**Using administrative registers to improve sampling of EU-SILC
in Austria**

Thomas Glaser, Statistics Austria, thomas.glaser@statistik.gv.at

Johannes Gussenbauer, Statistics Austria, johannes.gussenbauer@statistik.gv.at

Alexander Kowarik, Statistics Austria, alexander.kowarik@statistik.gv.at

**Abstract**

*The following paper describes how administrative data have been introduced for EU‑SILC in Austria with a regard to sampling design. The availability of a variety of data sources from administrative processes that can be matched to the survey by a pseudonymised key has improved measurement of household income as well as weighting for EU‑SILC. Since some administrative data are also available on the level of the corresponding population a further step is to enrich the sampling frame with characteristics that are related to the main indicators of EU-SILC. With this additional information, strata could be constructed that have the potential to allow for a more efficient sample with the expected result of a smaller standard error of the main income related indicators of EU-SILC. For EU‑SILC 2016 and 2017 first results for standard error estimation using these new strata are presented in this paper and conclusions for further improvements are laid out.*

**Keywords:** EU-SILC, administrative data, sampling, stratification, standard error, rich sampling frame

**1. Introduction**

EU-SILC (EU Statistics on Income and Living Conditions) is the main source for data on poverty, social inclusion and income of private households in Austria. Starting with a pilot in 2003, from 2004 onwards a rotational sampling design with four subsamples has been carried out on a yearly basis as a survey with voluntary participation. The first waves of each of the rotational subsamples are drawn from the central residence register (ZMR) by a probability sample. The most relevant indicator based on EU‑SILC is the rate of persons “at risk of poverty or social exclusion” (AROPE) and is surveyed for the Europe 2020 headline target for “lifting at least 20 million people out of the risk of poverty and social exclusion”.[[1]](#footnote-1) AROPE is defined as being affected by at least one of the following conditions: (1) “At-risk-of-poverty” (AROP) describing persons in households with an equivalised household income[[2]](#footnote-2) below 60% of the median of the referring distribution, (2) in a condition of severe material deprivation defined as a lack of resources (encountering at least four out of nine material deprivation items[[3]](#footnote-3)) or (3) living in a household with very low work intensity[[4]](#footnote-4). Among these three indicators only AROP is depending directly on the household income distribution.

**2. Usage of administrative registers**

Data collection for EU-SILC in Austria is carried out by Computer Assisted Personal Interviewing (CAPI) and Computer Assisted Telephone Interviewing (CATI). From 2012 onwards a national regulation (ELStV – *Einkommens- und Lebens-bedingungen-Statistikverordnung*)[[5]](#footnote-5) provides the legal basis to also use existing income information from administrative registers for measuring income components required for EU‑SILC. This is carried out by matching income components on personal level by a pseudonymised key (*bereichsspezifisches Personenkennzeichen* - bPK) to the sample of EU-SILC.[[6]](#footnote-6) In EU-SILC 2017 about 87% of the volume of the annual household income could be covered by registers. Register information can also be applied for imputation if an income component is missing in one register but additional information from another register (e.g. social insurance register) indicates that a person is in a specific form of occupation.

Since register data are available for all persons in the sampling frame and thus in the gross and net sample those characteristic also play a vital role in the weighting procedure of EU-SILC in Austria. For example, weighting to counter selective unit nonresponse relies on estimated response probabilities. For these models also predictors based on register information are used and some marginal distributions in the calibration of weights are derived from income registers too.

The main usage of income register data presented in this text concerns the improvement of the sample selection for the first wave of EU-SILC. Several scenarios were discussed during a Eurostat grant agreement on “Improving Methodology on Sampling, Weighting, Imputation and Variance Estimation in the Austrian EU‑SILC with regard to Administrative Data”. The corresponding work lead to the conclusion that a sub-stratification for the first wave sample selection by household income data from administrative registers which are correlated to the main variables of interest in EU-SILC is the most promising application.

**3. Sampling design of the first wave of EU-SILC**

*3.1. Overview*

As mentioned above EU-SILC in Austria uses an integrated rotational design meaning that each year about one fourth of the sample is replaced by a new rotational group. The first wave of such a new rotational group is selected by a one stage stratified probability sample with disproportional allocation by province (NUTS 2) and regional sub-stratification within provinces. The allocation of selected addresses varies by province in order to take into account different expected unit response rates (based on previous years’ response rates).

The sampling frame stems from the central residence register (*Zentrales Melderegister – ZMR*). The availability of income register data matched to the sampling frame by the pseudonymised key bPK described above opened up new opportunities for creating a more efficient sampling design. In the following sub-section the sample selection procedure for the first year wave as applied for EU-SILC 2015 to 2017 will be described in more detail.

*3.2. Motivation for new stratification criteria*

Up until EU‑SILC 2015 the sub-stratification within each province was designed according to interviewer units. These are regional divisions that may be approximately combined to NUTS 2 regions. Each interviewer unit is associated with the area that one CAPI interviewer has to cover.

The availability of income register data from administrative sources enables a variety of applications for improving the sampling design of the first wave. The main idea here is to introduce stratification criteria for the first wave sample based on income information and thus creating a “rich sampling frame” (cf. Groves, 2006, p. 654). More efficient estimation for the main indicators of EU‑SILC could be carried out, if stratifying by characteristics correlated with the main variables on income that are needed for AROP and AROPE was possible (cf. Särndal et al, 2003, pp. 100f.). Such stratification criteria should lead to smaller estimated standard errors compared to stratification criteria that are based on non-income related criteria (cf. Cochran, 1977, pp. 101ff.).

Ideally, only persons who are at risk of poverty or social exclusion are in one stratum and in the other stratum only those who are not affected are situated. Since this information is obviously not at hand before the survey takes place the closest we can get to the actual net household income is by using all net income components from adminsitrative registers aggregated over all persons in one household (HINCn\_REG) as the variable used for building efficient strata. Preliminary tests carried out with EU‑SILC 2014 showed that 63% of persons classified as AROP and 53% of persons classified as AROPE also fall under the 1st quartile of HINCn\_REG.

Thus it was concluded that stratifying by the 1st quartile (Q1) of HINCn\_REG could help to reduce the standard error of AROP and AROPE. However, EU-SILC still requires regional stratification criteria in order to carry out the fieldwork. Before EU‑SILC 2016 each interviewer unit was used as a stratum and so the exact number of households for the first wave sample to be contacted by the fieldwork department could be controlled. In order to incorporate register income information by stratifying by Q1 for each of the interviewer units a subdivision had to be carried out. Since there are interviewer units in some provinces (e.g. Burgenland) which only cover small regions a major concern was that the proposed new stratification by Q1 would very likely lead to strata with only one PSU (i.e. one address) and even empty strata in the net sample. This would contradict the construction of the primary strata target variable DB050 as described in the methodological guidelines for EU‑SILC (cf. Eurostat, 2017).

In order to be able to use the new sample selection scheme adjacent interviewer units were collapsed into larger regional strata, so called “regional units” (*Regionale Einheiten*). These strata are big enough to allow for a sub-stratification by Q1, because they were designed to contain on average about 60.000 addresses in the frame. They also can be combined to whole NUTS 2 regions. In other words, provinces can be broken down into sets of regional units.

*3.3. Application of the new stratification criteria*

For the first wave of EU-SILC 2016 and 2017 to each person registered at addresses of private households in the sampling frame income data from registers were matched using by the bPK. The sum of all net income components of an address in the frame was then used to calculate HINCn\_REG. Addresses with HINCn\_REG below the first quartile (Q1) were put into one stratum and the remaining addresses into the complementary stratum. This stratification criterion was applied to all 63 regional units resulting in a total number of 126 strata for the first wave of EU‑SILC 2016 and also the first wave of EU‑SILC 2017. Compared to 205 interviewer units used as geographical stratification criteria in EU‑SILC 2015 the larger regional units allow for a sub-stratification for each province by Q1 leading to a smaller number of strata as presented in table 1.

**Table 1. Strata including sub-stratification according to the first quartile (Q1) of household income from administrative registers**

|  |  |  |
| --- | --- | --- |
| **Province (NUTS 2)** | **Regional units** | **Regional units × Q1** |
| Burgenland | 2 | 4 |
| Niederösterreich | 11 | 22 |
| Wien | 15 | 30 |
| Kärnten | 4 | 8 |
| Steiermark | 9 | 18 |
| Oberösterreich | 10 | 20 |
| Salzburg | 4 | 8 |
| Tirol | 5 | 10 |
| Vorarlberg | 3 | 6 |
| **Total** | 63 | 126 |
| Source: Statistics Austria, EU-SILC 2015-2017 (unpublished results) |

**4. Results for standard error estimation**

In EU‑SILC 2016 only the first wave incorporated this new sampling design with sub-stratification by income register information, in 2017 the first and the second wave (previous year’s first wave) had the same new strata.

Table 2 and table 3 below compare the standard error and the coefficient of variation (CV) between different stratification criteria. In the simplest case only the NUTS 2 regions are used as stratification criteria, since disproportional allocation for the first wave samples is carried out through them and all smaller strata are nested within them. Furthermore, the NUTS 2 regions are strata definitions that did not change with the introduction of sub-stratification according to Q1 in 2016. DB050 denotes the actual strata also present in the Eurostat data files of EU‑SILC. They consist of interviewer units with the exceptions of the rotational subsamples that commenced in 2016 and 2017. For these the new first wave samples sub-stratification according to Q1 within regional units was applied. All standard error estimates presented below were carried out with SAS proc surveymeans, clustered on household level. For simplification, the at-risk-of-poverty threshold was assumed to have no sampling variation.

For both AROP and AROPE estimated based on the entire cross-section of a specific survey year of EU-SILC it is evident, that only for the years 2016 and 2017 stratifying by DB050 shows a slightly smaller standard error compared to using NUTS 2 as a stratification criterion. This result was expected, if sub-stratification by Q1 indeed produces more homogenous strata in terms of AROP and AROPE. However, for the whole cross-section the standard error is only about 2% smaller for both indicators in 2016 and about 2% to 3% smaller in 2017.

**Table 2. Standard error by stratification criteria for the at-risk-of-poverty indicator
(entire cross-section)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Year** | **AROP (in %)** | **Standard error** | **Coefficient of variation** |
| NUTS2 | DB050 | NUTS2 | DB050 |
| 2015 | 13.9 | 0.6351 | 0.6351 | 0.0457 | 0.0457 |
| 2016 | 14.1 | 0.7046 | 0.6929 | 0.0501 | 0.0493 |
| 2017 | 14.4 | 0.7226 | 0.7054 | 0.0501 | 0.0489 |
| Source: Statistics Austria, EU-SILC 2015-2017 (unpublished results) |

**Table 3. Standard error by stratification criteria for the at risk of poverty or social exclusion indicator (entire cross-section)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Year** | **AROPE (in %)** | **Standard error** | **Coefficient of variation** |
| NUTS2 | DB050 | NUTS2 | DB050 |
| 2015 | 18.3 | 0.6883 | 0.6871 | 0.0376 | 0.0375 |
| 2016 | 18.0 | 0.7343 | 0.7217 | 0.0409 | 0.0402 |
| 2017 | 18.1 | 0.7556 | 0.7362 | 0.0418 | 0.0407 |
| Source: Statistics Austria, EU-SILC 2015-2017 (unpublished results) |

The effect of using DB050 as a stratification criterion becomes more pronounced, if only the first year waves of 2015, 2016 and 2017 are considered independently. This means that the indicators are estimated using only the first waves and the weights produced for the first waves separately. This makes sure that the household weights are estimated as if each first wave sample was a separate survey. For the first wave of 2015 again, it is almost irrelevant if NUTS 2 or DB050 is the stratification criterion applied in the standard error estimation as is evident in table 4 and 5. In 2016 the effect of the new stratification is recognisable in a standard error that is 6% smaller for AROP and AROPE, in 2017 the corresponding standard errors are about 3% smaller.

**Table 4. Standard error by stratification criteria for the at-risk-of-poverty indicator
(first wave separately)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Year** | **AROP (in %)** | **Standard error** | **Coefficient of variation** |
| NUTS2 | DB050 | NUTS2 | DB050 |
| 2015 | 14.9 | 1.0750 | 1.0778 | 0.0722 | 0.0724 |
| 2016 | 13.8 | 0.9923 | 0.9315 | 0.0717 | 0.0673 |
| 2017 | 14.4 | 1.0798 | 1.0438 | 0.0750 | 0.0725 |
| Source: Statistics Austria, EU-SILC 2015-2017 (unpublished results) |

**Table 5. Standard error by stratification criteria for the at risk of poverty or social exclusion indicator (first wave separately)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Year** | **AROPE (in %)** | **Standard error** | **Coefficient of variation** |
| NUTS2 | DB050 | NUTS2 | DB050 |
| 2015 | 20.0 | 1.1864 | 1.1902 | 0.0594 | 0.0595 |
| 2016 | 17.5 | 1.0635 | 0.9945 | 0.0608 | 0.0569 |
| 2017 | 18.5 | 1.1799 | 1.1462 | 0.0639 | 0.0620 |
| Source: Statistics Austria, EU-SILC 2015-2017 (unpublished results) |

Another way to study the effect of different stratification criteria is by comparing the change of the coefficient of variation (CV) of AROP and AROPE over time. However, this is particularly challenging, because the standard errors of different years not only rely on the sample selection but many other effects that influence the distribution of the household weights (e.g. quality of predictor variables and models used for weighting, changing selectivity of nonresponse). For example, in SILC 2017, the household weights have a noticeable higher CV than in 2015 and 2016 (0.57 compared to 0.47 and 0.48).

For 2015 and 2016, where the dispersion of weights is similar, the usage of the new sub-stratification by Q1 was able to reduce the CV by 7% for AROP (cf. table 4) and by 4% for AROPE if the new stratification is taken into account (cf. table 5). This evidence shows that sub-stratification by income characteristics from administrative registers has potential to reduce the standard error for EU‑SILC. Nevertheless, in 2017 despite this new methodology the CV of AROP is almost the same as in 2015 and even slightly higher for AROPE. As mentioned above, for 2017 comparisons over time are a particular challenge because of the dispersion of weights and only future survey years will give a profound evaluation of the new stratification procedure presented in this paper.

**5. Outlook**

Results for stratifying the first wave sample of EU-SILC by income data from administrative registers show a moderate effect on the reduction of the standard error. However, the results described above also indicate that the applied sub-stratification criterion (first quartile of household income only from registers: Q1) may not capture the real situation on income and living conditions of households in the sample adequately enough, since it is missing important components of the household income, such as cash benefits or losses from self-employment or regular inter-household cash transfers.

For the first wave of EU-SILC 2018 the approach for stratifying the first wave sample of EU-SILC described in this paper was enhanced by using the newly available rich sampling frame. It is a quarterly generated frame of the whole Austrian population based on a couple of different registers supposed to be used by all household surveys carried out by Statistics Austria. Socio-demographic variables in combination with the available income information were used to train a machine learning algorithm for predicting the AROPE onto the whole frame. This predicted AROPE was afterwards used as a sub-stratification variable instead of Q1.

In 2019 all four rotational sub-samples are going to incorporate a sampling procedure that applies stratification by criteria that are correlated with the main indicators. Only then a complete evaluation of the presented new stratification procedures will be possible for the entire cross-sectional sample of EU‑SILC in Austria.

**6. References**

Cochran, W. G. (1977), Sampling Techniques, Wiley, New York.

Eurostat, (2017), Methodological guidelines and description of EU-SILC target variables 2016 Operation. Eurostat. Luxemburg. Available at: <https://circabc.europa.eu/sd/a/165c80b9-5631-4f5b-b847-29c638715c0e/DOCSILC065%20operation%202016%20VERSION%2022-05-2017.pdf> (Accessed: 14 May 2018).

Groves, R. M. (2006), Nonresponse Rates and Nonresponse Bias in Household Surveys, Public Opinion Quarterly 70 (5), pp. 646-675.

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

Särndal, C.-E., Swensson, B. and Wretman, J. (2003), Model assisted survey sampling, Springer, New York.

1. Cf. <http://ec.europa.eu/eurostat/statistics-explained/index.php/Europe_2020_headline_indicators>
(Accessed 17 May 2018) [↑](#footnote-ref-1)
2. The equivalised household income is defined as the available net household income divided by the number of consumption equivalents in the household which are calculated as a weighted sum of assumed resource requirements. A single adult is assigned a weight of 1, any additional adult receives a weight of 0.5 and every child below the age of 14 a weight of 0.3. [↑](#footnote-ref-2)
3. Cannot afford to pay rent or utility bills; keep home adequately warm; face unexpected expenses; eat meat, fish or a protein equivalent every second day; a week holiday away from home; a car; a washing machine; a colour TV; a telephone (including mobile phone). [↑](#footnote-ref-3)
4. Households where the work intensity of persons aged 18-59 is below 20% of the theoretically reachable work intensity. This indicator applies to household members younger than 60 years. [↑](#footnote-ref-4)
5. Cf. <http://www.statistik.at/web_de/static/eu-silc_nationale_statistik-verordnung_elstv_055277.pdf> (Accessed 12 May 2018). [↑](#footnote-ref-5)
6. For more information about the record linkage necessary for matching income data from adminstrative register to EU-SILC, please refer to Heuberger et al (2013), ch. 10.2. [↑](#footnote-ref-6)