Comparing income data from survey and register

**Richard Heuberger, Statistics Austria, Vienna**

Since EU-SILC 2012 Statistics Austria has been using mainly register data to calculate household incomes. As a consequence, results for2012 differed significantly from previous years. To understand the effect of register data on the results of EU-SILC, Statistics Austria decided to recalculate the data of EU-SILC 2011 with register data, so that for this survey year we have the opportunity to compare income data and the results of income-based indicators from both sources.

The paper compares income components, the household income and some income-based indicators for EU-SILC 2011. It aims at highlighting the structural differences between register and survey information. The presentation will also discuss implications for income measurement and consequences of the break in the time series for EU-SILC in Austria.

# Introduction

From EU-SILC 2012 onwards register data has been used to calculated income target variables, if adequate register information was available [1]. The motivation to use register information was threefold. First, the aim was to reduce the burden for respondents caused by many detailed and specific questions about the individual income [2]. Second, despite some exceptions, register data seem to deliver higher data quality, particularly register data on incomes [3, 4]. And third, the use of register data simplifies the data editing process and reduces the work on the editing of the data. The use of register data for the data production of EU-SILC in Austria is therefore seen as an advantage.

The use of register data, however, causes a break in the time series in the Austrian EU-SILC data. This break is induced by the change of the mode of data production and is therefore caused by methodological reasons. Changing the fundamental logic of data production inevitably changes the perspective in which the decisive properties of the resulting statistics in terms of e.g. accuracy, precision and comparability. Additionally, as EU-SILC comprises a significant panel component, a break in the time series can be observed on the individual level: at least for some respondents income data from register lead to different income structures compared to the results of the last years’ survey(s). This not only causes concerns about the consistency of the data but also problems for longitudinal research. Finally, a break in the time series can be observed in terms of results of EU-SILC. For example, the at-risk-of-poverty rate, one of the main indicators of interest of this statistic, is 14.4% in EU-SILC 2012, a value more than about 2 percentage points higher than the value of EU-SILC 2011. Taking into account the time series of the at-risk-of-poverty rate in Austria from 2004 onwards it is clear that this increase is not mainly due to disadvantageous societal processes, but mainly due to the methodological change of the data production process.

Evidently, breaks in time series are always a matter of concern, since they hamper the comparability of statistics in a longitudinal perspective. Breaks also question the reliability of the statistics: if results are strongly dependent on the way of measurement, how can one be sure that the statistic accurately represents the real world? Moreover, data from EU-SILC are used in many national and transnational political processes, among them the EU 2020-strategy. Here, a path of target achievement is measured from EU-SILC 2008 onwards. A break in the time series, no matter in which direction, complicates and challenges the measurement of target achievement.

Thus, even though the use of register data for the production of EU-SILC is deemed preferable, the break of the time series causes problems. One solution to avoid these problems would be not to use register data for EU-SILC at all. However, this option is not available for several reasons, and, keeping the data quality in mind, is not preferable. Another solution is to shift the break in the time series historically backwards, that is, to use register data for the waves of EU-SILC prior to EU-SILC 2012 too. Ultimately, this would mean to recalculate the data of EU-SILC from 2004 onwards. This is, however, not feasible ( due to practical reasons), Statistics Austria decided to recalculate until EU-SILC 2008 – an important year given the political process mentioned above.

The recalculation of EU-SILC data with register data is a cumbersome process, even more for four years in a row (and parallel to the everyday work on the ongoing EU-SILC projects). Up until now (May 2014), recalculated data are only available for EU-SILC 2011; data for EU-SILC 2008, 2009 and 2010 will be available in autumn/winter 2014.

The availability of the recalculated data for EU-SILC 2011 is not only valuable with reference to the time series of EU-SILC and the comparability with EU-SILC 2012 data, but also provides the opportunity for comparisons of survey and register data. The two datasets of EU-SILC, thus, can be compared with regard to particular variables and derived indicators, on individual and aggregate level likewise. In the following, the framework of the recalculation of EU-SILC 2011 with register data will be described. The subsequent sections of this paper present some comparisons on the level of income components and on the level of indicators. The final section will discuss the main insights for the data production process and the consequences for the recalculation of EU-SILC 2008 onwards.

# Recalculation of EU-SILC 2011

The recalculation of EU-SILC 2011 demanded the repetition of the data editing process in such a way that for the calculation of income variables register information can be used. Since the structure of the editing process for EU-SILC 2012 was revised, the editing process for the recalculation should mirror the editing process of 2012, though constricted to income variables and adapted to the available register data. Hence, in this section the recalculation of EU-SILC 2011 is described (particularly the link with register data) and differences between register use in EU-SILC 2011 and 2012 are highlighted.

For using register data the link between survey and register data is crucial since legal, practical and methodological limitations may affect this linkage. In Austria, a branch-specific personal identifier bPK (“bereichspezifisches Personenkennzeichen”), introduced with the Austrian e-government strategy in 2004 for the secure communication of personal related data between administrative authorities, allows for a practical, secure and “pseudonymisied” linkage between register and survey data [5]. This PIN is – principally - available for (almost) every person living in Austria, for all registers and therefore also in the sampling frame of EU-SILC, the central population register ZMR (“Zentrales Melderegister”). However, it was not possible to assign a PIN for every person in the net sample on the basis of the gross sample. As a consequence for persons that were surveyed with EU-SILC but who did not acquire a PIN via the gross sample, the PIN has do be obtained from the Ministry of the Interior.

Finding or applying PINs for the link between register and survey data is a twofold process. It necessitates not only that PINs can be applied to sample persons in the survey dataset by that all these survey persons are traceable in the register data (meaning that these persons are also provided with a PIN in the register dataset). In most of the registers used about 5% of the records did not have a PIN, whereas for more than 99% of the persons in the survey of EU-SILC 2011 a PIN can be assigned. Therefore it is not possible to distinguish between cases, i) where there is no match between register and survey data because of a missing PIN and ii) cases, where there is no record for this particular person in the register. Additionally, it can be assumed that the lack of PINs in register data is not random.

For EU-SILC 2011 almost all register data available for EU-SILC 2012 could be used, since most of the datasets are already available for statistical purposes at Statistics Austria. Exceptions were the dataset on student grants, the dataset on pupil allowances and the dataset on childcare benefits. For these income components, where no register information was available for the recalculation of the 2011 data, income information was used from the survey. Another difference between EU-SILC 2012 and EU-SILC 2011 was that due to the fully developed use of register data in EU-SILC 2012 the questionnaire on income was reorganised and redesigned; this included shifting variables from household to personal level of the questionnaire or the other way round and including different questions to account for incomes not included in registers. Albeit these differences, the income variables of the recalculated EU-SILC 2011 are comparable to EU-SILC 2012. Moreover, the recalculated dataset of EU-SILC 2011 offers the opportunity to compare register information and survey responses on the individual and aggregate level and on the level of indicator results.

# Results for income variables

In this section, income variables from register and survey data are compared. Although the aim of this paper is mainly descriptive, one should bear in mind the differences between register and survey information. Answers in surveys are subjected to non-response errors, memory lapses and desirability effects. Due to the sensitive nature of income questions, the aforementioned processes are of particular importance. These effects, conservatively speaking, affect the distance between the surveyed distribution and the “true” distribution of the variable. Register information on the other hand, are mostly designed to fulfil administrative aims. More often than not, this aim is different from the purpose of specifically designed surveys, for which the register information is used – for EU-SILC this would be the available personal or household income. For the assessment of the suitability of register data a critical assessment of the content and scope of the data is necessary. The discussion of the comparison should at least shed some light on the different nature of register and survey information, allowing for a conclusion about data quality in the final chapter of this paper.

The collection of the available household income is one main aim of EU-SILC. EU-SILC is output-harmonised, which means that the variables of the survey (from registers or the questionnaire) are filled into predefined (income) target variables. For the following comparisons both types of variables - detailed and target income variables - are used. Evidently, not all possible variables can be discussed in this paper; the comparisons in this section are restricted to incomes from employment, pension income and the (equivalised household).

Income from employment is the single most important income component, accounting for more than 50% of the total household income. In Austria, the target variable “employee cash or near cash income” also includes income from civilian or military services. Therefore it seems reasonable first to compare only income from employment in the narrow sense, and refer to the more detailed variables. Income from employment is derived from the income tax register.

First, the income distribution from survey and register data from 2011 are compared. The distributions are presented unweighted, since for the register-based income calculation also the weights (using register data) were recalculated and these effects should not be mixed. The median of the register income is somewhat lower, and overall the distribution is more unequal, whereas the survey distribution tends to the middle. The higher number of income recipients in register data indicate a better coverage. But evidently in at least two cases also register data are incomplete: First, clandestine employment or illegal employment and incomes that are not liable to taxation and are not registered (e.g. in the Austrian case: tips). These incomes are not available in register data. Second, incomes from persons living in Austria but working abroad (“Grenzgänger”, commuters across the border) are not registered in the income tax register. These incomes would be available in other tax registers but are not available in due time and/or in desirable quality.

Table 1: Income from employment in EU-SILC 2011 (Questionnaire and register income)



However, due to the higher number of recipients of income from employment (weighted and unweighted) it can be assumed that the coverage with register data is more complete. This raises the question, whether the change in the distribution is due to these additional cases or also caused by those cases that also stated an income in the questionnaire. If the latter is the case, this would signify the effect of the difference in the income measurement.

The difference between register and survey income is not negligible and also depends on the size of the income. In Table 2 the median and the mean of the difference between register and survey income from employment is presented in deciles of the register income. The lower the register income, the lower is the difference between register and survey income. In the lower deciles the difference is actually negative; the corresponding survey incomes are therefore higher than the register incomes. On the other end of the distribution the difference is higher; the survey incomes are lower than the register incomes. Apparently, survey incomes display a tendency towards the middle of the distribution compared to register incomes. This is also true, for example, for pension incomes (PY100) before retirement age[[1]](#footnote-1) and for the equivalised household income. Looking at the correlation between the equivalised household income from register and survey and the difference between these household incomes, the correlation between the register income and the difference is positive, the correlation between the survey income and difference is negative.

Table 2: Difference between questionnaire and survey equivalised household income by decile groups of register income



The difference between register and survey income is not equally distributed among income groups and, maybe even more important, it is not equally distributed among social groups. The following table presents the differences of the median of the equivalised household income between register and survey information by household types. Though these differences are not dramatic, it seems evident that the change of the method influences the picture of social reporting. Admittedly, these findings are based on a cross-sectional view alone. Thus, there might come other influences into play as well (like sampling errors), so that only repetition of measurement (additional years with EU-SILC based on register data) could assure about the trends found here.

Table 3: Median of register and survey equivalised household income by household type



Evidently, differences of the equivalised household income between these two data sources also do have certain effects on (income-based) indicators derived from EU-SILC. Additionally, register data were also used for the recalculation of weights, so that the use of register data also affects non-monetary indicators as well.

# Results for indicators

One of the most important indicators of EU-SILC is the at-risk-of-poverty rate (AROPEE). This indicator is also integrated in the at-risk-of-poverty or social exclusion indicator (AROSE) which is of high political relevance for the Europe 2020 strategy. This latter indicator combines AROPEE with severe material deprivation and the rate of persons living in households with very low work intensity. This chapter will focus on these indicators to discuss the effect of the use of register data for EU-SILC 2011.

On the basis of the distribution of the equivalised household income the effect of the use of register data is not immediately visible: though the register-based household income is distributed somewhat more unequal, the median of the equivalised household income and therefore the at-risk-of-poverty threshold does not change greatly.

Table shows that the register-based AROPE (weighted) is higher by two percentage points. The main reason is the use of register-data or, to be more precise, the change of the distribution at the lower margin of the distribution of the equivalised household income or the more unequal distribution of the register-based income. The increase of the AROPE, then, affects directly the AROSE, which is by more than 2 percentage points higher. The change due to the use of register data for the weighting procedure yet is small, but also increases the AROSE indicator.

Table 4: Dimensions of at-risk-of-poverty or social exclusion – register and survey data



The effect of register data, then, is mainly on the distribution of the equivalised household income alone. The increase of the AROPE is not equally distributed but affects the structure of the AROPE in Austria. An extensive description or interpretation would overstretch the scope of this paper, so only the change by household type is presented.

Table 5: At-risk-of-poverty rate and At-risk-of-poverty rate or social inclusion indicator from survey and register information



Interestingly, the effect of the use of register data seems to be less profound for households with old-age benefits as main source of income. This might be because pension incomes were calculated based on register data in the original dataset of EU-SILC 2011 already– but the register income now is affecting also the calculation of the at-risk-of-poverty threshold.

# Conclusions

The introduction of the use of register data in the data production process is a significant change for EU-SILC in Austria. The use of register data causes a break in the time series of annual EU-SILC survey, be it on the level of income variables, on the level of target variables or income-based indicators and on the individual or aggregate level. The definition of the EU 2020 strategy with EU-SILC 2008 as the starting year demands the recalculation of EU-SILC data from 2008 onwards. The experience with the data of EU-SILC 2011 provides an insight into the differences between register and survey information and allows for the analysis of the underlying explanatory mechanisms. For the recalculation of prior waves (2008 – 2010) these insights can be used for data preparation. Additionally these waves can be used to confirm the current findings.

Differences between register and survey incomes are significant and profound, questioning the quality of income data from surveys – is income information from a survey problematic? There is no easy answer to this question. Deeper understanding of the limits and scope of register data shows that the also register data do not provide a clear, unbiased, untainted view on reality (in this case: the income situation of persons and households). This does not come unexpected since register data are rarely designed to please the expectations of statisticians or social researchers. And maybe there are good reasons for that. may they be practical or epistemological - to believe, that there is no clear, unbiased untainted view on reality in statistic or social science at all. But what can be shown is that there is room for improvement and that there are possibilities to understand the causes and effects of the bias mechanism that blur the view. Though this paper does not exploit the possibilities provided by the existence of two identical datasets – with survey and with register income information -, it may hint to the fruitfulness of researching these differences. If this paper can be an invitation for further research, its main objective is achieved.

# References

[1] Heuberger, R./Glaser, T./Kafka, E. (2013), The use of register data in the Austrian SILC survey. In: Jäntti, M./Törmaälehto, V.-M./Marlier, E., The use of registers in the context of EU-SILC: challenges and opportunities. Eurostat Statistical Working Papers. Luxembourg

[2] Wallgren, A./Wallgren, B. (2007), Register-based Statistics. Administrative Data for Statistical Purposes. Chichester, Wiley

[3] Rendtel, U./Nordberg, L./Jäntti, M./Hanisch, M./Basic, E. (2004), Report on quality of income data, *Chintex Working Paper 21*.

[4] Lohmann, H. (2011), Comparability of EU-SILC survey and register data: The relationship among employment, earnings and poverty, Journal of European Social Policy, 21 (1): 37-54.

[5] Hackl, P. (2009), Using Administrative Data at Statistics Austria: Legal Provisions. Paper presented at the 95th DGINS Conference, 1st October, Malta, MT

1. In EU-SILC 2011 pension income after the legal retirement age were already calculated on the basis of register data [↑](#footnote-ref-1)