The use of administrative data for weighting (education level) and evaluation of survey data quality

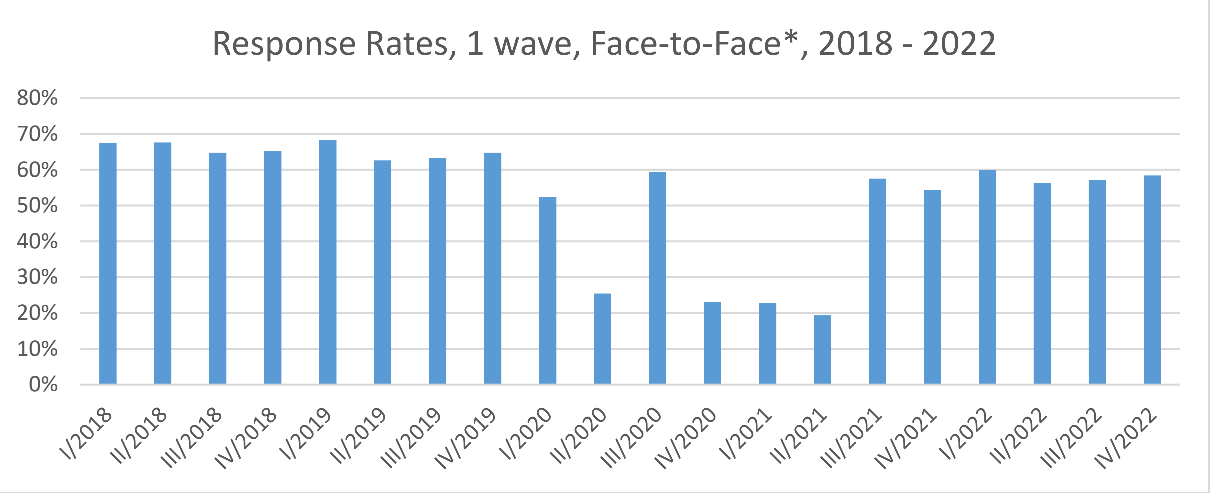
Katja.Rutar@gov.si , Statistical Office of the Republic of Slovenia

The promise of new ESS regulation, relevant for LFS, made us wait with all possible improvements to the date of implementation of ESS. In 2021, when it came in force, we basically renewed the survey. Survey data base was organized that way that all possible administrative variables are added to survey variables.

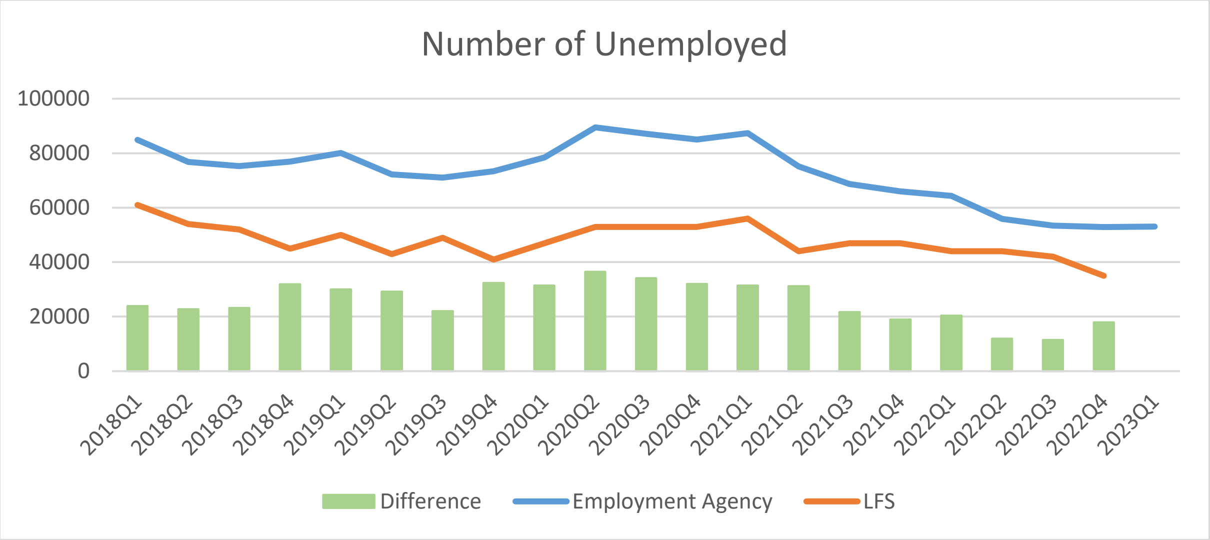
Slovenian Statistical Office has long tradition of collecting administrative demographic data and really many of data. National and EU legislation enables that. In 2011 exclusively register-based Population Census was conducted (similar as in Austria and Belgium). The basis were two main registers (Central Population Register, Real Estate Register) plus additional 18 data sources (employed, unemployed, retired, education, households composition, social transfers, tax data). Unique Personal Identification Code is used in all of them, and it makes merging data easier. For that purpose, in the time before census some improvements were introduced in the administrative sources (e.g., introduction of dwelling number in the Population register). Additionally, some quality checks were done to administrative sources and consequently some improvements were introduced (e.g., checking for very old persons from Central population register with no administrative activities in the country in last years if they are still living in the country). Register-based Population Censuses were repeated in 2015, 2018 and 2021. Data about socioeconomic characteristics (activity status, education, average net income per capita) are prepared and published every year.

Some social surveys, especially surveys of individuals, soon started using those demographic data to lower the burden of respondents in surveys (Consumer Opinion Survey, ICT Usage in Households). There are many reasons, why LFS was more conservative. In our case it is a survey of all household members and when interviewers go to households to collect data, they don't know who is living in the household beside selected person (e.g., year of birth for every member, which guides the questionnaire, can't be known in the first wave). As it is the survey with the biggest sample, we used LFS also to compare survey and administrative data. And as mentioned before, we waited to the introduction of new ESS regulation to modernize the survey to have only one (big) break in time.

Response rates in first wave face-to-face data collection, which fall from above 60 % to below 30 % in corona times and in last quarters rose to above 50 % are reason for skepsis in the LFS data. Assumption, underlying weighting for non-response, that non-respondents are similar to respondents (missing at random), are treated by so high non-participation. It is more probable that those who do not respond have a reason that they do not respond (non-missing at random) and that they are different from those who respond. Data shows, that households with ILO unemployed members in the 1st wave have lover response rates in 2nd wave than households without unemployed members (informative/panel non-response).

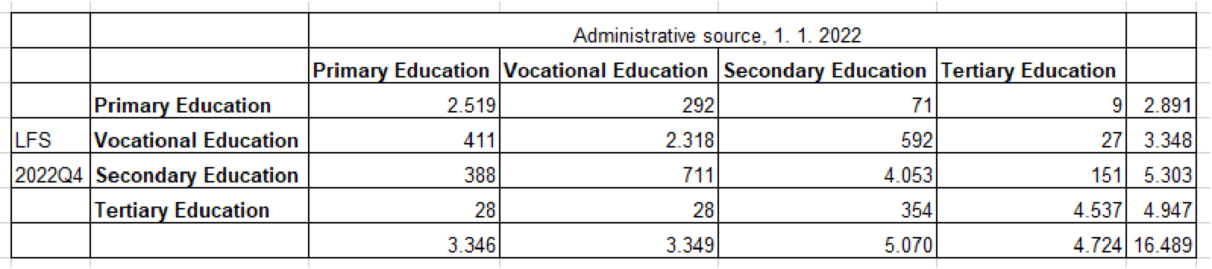


Another reason for skepsis in quality of survey estimates is comparison of LFS estimates of number of unemployed people in the country with Employment Agency number of unemployed. There is quite big difference in level and at the same time also in the direction of movement. The difference is especially visible when the Register Based Population Census data is published (e.g. 2021 – 93.000 vs. 56.000 LFS 2021Q1).



Such findings are a reason to find some possible improvements of estimates with weighting. Administrative data shows that unemployed persons are on average lower educated. Non-response analysis data shows that higher educated persons are more prepared to participate in survey request. Survey estimates for education level in Slovenia from LFS in the past were volatile to some amount. So, education level seems very relevant variable to be used in weighting procedure and we have reliable population totals available once a year.

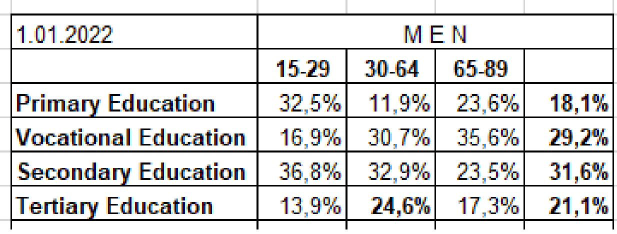
First step was to compare survey data about education level with administrative data for every respondent. We decided to use four education level classes with approximately equal distribution. 81 % of respondents are in the same education class, while 19 % respondents are in different education group in two different sources. For self-reports the correlation is a little higher, but as long as we accept proxy reports (almost half of answers), this finding is irrelevant. Survey data shows lower proportion of respondents with Primary Education and higher proportion of respondents with Tertiary education, comparing to administrative data, but on individual level there are discrepancies in the direction of overestimating and underestimating of education level. There is some time lag between administrative data and reference period, but this can be explanation only for small part of individuals. The question about exact education level is in fact difficult question, also because the classification changed after Bologna reform. Some of respondents are also joking when interviewer is asking them questions for which they are sure that the government has the data (e.g., in which company you work). Majority of inconsistencies are probably measurement errors of different types and from different sources. (Average difference between the end of reference week and the day of interview in our case is two weeks.)



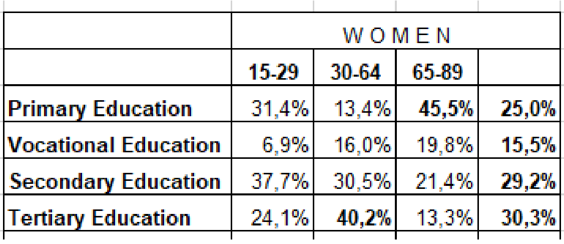
		Administrative source, 1. 1. 2022				
		Primary Education	**Vocational Education**	**Secondary Education**	**Tertiary Education**	
	Primary Education	2.519	292	71	9	2.891
LFS	**Vocational Education**	411	2.318	592	27	3.348
2022Q4	**Secondary Education**	388	711	4.053	151	5.303
	Tertiary Education	28	28	354	4.537	4.947
		3.346	3.349	5.070	4.724	16.489

In the first experiment, we decided to use survey data about education level for weighting.

We know that men and women in different age groups have different education level distribution. Among older population men are better educated, while in younger age groups women are better educated. The definition of weighting classes was the next challenge and we decided for three age groups in combination with four education levels, divided by men and women.



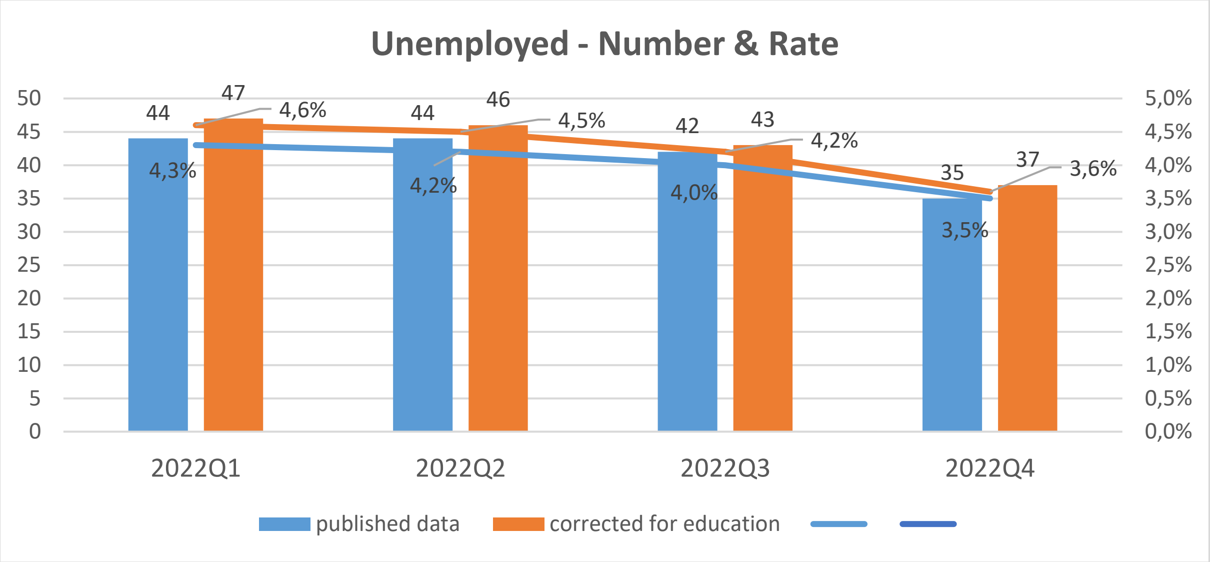
1.01.2022	MEN			
	15-29	**30-64**	**65-89**	
Primary Education	32,5%	11,9%	23,6%	**18,1%**
Vocational Education	16,9%	30,7%	35,6%	**29,2%**
Secondary Education	36,8%	32,9%	23,5%	**31,6%**
Tertiary Education	13,9%	**24,6%**	17,3%	**21,1%**



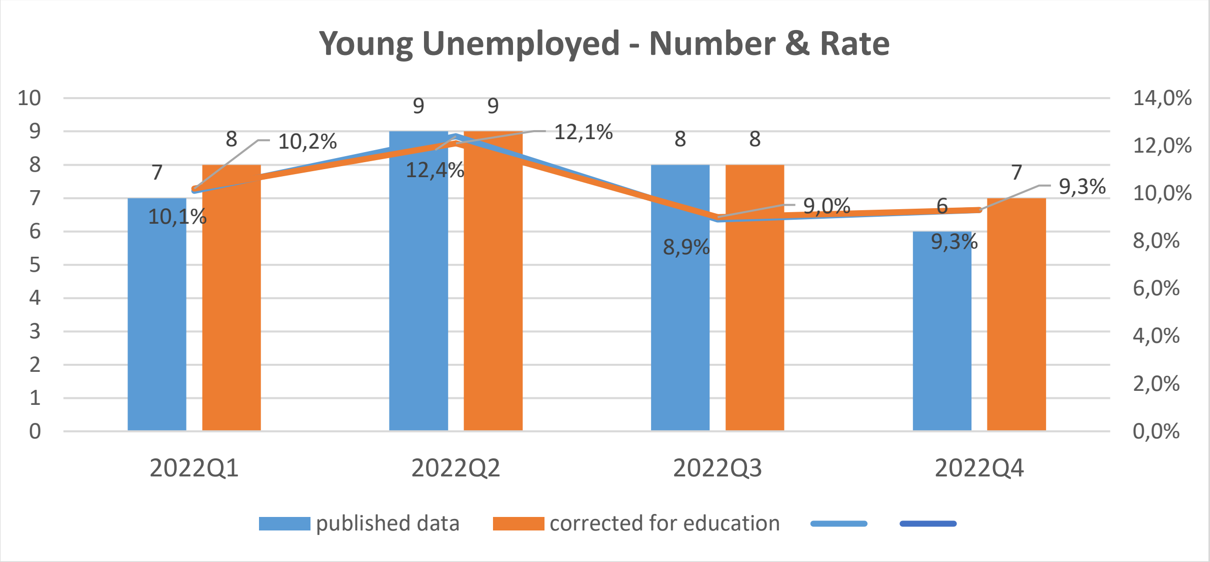
	WOMEN			
	15-29	**30-64**	**65-89**	
Primary Education	31,4%	13,4%	**45,5%**	**25,0%**
Vocational Education	6,9%	16,0%	19,8%	**15,5%**
Secondary Education	37,7%	30,5%	21,4%	**29,2%**
Tertiary Education	24,1%	**40,2%**	13,3%	**30,3%**

In the year 2022 LFS estimate for proportion of people with primary education was 17 %, which is 4 percentage points lower than demographer’s estimate, similar - proportion of people with vocational education level 19 %, which is 4 percentage points lower than demographer’s estimate. Proportion of people with secondary education was 32 % and only 2 percentage points higher than demographer’s estimate, while proportion of people with tertiary education was 32 %, which is 6 percentage points higher than demographer’s estimate. When we look at single quarters, e.g. proportion of tertiary educated population sometimes also decries.

The step with calibration to education level was added to the process of calibration, after calculating sampling and non-response weights. As expected, the consequence is higher unemployment rate (in all quarters of 2021 and 2022, for men and women), but on average only for 0,2 % or in around few thousand individuals.



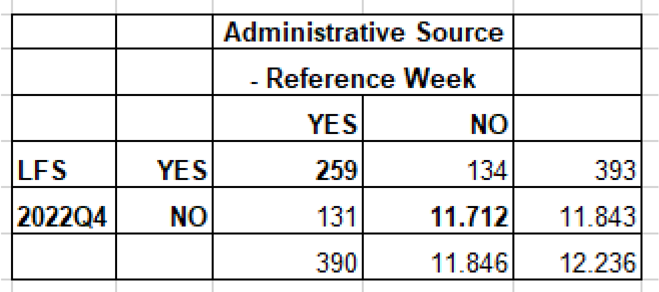
We checked estimates also for small sub-group of respondents – youth, aged between 15 and 24 years (cca. 1500 respondents from which one third active). In this group, the estimate is published with warning (wide confidence intervals, cv > 10%) and the results are probably connected with that fact. The pattern of differences between published estimates and estimates when we include education level in weighting is less regular as for all respondents together or for subgroup of men or woman. Variability of estimates for so small groups (both young and unemployed) is probably the most important reason.



The decision if we will include education level in weighting isn't taken yet. Anyway, the difference between census/demographic/administrative data about the same phenomena would remain considerable.

Besides diminishing response rates also proportion of unemployed people is diminishing in aging European societies (shortage of labour force seems to be more relevant) and it is every year more difficult to measure such rare phenomena with a sample survey.

As the title promises the use of administrative data for evaluation of survey data quality let’s conclude with the comparison of data to the question Are you registered at the Employment Agency?



		Administrative Source		
		- Reference Week		
		YES	NO	
LFS	YES	**259**	134	393
2022Q4	NO	131	**11.712**	11.843
		390	11.846	12.236

Numbers of registered at Employment Agency are very low. But one third of those who responded in the survey that they are registered there in reality are not registered. And vice-versa one third of respondents who according to Employment Agency data were registered there responded that they weren’t registered. In so small groups, every mistake can have big influence (they represent thousands of not-selected and non-responding people). To explain it another way –survey estimate for persons, registered at Employment Agency in fourth quarter last year is 44 thousand while in reality there were 53 individuals registered as unemployed in the same time period. The consequence of this experiment is that we will exclude the questions about registration at the Employment Agency and receiving subsidies for being unemployed from the questionnaire and use administrative data in the survey. (We think this variable is too correlated to WKSTAT to use it for weighting.)

Instead of conclusion we would like to open some issues for discussion. Will we define some minimum threshold for response rate for (official social) surveys under which it is not acceptable to publish the estimates? When majority of sample does not respond, this majority is probably not missing at random. Census and LFS underlie same UN regulation and harmonization rules. Could LFS in visible future in register-oriented countries also be register based? Regarding variable Registration at the Employment Agency – isn’t it to correlated to WKSTAT to use it for weighting?