatory notes tells that it “should focus on the usual situation around the reference week”. In STAPRO, “the reference period is the current situation” – this does not coincide with the strict reference week rule as in WKSTAT. On the other hand, the use of expression “current situation” is quite understandable in a practical sense of view.

In REGISTER, the reference period is clear “the end of the reference week”, but there may be many kinds of interpretation about receiving the benefit or not at that moment (at the end of the reference week).

Bringing administrative source data onto the questionnaire for the respondent to verify (“is this data wright or wrong”) would not be useful in this REGISTER variable, as it would probably take more time than simply answering yes/no for the question “Are you registered in the public employment service” and so on.

Administrative data (register, data reserves)

Clearly, only few variables are taken directly from administrative sources like SEX or YEARBIR (Year of birth). However, the importance of the register comes out when deducting some more background variable or using them as an alternative source together with the interview.

FI-LFS uses different registers: Statistics Finland's population database, Register of Completed Education and Degrees, and employment statistics. Also are in use: Ministry of Economic Affairs and Employment’s Jobseeker register, and Tax Administration’s Income Register (and membership in a trade union, although not in the EU variables).

Using both register and interview

NACE3D and ISCO4D

We will concentrate on how NACE3D (Economic activity of the local unit for main job) and ISCO4D (Occupation in main job) are collected in the FI-LFS.

NACE is one of the most demanding variable because the local unit must be determined in the first place and then its economic activity. As explanatory notes remarks, "…*respondent can be asked for the name and address of the firm where he/she has his/her main job, if this can be linked to a database of all firms in a country like a Statistical Business Register…*", this has been the traditional way.

Nowadays, FI-LFS uses pre-filled information from Statistics Finland's data of workplaces. An example of question is given here:

"According to the employment statistics, your workplace is:
KONE corporation
Keilasatama 3 [address]
02150 Espoo [post code and city]
Is this information still correct?"
Y/N

If yes, the industry can be easily linked with the business register and no need to bother respondent with more questions. If the answer is no, then detailed questions about workplace will follow.

Consequently, the occupation is given to confirm after the workplace:

"According to employment statistics, your occupation is:
Lift assembler
Is this information still correct?"
Y/N

If yes, the ISCO4D is classified by this information. If answer is no, then respondent must give an occupation, browse the list of occupations and pick it up.

Looking the results (table 2) we will find that relatively high share of register information is not correct, according to the respondents.

Table 2: Correctness of register data in NACE3D and ISCO4D variables according to interviewees, FI-LFS 2023 1st round data collection.

	Workplace (NACE3D)	Occupation (ISCO4D)
Right (yes)	54.3	45.1
Wrong (no)	25.4	18.1
Can't say	0.0	0.0
No data	20.1	36.8

Results are quite understandable because used data reserve are not very accurate – time cap is one to two years. On the other hand, register information especially regarding the occupation does not cover all employed, especially self-employed.

HATLEVEL, HATFIELD, HATYEAR

Another example contains educational attainments (the highest level, field, and the year of successfully completed degree). The basis of data comes from the register of completed degrees for all sampled persons.

Only if no such data exist, questions about degrees are asked from foreign or unknown country born sampled persons. Also, all sampled persons are asked about studies below tertiary education, as a certainty, if there is a lack of information from register. Table 3 shows the distribution between register and interview data.

Table 3: The share of source type (%) for highest degree of education in two ISCED groups, FI-LFS 2023 1st round data collection.

ISCED 100-399		ISCED 440-800	
Source	**Percent**	**Source**	**Percent**
Question	20.6	Question	3.3
Register	79.4	Register	96.7
Total	100.0	Total	100.0

The register data is accepted for the EU data without the respondent checking it (except for the groups and questions mentioned above). We have found that some respondents are not always aware of what counts as a degree and what does not. As a result, it is wiser to rely on register data here than on the respondent's own assessment.

Reasoning (deduction)

Reasoning could also be called “deduction”. For example, we easily know what the variable YEAR is, but it is not the same as a calendar year – so it must be formulated according to the LFS rules. Or we need to do some arrangement how to formulate HHNUM (Serial number of the household), or, to make calculations for COEFFY (Yearly weighting factor).

Furthermore, CITIZENSHIP (Country of main citizenship) obviously comes from register, but there also could be stateless or unknown cases, so the variable is ultimately deducted. Common for these types of variables is that they are deducted either from the direct survey or from administrative data, combining the special rules of the LFS (e.g., what the YEAR is in the LFS definition).

On the other hand, some variables are derived from the responses without asking specifically that variable, like EMPSTAT (Being in employment). MODE (Interviewing mode used) or PROXY (Nature of participation in the survey) are the paradata information from the data collection.

In the sense of respondent burden, derived variables are “good” because they do not increase any burden for interviewees.

Future challenge: STAPRO to ICSE-18

STAPRO will remain as a base variable of professional status, although in the future there will be a new variable “ICSE-18”. It is premature to go to all details in the renewal of professional status, but we already know something how the reform will affect the source data in the FI-LFS.

Today, STAPRO is based directly on the respondent's own assessment of professional status with a direct question:

“Are you:
1. employee
2. self-employed in agriculture (including forestry, horticulture, etc.)
3. self-employed (other than in agriculture)
4. a. own-account worker b. freelancer c. grant recipient
5. unpaid worker on the farm of a family member
6. unpaid worker in a business of a family member?
7. Other”

This method of asking STAPRO question is already becoming obsolete for many of respondent because basically registers tell whether a person receives a salary, i.e., is an employee in most cases (at least, if a person does not have multiple jobs [variable NUMJOB] with different statuses).

On the other hand, the pension contributions of entrepreneurs are also obtained from registers, so the person's entrepreneurial status is often known.

However, since the professional status module of the questionnaire will be renewed in the future due to the ICSE-18, we have not yet made any changes in FI-LFS.

Regarding ICSE-18, the use of register wage data brings additional benefits because it possibly helps to eliminate wage earners from dependent contractors. On the other hand, the corporate form is an essential part of the future ICSE-18 classification. It is possible to make use of by combining the information in the business register with the status of an entrepreneur.

Use of registry is not simple but necessary

The above-mentioned ICSE reform may sound easy, but combining register data with survey data as a pre-fill is hardly going to be technically simple.

Combining of register data with the requirements of direct data collection is not necessarily either completely timesaving. For example, the job data cannot be taken from the register without being checked by the respondent because the data in the employment statistics is often not the same as the job performed by the person at the time of interview.

The period of reference time is not entirely clear in the LFS, and if is, the definition of reference period does not solely fits for the respondent. Questionnaire designers – or register pickers – must sometimes be “creative” to get the data.

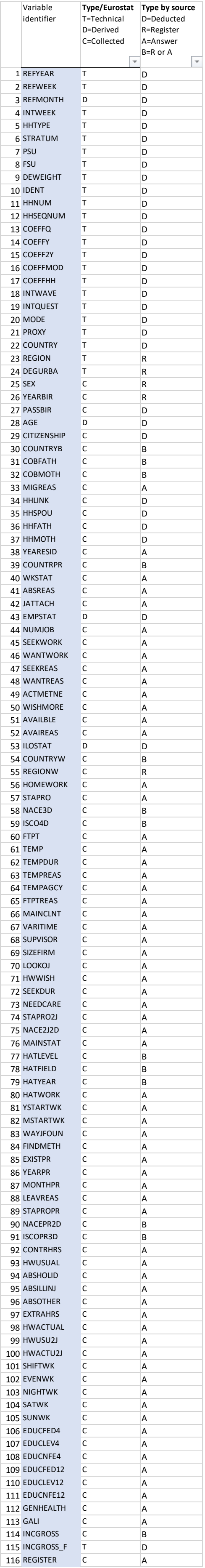
Money savings is not often vast reproducing register instead of answering. E.g., currently, questions related to the variable REGISTER are asked over the telephone about 33,000 times a year. The savings in register picking would be calculated to be about six man-days as interviewer costs.

Combining register information with pre-filled questionnaires is a demanding and error-prone activity, especially when combined with supplementary questions. Recently, this has been done with pensions, and degree register data related to the

8-yearly modules of the Labour Force Survey in 2023. However, the combinations have been made so that we do not collect the data that already exists.

All in all, using register data is necessary today. People are coming more and more aware the existence e.g., of Income register. We should also show for respondents why taking samples and asking about employment –not asking the data which already exist in numerous administrative sources.

APPENDIX: List of EU variables in the FI-LFS by type of source and Eurostat variable type.



	Variable identifier	**Type/Eurostat**
		T=Technical
		D=Derived
		C=Collected
			Type by source
			D=Deducted
			R=Register
			A=Answer
			B=R or A

1	REFYEAR	T	D
2	REFWEEK	T	D
3	REFMONTH	D	D
4	INTWEEK	T	D
5	HHTYPE	T	D
6	STRATUM	T	D
7	PSU	T	D
8	FSU	T	D
9	DEWEIGHT	T	D
10	IDENT	T	D
11	HHNUM	T	D
12	HHSEQNUM	T	D
13	COEFFQ	T	D
14	COEFFY	T	D
15	COEFF2Y	T	D
16	COEFFMOD	T	D
17	COEFFHH	T	D
18	INTWAVE	T	D
19	INTQUEST	T	D
20	MODE	T	D
21	PROXY	T	D
22	COUNTRY	T	D
23	REGION	T	R
24	DEGURBA	T	R
25	SEX	C	R
26	YEARBIR	C	R
27	PASSBIR	C	D
28	AGE	D	D
29	CITIZENSHIP	C	D
30	COUNTRYB	C	B
31	COBFATH	C	B
32	COBMOTH	C	B
33	MIGREAS	C	A
34	HHLINK	C	D
35	HHSPOU	C	D
36	HHFATH	C	D
37	HHMOTH	C	D
38	YEARESID	C	A
39	COUNTRPR	C	B
40	WKSTAT	C	A
41	ABSREAS	C	A
42	JATTACH	C	A
43	EMPSTAT	D	D
44	NUMJOB	C	A
45	SEEKWORK	C	A
46	WANTWORK	C	A
47	SEEKREAS	C	A
48	WANTREAS	C	A
49	ACTMETNE	C	A
50	WISHMORE	C	A
51	AVAILBLE	C	A
52	AVAIREAS	C	A
53	ILOSTAT	D	D
54	COUNTRYW	C	B
55	REGIONW	C	R
56	HOMEWORK	C	A
57	STAPRO	C	A
58	NACE3D	C	B
59	ISCO4D	C	B
60	FTPT	C	A
61	TEMP	C	A
62	TEMPDUR	C	A
63	TEMPREAS	C	A
64	TEMPAGCY	C	A
65	FTPTREAS	C	A
66	MAINCLNT	C	A
67	VARITIME	C	A
68	SUPVISOR	C	A
69	SIZEFIRM	C	A
70	LOOKOJ	C	A
71	HWWISH	C	A
72	SEEKDUR	C	A
73	NEEDCARE	C	A
74	STAPRO2J	C	A
75	NACE2J2D	C	A
76	MAINSTAT	C	A
77	HATLEVEL	C	B
78	HATFIELD	C	B
79	HATYEAR	C	B
80	HATWORK	C	A
81	YSTARTWK	C	A
82	MSTARTWK	C	A
83	WAYJFOUN	C	A
84	FINDMETH	C	A
85	EXISTPR	C	A
86	YEARPR	C	A
87	MONTHPR	C	A
88	LEAVREAS	C	A
89	STAPROPR	C	A
90	NACEPR2D	C	B
91	ISCOPR3D	C	B
92	CONTRHRS	C	A
93	HWUSUAL	C	A
94	ABSHOLID	C	A
95	ABSILLINJ	C	A
96	ABSOTHER	C	A
97	EXTRAHRS	C	A
98	HWACTUAL	C	A
99	HWUSU2J	C	A
100	HWACTU2J	C	A
101	SHIFTWK	C	A
102	EVENWK	C	A
103	NIGHTWK	C	A
104	SATWK	C	A
105	SUNWK	C	A
106	EDUCFED4	C	A
107	EDUCLEV4	C	A
108	EDUCNFE4	C	A
109	EDUCFED12	C	A
110	EDUCLEV12	C	A
111	EDUCNFE12	C	A
112	GENHEALTH	C	A
113	GALI	C	A
114	INCGROSS	C	B
115	INCGROSS_F	T	D
116	REGISTER	C	A