Using registers in BE-SILC to construct income variables

Eurostat Grant: Action plan for EU-SILC improvements

Version 12/02/2018

1 Introduction

In the context of the modernization of European social statistics in general and the revision of the EU-SILC legal basis in particular, there is a need to change considerably the way of producing EU-SILC data in Belgium. Statistics Belgium signed up for the Eurostat Grant "Action plan for EU-SILC improvements". One of the central pillars in this plan is to intensify the use of registers in BE-SILC. Up till now registers are used for data-collection and methodological purposes, but not yet to deliver data instead of direct questioning, except from some basic demographic information. Several SILC-variables might be substituted by register-information, but this report is restricted to the income variables at individual level, and some at household level.

Information from two tax registers is evaluated: IPCAL, containing the official tax data, and Belcotax, containing the provisional tax data. The Belgian legislator obliges most debtors of income components to send the tax fiche not only to the tax payers themselves to prepare their tax declaration, but also to the tax authorities. These fiches all together constitute Belcotax, and are used as prefill for the online tax declaration with 'tax on web'. Unfortunately, for some income components this is not mandatory but optional. So, tax payers must still add some other types of income during declaration. Consequently, this means that IPCAL is more complete and more correct than Belcotax, as it is supplemented and confirmed by the tax payer. On the other hand, Belcotax is available at the start of the tax declaration (i.e. June income year +1), while IPCAL is only available after processing all tax declarations (i.e. June income year +2). Because of the strict deadlines for data delivery, Belcotax is the only option. The comparison with IPCAL remains important to assess the 'loss' of information due to a provisional tax file instead of a finale file.

SILC data from 2009 until 2014 is analyzed for all concerned PY- and HY- variables separately. All analyses together result in a concrete advise concerning the use of administrative information in the future SILC, and a recalculation of the poverty indicators. However, these recalculations cannot be interpreted as back casted indicators because of two reasons. First, at some points SILC questionnaire information remains necessary. Sometimes we already dispose of this data in the current SILC questionnaires (e.g. illicit work, tax free allowances, missing identification number for social security), sometimes we do not yet (e.g. identification of some PhD. bursaries and work for an international of foreign employer). Based on the results of this exercise, the SILC questionnaires will be modified to deliver the necessary data for the future SILC. Second, in calculating the weights panel attrition is taken into account. One of the variables in these models is the at-risk-of-poverty-status of respondents. As this status might change, the weight calculations might change. These 'new' and 'real' weights are not taken into account in this exercise. Once administrative data will be used to deliver income data, this has to be seen as a break in series and this text serves as information of and argumentation for this break.

In a first step SILC datasets were linked to IPCAL and Belcotax datasets based on the respondents' personal identification number (INSZ). Unfortunately there is no 1-to-1 match between these datasets. SILC samples consist of respondents whose INSZ is unknown, therefore their interview data cannot be match with fiscal data by default. Additionally, SILC also contains respondents who do not have to pay taxes in Belgium, so again fiscal data on their income is by default not available. These remarks were

taken into account during analysis, interpretation and decision making – and will be taken into account when the new questionnaire will be constructed. For each variable SILC and tax definitions are compared and based on Belcotax and IPCAL a reconstruction is made as close as possible to the SILC definition. Below a summary of the results is presented.

2 Summary of the results

Employee cash or near cash income (PY010) is the income variable most frequently reported during SILC interview. It consists of many types of income obtained by an employee status. We were unable to match the SILC definition 100% with Belcotax/IPCAL information, so there are some types of income components exclusively present in SILC (international civil servants, undeclared work, some PhD. bursaries, foreigners) or in tax registers (part of royalties). Overall it concerns not that many respondents, and for decision making we were able to correct for most of these income components. The analysis shows that employee income is more frequently declared in the registers than reported in SILC, even though there are respondents exclusively reporting PY010 in Belcotax/IPCAL and also respondents exclusively reporting PY010 in SILC. Those exclusively reported in registers often concern small forgotten employee incomes. Those exclusively reported in SILC often concern 'misplaced' income components, i.e. other income components (e.g. unemployment benefits) falsely reported as employee income. At the aggregate level, net SILC mean is each year higher than tax mean – only taking respondents in all three datasets into account. Only comparing those in the registers, shows that Belcotax mean is each year higher than IPCAL mean, with differences between 2,2% and 2,7%. At the individual level, median absolute difference score is around 1.600 euro a year, and the mean absolute difference score is 3.500 euro a year. This shows that for most respondents differences are rather small, but there are some extremely high outliers. Differences between IPCAL and Belcotax on the individual level show that most respondents copy Belcotax into IPCAL. Yet, net differences are larger than gross differences because of difficulties in ICPAL with tax calculations. This is a problem that occurred for all PY-variables. Overall, differences between SILC and taxes are acceptable, and we will use Belcotax information in the future.

A second employee income variable is **PY020, referring to benefits in kind**. Analysis showed that only a fraction of the benefits in kind are taxable and present in the tax registers. Luncheon vouchers – the most popular benefits – is not, and will need to be kept in the questionnaire. Additionally, in IPCAL there is no separate code for these benefits in kind, while in Belcotax there is. The company car (PY021) is the most popular taxable benefit, but we concluded to a very poor coverage in Belcotax, compared to SILC. Therefore, we do not prefer to use this tax information in SILC.

Contributions to individual private pension plans (PY035) were analysed as well, even thought, strictly speaking this is not an income variable. Information in Belcotax is only partially complete, as there is no obligation for the banks or insurance companies to provide the tax authorities with this information. Information in IPCAL is only partially compete as well, as tax payers can opt not to receive the tax bonus, and not to declare these contributions. Yet, taking everything together, the results show that tax information is more correct in terms of amounts, and that we only miss a small proportion of

contributors – but also gain some contributors. In the future we will use Belcotax data instead of interview data from SILC.

Cash benefits or losses from self-employment (PY050) is the next employment variable that was analyzed. It consists of many different types of self-employment: business executives, merchants and farmers, professionals, helpers, royalties, and profits and losses from previous self-employment. All types should be fiscally declared with different codes – who were used to reconstruct the SILC definition based on IPCAL – as there is no information in Belcotax. Again there was no 100% match between both definitions possible, but overall differences remain small. Like with employee income, more respondents declare income from self-employment fiscally than they report it during SILC-interview, these are again often misplaced or (mostly small) forgotten self-employment income components. Comparisons on the aggregate level for respondents in both datasets show, like with net employee income, that SILC mean of net self-employment income is each year higher than IPCAL mean. Yet, these differences are much higher now: between 20% and 30%. On the individual level difference are also very high: median absolute difference score is approximately 5.500 euro a year, and the mean is 10.000 euro. The differences at the individual level seem at this point too high to favor the use of registers above direct questioning. Therefore we decided not to use tax information for PY050. An additional factor is the problematic timeliness of IPCAL.

The next PY-variable in the analysis **is pension from individual private pension plan (PY080).** Compared to employee and self-employment income, this is a rather atypical income component. This third pillar pension is normally withdrawn at the age of retirement. In that case taxes are withheld at the source and the capital received should not be fiscally declared. In rare cases people below the retirement age can withdraw their savings. Yet, only in these cases the capital is taxed with income taxation. Analysis of SILC on the one hand and registers on the other hand shows that both cover a different type of beneficiary. On the one hand, most respondents in SILC withdrew at the retirement age and are consequently not present in IPCAL/Belcotax. On the other hand, the younger respondents in IPCAL/Belcotax did not report their third pillar pension in SILC. Consequently, tax information can be used in SILC production to complete the information, but not to impute. Using Belcotax improves the quality but does not allow to suppress questions from the questionnaire.

With the analysis of the **unemployment benefits (PY090)**, we discuss the first social benefit. Again, we need some assumptions and calculation rules to embody SILC's definition. Where for employee income and self-employment income the overall resemblance in number of beneficiaries in SILC and the registers is quite well, there are huge differences for unemployment benefits: each year Belcotax/IPCAL contains more than 1.000 additional beneficiaries. These consists often of small unemployment benefits probably forgotten during SILC-interview. Using tax information would thus improve the coverage substantially. At the aggregate level differences in net unemployment benefits for respondents in both datasets are rather small; but for gross there are practically no difference. Again, this is caused by the problems with tax calculations based on the register information. At the individual level absolute differences in net unemployment benefits are also smaller than those of the variables discussed above. Yet, we need to take into account that overall unemployment benefits are lower than employee and self-employment income. Calculating SILC PY090 based on Belcotax would not only improve the number

of beneficiaries, it would also allow the removal of many questions in the Belgian questionnaire. This will be the way forward for the future.

Next, we studied the **old-age benefits (PY100)**, combining the first and second pillar pensions – and for persons older than the retirement age, also the survivor's benefits. Based on Belcotax/IPCAL we were able to approach the SILC definition, with an exception of a tax-free care allowance, for which SILC information was used to supplement the tax reconstructions. Overall there is a good fit between the beneficiaries in the registers and SILC, but there are more beneficiaries in the tax files. Those probably have forgotten their old-age benefit during SILC-interview. On the aggregate level the net old-age benefits in IPCAL are very close to those in SILC in 2009; yet, in the other five years differences are larger. These differences are mostly caused by the second pillar pensions, i.e. larger – often single – payments that might be easily forgotten or misreported during interview. On the individual level, the differences are for the majority of beneficiaries rather small, and the mean is higher indicating some larger outliers, therefore we will use Belcotax in the future.

The **survivor's benefits (PY110)** are already included in the old age benefits for respondents above the retirement age, but serve as a separate variable for younger respondents. It only consists of a small fraction of the respondents, but overall the coverage of SILC, IPCAL and Belcotax is similar. Also in terms of amounts the differences are rather small, with higher tax means than SILC means; but IPCAL and Belcotax themselves are quasi identical. On the individual level, most respondents have very small difference scores, but there are some outliers as exception. All these results show that we can confidently use Belcotax in the future.

Sickness and disability benefits (PY120 and PY130) were the last social benefits to be analyzed. They were taken together because they share the same codes in IPCAL and Belcotax. Some elements of these variables are non-taxable, and will thus be kept in the questionnaire. Yet, for all others, a relevant tax code was found. In number of beneficiaries, the situation is similar to the unemployment benefits. Using tax information leads to a large increase. Especially small amounts seem to be forgotten during SILC interview. At the aggregate level, differences in mean are comparable to the unemployment results as well. Net tax means were each year higher than SILC means, and vary between 2,5% and 6,8%. Yet differences between both tax files remain minimum in gross terms, and larger in net terms because of the tax calculations. At the individual the results are again the same: for most respondents the difference with SILC is very small, but there are some higher extremes. As a last step, we used the calendar question present (and maintained) in SILC interview to separate between PY120 an PY103. All this together suggest that Belcotax is feasible for the future.

Family allowances (HY050) is the first variable to be discussed at the household level. They consist of several very specific parts (e.g. allowance for maternity leave, paternity leave, parental leave, birth allowance, child support). First of all, the amounts received as sickness or unemployment benefit for maternity, paternity and parental leave are taken from the tax codes in IPCAL and Belcotax using the household composition and the calendar questions where respondents indicate month by month their activity status and income received for the entire income reference period. As such, these amounts were not included in unemployment (PY090) and sickness benefit calculations (PY120). For those we will

use Belcotax in the future. Birth allowances and child support on the other hand are non-taxable, and more complex. For example the birth allowance can be paid in the year preceding the year of birth, or in the years afterwards. Additionally, in split-up and decomposed families the child support is extremely difficult the simulate, as there might be specific arrangements between ex-partners. Therefore we will keep asking questions during SILC interview.

Social exclusion not elsewhere classified (HY060) is a non-taxable income, and therefore IPCAL nor Belcotax can be used. Register information is however available from the public planning service social integration. The analysis showed no 1-to-1-relationship between SILC interview data and the register because in Belgium we have two types of this integration income; one that is nationally organized, and one that is locally organized. Only national help is available in the registers. The comparison showed that there are cases were no such income or a lower income is reported in SILC. In these cases, we can use the register. In other cases, we have only an income reported in SILC (probably local help) or a higher income reported in SILC (combination of national and local help). Therefore, register data will be used together with SILC interview data.

The last variable at the household level is the **income of children aged under 16 (HY110)**. In SILC there are no details, but based on the registers it is clear that the majority of these incomes come from death grants, and that most of them are not reported during SILC interview. Therefore this income will be collected from Belcotax in the future.

The most important question is then about the impact of the changes in data sources on the poverty indicators. Results show that – taking all the fundamental changes with PY010, PY035, PY080, PY090, PY100, PY110, PY120, PY130, HY050, HY060 and HY110 – the impact on the poverty indicators is minimal. AROP is the most important one, and some years the Belcotax construct is closer to SILC than the other years, some years the difference is positive (i.e. higher in Belcotax) and some years negative (i.e. lower in Belcotax). This shows that there is no systematic bias based on the tax registers. Additionally, we see large improvements in the AROP for the unemployed, which shows that adding the 'forgotten' income information from SILC has in fact a substantial impact on the indicators. Even though IPCAL is 'more correct' as it contains the final tax data, analysis shows that the provisional data of Belcotax works as well. Because of the timeliness constraints, Belcotax is our only option. Overall, **these minimal differences prove that we can be confident to use Belcotax in the future SILC for these PY- and HY-variables**. Yet, this implies that the questionnaire should be modified, and that specific filter questions need to be provided to identify case where Belcotax data will not be available. In these situations, all income questions will be kept.

After finalizing the study, a Belcotax analyses on SILC 2015 and 2016 was conducted as well, and results are similar. As was already mentioned, the results presented below cannot serve as back casted results, only as an illustration and argumentation.

	AROP SILC	AROP Belcotax	Δ
2009	14,57%	14,63%	0,06pp
2010	14,59%	15,78%	1,19pp
2011	15,30%	14,97%	-0,33pp
2012	15,29%	14,35%	-0,94pp
*2013	15,06%	15,29%	0,23pp
*2014	15,46%	15,50%	0,04pp
*2015	14,89%	14,65%	-0,24pp
*2016	15,48%	15,13%	-0,35pp

* HY060 was only available for this study until 2012