

**Outliers at upper end of income distribution
(EU-SILC 2007)**

Laura Neri, Francesca Gagliardi, Giulia Ciampalini,
Vijay Verma and Gianni Betti

Working Paper n. 86, November 2009

Outliers at upper end of income distribution (EU-SILC 2007)

Laura Neri, Francesca Gagliardi, Giulia Ciampalini, Vijay Verma and Gianni Betti

Abstract

The presence of extreme values at the upper end of the household income distribution, even if few in number, can greatly affect the inequality measures constructed from the data. It is necessary to control the impact of such values in order to improve comparability of the results from different countries. As a continuation of our earlier work presented in a number of reports, this paper explores procedures for the identification, examination and treatment of outliers in the empirical income distribution. Illustrations are given from six countries using EU-SILC 2007 data.

1. Introduction

This work presents an empirical analysis of very large values and other outliers in the income distribution from EU-SILC 2007 surveys¹. The objective is to develop proposals for the treatment of such values in using the data for constructing poverty and inequality measures, in particular Laeken indicators².

In the literature, there exist various methods to treat extreme (income) values and to adjust them, such as trimming, winsorizing, drawing from parametric tails, dropping unreliable income recipients (see for instance Van Kerm, 2006; Cowell and Victoria-Feser, 1996).

Of course before applying any solution – whether empirical or model-based – it is always necessary to empirically investigate the nature and extent of the outlier problem.

In two earlier studies prepared for Eurostat,³ we presented an empirical investigation of the nature and extent of the outliers problem in income data from 14 surveys for which 2004 EU-SILC data were available and from all the 26 countries for which 2005 EU-SILC data were available. We also proposed an approach to the treatment of extreme values, an approach which is largely empirical.

¹ EU-SILC is the European Union – Income and Living Condition survey. In this paper we have used the 2007 round cross-sectional component of the User Data Base (UDB) available to researchers.

² European Commission (2003), Laeken indicators. Detailed calculation methodology. DOC. E2/IPSE/2003.

³ Eurostat (2007b) and Eurostat (2007c).

The approach in the present analysis is more purely empirical beyond the above. *We take the view that, for routine application in a large system such as EU-SILC, not only the investigation but also the treatment should be simple, robust and purely empirically – based.*

In order to assist other researchers to conducting similar empirical analysis, we list in Annex B the SAS code with comments we have used.

2. General characteristics of the extreme ends of the income distribution

It is instructive to begin with a look at extreme ends of the income distribution in individual countries. Tables 1, 2 and 3 present some useful indicators.

Table 1 shows, for total net household income (HY020), the mean and median values, the minimum and maximum values in the data set, and percentiles P80, P90, P95 and P99 relative to the national median value.

Bottom end of the income distribution

In most countries, negative values of total disposable income have been allowed in EU-SILC 2007 data. The only exceptions are: Czech Republic where all recorded values are non negative; and Portugal where all recorded values are strictly positive. (The national Quality Report specifies that total disposable income corresponds to the sum of various components so that it does not present missing values; however no reference is made to negative income). In some countries, very large negative values appears: for instance the (algebraically) minimum value exceeds five times in magnitude the national median values in Germany, Denmark, Norway, and Slovenia.

Such differences in the treatment of negative (and zero) values can affect comparability across countries of the poverty and inequality statistics generated.

Elsewhere we have recommended “bottom-coding” of the variable “equivalised disposable income” used for constructing poverty and inequality indicators - such as bottom coding of total disposable income to 15% of the national median (Eurostat, 2006).

Problems in the empirical data at the bottom end of the income distribution are not discussed further in this paper.

Table 1 HY020 distribution by country, main indicators and ratio between percentiles and median

Countries	num	min	median	mean	max	R_mean	R_P_80	R_P_90	R_P_95	R_P_99
1 AT	6,806	-4,200	27,971	32,698	284,077	1.17	1.63	2.07	2.54	4.25
2 BE	6,348	-22,500	25,016	29,978	507,000	1.20	1.75	2.17	2.65	3.85
3 CY	3,505	-35	29,141	34,424	663,180	1.18	1.63	2.07	2.52	4.44
4 CZ	9,675	0	8,810	10,367	196,136	1.18	1.62	2.06	2.52	3.83
5 DE	14,153	-186,249	24,445	29,742	647,312	1.22	1.73	2.23	2.76	4.64
6 DK	5,783	-618,129	28,824	35,367	1,755,973	1.23	1.81	2.17	2.52	3.68
7 EE	5,146	-1,296	6,404	8,306	256,511	1.30	1.89	2.53	3.21	5.05
8 ES	12,329	-11,171	20,820	24,352	212,393	1.17	1.70	2.16	2.67	3.95
9 FI	10,624	-3,261	25,702	30,609	855,602	1.19	1.71	2.10	2.52	3.85
10 FR	10,498	-79,373	24,875	28,849	1,041,917	1.16	1.62	2.03	2.46	3.75
11 GR	5,643	-21,759	16,889	21,132	359,408	1.25	1.79	2.37	2.99	5.10
12 HU	8,737	-45	6,540	7,565	135,851	1.16	1.60	2.00	2.45	3.63
13 IE	5,608	-736	37,880	46,734	1,510,760	1.23	1.78	2.29	2.80	4.36
14 IS	2,872	-228	45,668	55,642	1,070,147	1.22	1.64	2.11	2.64	4.63
15 IT	20,982	-60,090	23,051	28,529	591,411	1.24	1.78	2.31	2.88	4.72
16 LT	4,975	-579	5,213	6,670	77,466	1.28	1.93	2.57	3.23	5.02
17 LU	3,885	-22,977	47,870	57,104	3,629,892	1.19	1.66	2.07	2.57	3.92
18 LV	4,471	-248	5,166	6,863	123,618	1.33	2.02	2.81	3.58	5.56
19 NL	10,219	-62,602	27,179	32,321	449,625	1.19	1.64	2.07	2.47	4.44
20 NO	6,013	-382,721	37,024	43,551	882,264	1.18	1.74	2.10	2.44	3.50
21 PL	14,286	-2,380	6,285	7,707	178,937	1.23	1.73	2.29	2.86	4.48
22 PT	4,310	469	13,800	17,850	275,900	1.29	1.81	2.54	3.51	5.67
23 SE	7,183	-8,289	25,230	29,145	416,955	1.16	1.68	2.03	2.36	3.31
24 SI	8,707	-241,292	18,205	19,926	161,511	1.09	1.56	1.91	2.28	3.17
25 SK	4,941	-50	6,747	7,827	57,973	1.16	1.65	2.10	2.52	3.84
26 UK	9,275	-12,669	31,891	39,423	1,835,132	1.24	1.76	2.31	2.89	4.76
mean						<i>1.21</i>	<i>1.72</i>	<i>2.21</i>	<i>2.72</i>	<i>4.29</i>

Table 2 Ratio between upper 10% percentiles of HY020 and median

Countries	R_P_90	R_P_91	R_P_92	R_P_93	R_P_94	R_P_95	R_P_96	R_P_97	R_P_98	R_P_99	R_P_100
1 AT	2.07	2.14	2.23	2.33	2.42	2.54	2.67	2.92	3.29	4.25	10.16
2 BE	2.17	2.25	2.31	2.42	2.51	2.65	2.79	2.96	3.27	3.85	20.27
3 CY	2.07	2.13	2.20	2.30	2.43	2.52	2.66	2.91	3.49	4.44	22.76
4 CZ	2.06	2.11	2.19	2.30	2.39	2.52	2.69	2.93	3.35	3.83	22.26
5 DE	2.23	2.30	2.39	2.49	2.61	2.76	2.94	3.18	3.61	4.64	26.48
6 DK	2.17	2.22	2.28	2.34	2.42	2.52	2.63	2.80	3.08	3.68	60.92
7 EE	2.53	2.64	2.73	2.84	3.00	3.21	3.44	3.75	4.28	5.05	40.06
8 ES	2.16	2.25	2.34	2.44	2.53	2.67	2.82	3.01	3.33	3.95	10.20
9 FI	2.10	2.15	2.22	2.31	2.42	2.52	2.65	2.88	3.21	3.85	33.29
10 FR	2.03	2.09	2.16	2.24	2.34	2.46	2.61	2.81	3.05	3.75	41.89
11 GR	2.37	2.45	2.54	2.65	2.80	2.99	3.18	3.53	3.98	5.10	21.28
12 HU	2.00	2.05	2.15	2.22	2.34	2.45	2.62	2.79	3.09	3.63	20.77
13 IE	2.29	2.36	2.46	2.58	2.68	2.80	3.02	3.31	3.71	4.36	39.88
14 IS	2.11	2.18	2.25	2.35	2.48	2.64	2.86	3.16	3.55	4.63	23.43
15 IT	2.31	2.40	2.50	2.60	2.73	2.88	3.09	3.41	3.79	4.72	25.66
16 LT	2.57	2.67	2.78	2.90	3.09	3.23	3.50	3.88	4.31	5.02	14.86
17 LU	2.07	2.14	2.21	2.31	2.43	2.57	2.75	3.00	3.18	3.92	75.83
18 LV	2.81	2.91	3.05	3.17	3.28	3.58	3.73	3.96	4.37	5.56	23.93
19 NL	2.07	2.12	2.18	2.25	2.36	2.47	2.62	2.89	3.30	4.44	16.54
20 NO	2.10	2.15	2.20	2.27	2.34	2.44	2.57	2.74	3.01	3.50	23.83
21 PL	2.29	2.38	2.49	2.60	2.72	2.86	3.10	3.42	3.85	4.48	28.47
22 PT	2.54	2.64	2.80	2.98	3.25	3.51	3.78	4.06	4.56	5.67	19.99
23 SE	2.03	2.08	2.14	2.20	2.28	2.36	2.47	2.60	2.81	3.31	16.53
24 SI	1.91	1.97	2.04	2.11	2.19	2.28	2.41	2.56	2.79	3.17	8.87
25 SK	2.10	2.16	2.23	2.31	2.40	2.52	2.67	2.86	3.19	3.84	8.59
26 UK	2.31	2.41	2.50	2.61	2.75	2.89	3.12	3.41	3.85	4.76	57.54

Table 3 Ratio between upper 1% percentiles of HY020 and median

Countries	R_P_99	R_P_99_1	R_P_99_2	R_P_99_3	R_P_99_4	R_P_99_5	R_P_99_6	R_P_99_7	R_P_99_8	R_P_99_9	R_P_100
1 AT	4.25	4.29	4.40	4.56	4.76	5.12	5.28	5.74	6.40	7.61	10.16
2 BE	3.85	4.00	4.04	4.10	4.45	4.56	4.80	5.09	5.85	7.39	20.27
3 CY	4.44	4.67	5.16	6.29	6.60	7.09	8.00	8.97	11.88	15.45	22.76
4 CZ	3.83	3.89	3.96	4.15	4.32	4.47	4.79	5.47	6.36	7.79	22.26
5 DE	4.64	4.82	5.00	5.27	5.65	6.14	6.60	7.50	8.92	11.24	26.48
6 DK	3.68	3.83	4.01	4.30	4.45	4.91	5.73	6.23	8.29	11.92	60.92
7 EE	5.05	5.16	5.27	5.68	5.79	6.12	6.14	6.39	7.17	7.92	40.06
8 ES	3.95	4.04	4.16	4.35	4.53	4.74	4.97	5.40	5.85	6.87	10.20
9 FI	3.85	3.98	4.14	4.26	4.42	4.67	5.41	5.84	7.06	11.22	33.29
10 FR	3.75	3.86	3.96	4.07	4.26	4.41	4.66	5.07	5.55	7.59	41.89
11 GR	5.10	5.30	5.47	5.65	5.72	6.33	6.53	7.36	8.75	9.73	21.28
12 HU	3.63	3.84	3.99	4.20	4.59	4.80	4.96	5.31	6.27	7.27	20.77
13 IE	4.36	4.50	4.64	5.03	5.43	5.87	6.47	7.36	7.90	11.11	39.88
14 IS	4.63	5.25	6.02	6.60	6.84	7.35	8.06	8.98	11.01	13.25	23.43
15 IT	4.72	4.87	5.06	5.24	5.55	5.83	6.19	6.58	7.24	8.89	25.66
16 LT	5.02	5.09	5.09	5.20	5.48	5.77	6.16	6.33	6.72	8.05	14.86
17 LU	3.92	4.00	4.14	4.60	4.96	5.04	5.56	5.68	5.68	5.87	75.83
18 LV	5.56	5.81	5.97	6.25	6.63	7.06	7.35	7.60	7.84	8.95	23.93
19 NL	4.44	4.67	4.90	5.27	5.48	6.02	6.88	8.84	10.35	11.61	16.54
20 NO	3.50	3.58	3.67	3.82	4.06	4.33	4.70	5.01	5.99	7.66	23.83
21 PL	4.48	4.61	4.74	4.90	5.09	5.43	5.68	5.99	6.73	8.15	28.47
22 PT	5.67	5.73	5.81	6.46	6.87	7.45	7.82	7.86	9.49	9.88	19.99
23 SE	3.31	3.39	3.51	3.62	3.77	4.03	4.22	4.58	5.21	7.12	16.53
24 SI	3.17	3.23	3.29	3.36	3.46	3.56	3.66	3.86	4.18	5.40	8.87
25 SK	3.84	4.04	4.08	4.25	4.34	4.52	4.60	4.87	5.19	5.96	8.59
26 UK	4.76	4.88	5.11	5.28	5.51	5.92	6.37	6.93	7.65	10.99	57.54

Top end of the income distribution

Similarly at the upper end, extremely large maximum values appears in some countries (see Table 2), for instance considerably exceeding 76 times the national median in Luxembourg, 61 the median in Denmark, and nearly 58 the median in the United Kingdom. Among the remaining countries, the ratio (maximum/median) varies from nearly 42 in France, down to 8.6 in Slovakia to. It is worth, pointing out that in EU-SILC 2005 the upper end values was much more extreme, for example the maximum values of the total disposable income exceeded 200 times the median in Belgium and nearly 300 times in Norway. Clearly here as well, some “top coding” to control against the occurrence of ‘uncontrolled’ extremely large values is desirable. Investigating these issues further is the objective of the present paper. This develops further the methodology we proposed earlier (Eurostat, 2007b). Ratio (P99/Median) exceeds 5.0 in Lithuania, Latvia, Portugal and the United Kingdom, and is approaching 5.0 in a number of other countries (Estonia, Greece, Ireland, Italy and Poland). Tables 2 and 3 examine this ratio at the extreme upper end of the distribution, from P90 to P99 (Table 2) and from P99.1 to P99.9 (Table 3). The patterns of occurrence of extreme values can be quite different in different countries. For instance we may note that neither Cyprus nor Luxembourg appear in the set of countries identified above for which the ratio P99.9/Median is approaching or exceeds 5.0 (this ratio is 4.4 in Cyprus and 3.9 in Luxembourg).

3. Procedure for the identification of extreme values

In the empirical analysis, we identify three classes of extreme values in the net household incomes:

- (a) “very large” values, such as values more than 3 or 4 times larger than the median net household income.
- (b) “outliers”, a subset of (a), with values exceeding 4 or 5 times the median.

Hitherto, in EU-SILC data sets, typically the “very large” set, (a), comprises around 1% of the households in the country, while the “outliers” subset, (b), is around 0.1-0.2% of the households. In a sample of 8,000 households, for instance, we may expect around 80 cases with “very large” income values, and 10-15 “outliers”.

- (c) “Implausible” values. On the basis of a closer examination of very large values, such as of individual components making up the total amount, it is sometimes possible to identify values which are *highly likely to be erroneous or implausible*. These should normally be discarded (treated as “missing”) and excluded from analysis.

In more details, we have identified the above sets in each country data set on the basis of empirical criteria as follows.

4. Identifying the ‘very large’ value set

The lower boundary of the very large-values set, (a), can be identified, for example, on the basis of the largest gap in the income level near the point of inflection at the upper end of the income distribution.

In a previous report (Eurostat, 2007b), we reported the results of a country-by-country empirical investigation to identify the boundary of the very large values set. Our objective was to seek a *general* and, as far as possible, a *simple* strategy to identify this boundary. For this purpose, we used linear regression to identify the relationship between different percentiles (P98.0, P98.5, P99.0, P99.5 etc.) of the HY020 distribution and the required boundary – i.e. the boundary defined as the point at which the largest jump in the income value occurred near the point of inflection at the upper end of the distribution. (Countries served as the data points for this analysis.)

The results clearly indicated that in most cases the required boundary (above which the income values may be considered very large in the above sense) was quite close to 99th percentiles (P99).

Based on the above experience, albeit only from the 14 countries in the 2004 survey, we recommend to adopt a very simple definition of what constitutes “very large values”: these may consist of the top 1% of the households, i.e. of households with total disposable income Y at or exceeding the 99th percentile: $Y \geq P99$.

Table 4 is very much focused on extreme values, and shows the number of households with disposable income above specified multiples (2 to 12) of the national median. It also reports the number of households in each country with total disposable income Y at or exceeding the 99th percentile: $Y \geq P99$.

As already seen in Table 3, 99th percentile (P99) lies close to 4-5 times the median. For this reason, the number (of households) with $Y > P99$ is shown between the columns for $Y > 4Y_m$ and $Y > 5Y_m$ in Table 4. The largest number of cases with household income exceeding 10 times the national median were found in the Netherlands (26) and Italy, Germany and Finland (around 20). In Slovenia and Slovakia no incomes have been reported to reach or exceed 9 times the national median.

Table 4 Number of households with disposable income (HY020) above specified multiples of national median.

Countries	num	HY020_2m	HY020_3m	HY020_4m	HY020_P99	HY020_5m	HY020_6m	HY020_7m	HY020_8m	HY020_9m	HY020_10m	HY020_12m
1 AT	6,806	801	192	72	61	34	17	7	6	4	1	
2 BE	6,348	898	197	64	74	27	14	9	6	5	2	2
3 CY	3,505	389	108	51	41	35	29	18	13	10	9	5
4 CZ	9,675	836	190	54	69	25	14	9	4	3	2	2
5 DE	14,153	2,080	553	211	142	116	73	45	36	26	19	13
6 DK	5,783	1,626	278	96	119	57	40	32	25	19	16	10
7 EE	5,146	997	315	129	46	48	25	9	3	3	3	3
8 ES	12,329	1,530	361	113	117	42	17	9	3	2	1	
9 FI	10,624	2,350	592	204	231	102	57	41	32	25	19	11
10 FR	10,498	1,359	311	111	147	43	25	17	9	7	5	1
11 GR	5,643	803	249	101	53	57	30	20	16	9	5	3
12 HU	8,737	833	180	66	85	34	18	9	6	5	3	2
13 IE	5,608	656	189	79	59	47	29	22	14	10	8	5
14 IS	2,872	470	137	53	37	32	28	20	17	12	11	6
15 IT	20,982	3,430	1,057	410	240	197	99	54	41	26	20	10
16 LT	4,975	858	292	130	48	49	26	11	7	5	4	2
17 LU	3,885	453	112	36	41	21	11	8	6	6	3	3
18 LV	4,471	741	283	108	38	61	29	19	9	3	2	2
19 NL	10,219	1,478	333	134	110	85	52	41	34	31	26	8
20 NO	6,013	1,188	197	59	94	26	16	11	6	5	5	4
21 PL	14,286	1,879	532	198	116	73	35	17	11	7	6	4
22 PT	4,310	643	267	121	38	59	32	22	10	9	4	3
23 SE	7,183	1,111	160	48	101	20	11	10	6	5	4	3
24 SI	8,707	1,109	174	40	135	22	10	5	1			
25 SK	4,941	634	126	46	50	11	4	2	1			
26 UK	9,275	1,336	419	169	95	82	44	28	15	12	11	6

5. Identifying the ‘outliers’ and implausible or “suspect” values

Outliers refer to the values which are clearly extreme. Of course, such values are not necessarily erroneous, but nevertheless have a high chance of being so. In any case, uncontrolled presence of even a small number of extreme values can adversely affect data comparability across countries and over time. Generally, the concept ‘outliers’ is useful if only a small proportion of the cases in the data are so classified. The incidence of outlying cases varies across countries. We therefore take an entirely empirical approach in order to identify them, as follows.

The following analysis concern just six MS taken as examples.

Graphs A in the annex, one for each country, examine the upper end (Y>P99) of the income distribution. *We take ‘outliers’ as the data points which appear clearly separated from the main body at the upper end of the income distribution.* In most cases, the boundary defining the outlier set so defined is easily identified.

Table 5 shows this boundary and the number of cases identified as outliers in this way in each national income distribution: outliers account for less than 0.2% of the total households.

The variation across countries is not marked. It is possible that some of the very large or at least the outlier values are merely the results of data errors, in which case it would be appropriate to exclude them in data analysis. We do not attempt such data editing here.

Nevertheless, it is useful to identify the component(s) which contribute towards making the total income value an extreme amount. Large values recorded for some components can be less plausible than large values for some other components.

The full information is displayed in Annex Tables A, one for each of the six selected countries.

Table 5 The number and boundaries of outlier set (HY020)

Countries	num	median	outliers	outlier_min	outlier_max	boundary_outliers	perc outliers
DK	5,783	28,824	7	482,015	1,755,973	16.72	0.121%
FR	10,498	24,875	1	1,041,917	1,041,917	41.89	0.010%
IE	5,608	37,880	8	406,336	1,510,760	10.73	0.143%
LU	3,885	47,870	6	452,187	3,629,892	9.45	0.154%
PT	4,310	13,800	3	181,484	275,900	13.15	0.070%
UK	9,275	31,891	6	459,758	1,835,132	14.42	0.065%

6. Impact of adjusting extreme values on indicators

In an earlier report (Eurostat, 2007c), we examined five possible approaches to treating extreme values in total household disposable income (HY020):

- (i) Leaving all extreme (outliers and very large) values unchanged⁴
- (ii) Removing outliers from the dataset
- (iii) Trimming the outliers to the value immediately below the smallest outlier (TRIM1)
- (iv) Trimming all the very large values to the value immediately below the lower boundary of the very large value set. As noted earlier, we define that boundary as P99.
- (v) Trimming all very large values using linear extrapolation of $\log(\text{HY020})$ from the region immediately below the lower boundary. Linear trend in $\log(\text{HY020})$ is estimated from data points such as P98 to P99, and then extrapolated to replace the existing values above P99. The rank of households sorted according to ascending HY020 can serve as the independent variable and $\log(\text{HY020})$ as the dependent variable in the regression.

Option (ii) amounts to assuming that all outliers values are erroneous and therefore are to be rejected. This is too strong as a general assumption, and can affect the value of the resulting inequality indicators too much, as has been confirmed by our previous work. Hence we do not recommend to follow this procedure – except for possibly deleting data points judged to be erroneous on the basis of other considerations.

Option (v) is a refinement of (iv). It seems from previous analysis that usually the difference in the results from these two options is only minor. We have not followed up (v) further here for the sake of simplicity (which is very important practical consideration), though in principle it may be preferable to (iv).

Tables 6 (a) and (b) compare the impact of the choice among options (i)-(iv) on two basic measures of inequality: Gini and the share ratio S80/S20.⁵

For computing these indexes a bottom coding strategy to the lowest values of the distribution have been applied throughout: all values below 15% of the median household income have been set equal to the 15% of the median. The following observations may be made.

⁴ By ‘very large’ we mean values in the top one percent of the income distribution. ‘Outliers’ are generally the much smaller set clearly separated from the main body of the income distribution described above.

⁵ In contrast to these, poverty measures such as the conventional HCR (at-risk-of-poverty rate) are not affected by extreme values, except very marginally in case (ii).

- Generally the effect of the data adjustments on the computed indices is not very large.
- The strategy (ii) reduces the Gini index of less than 3%, the strategy (iii) is even less severe.
- The most evident impact is the one induced by strategy (iv). In particular in Denmark, both the inequality indices after the trimming are reduced by more than 7.5%.

Table 6a The impact of trimming on computed Gini index

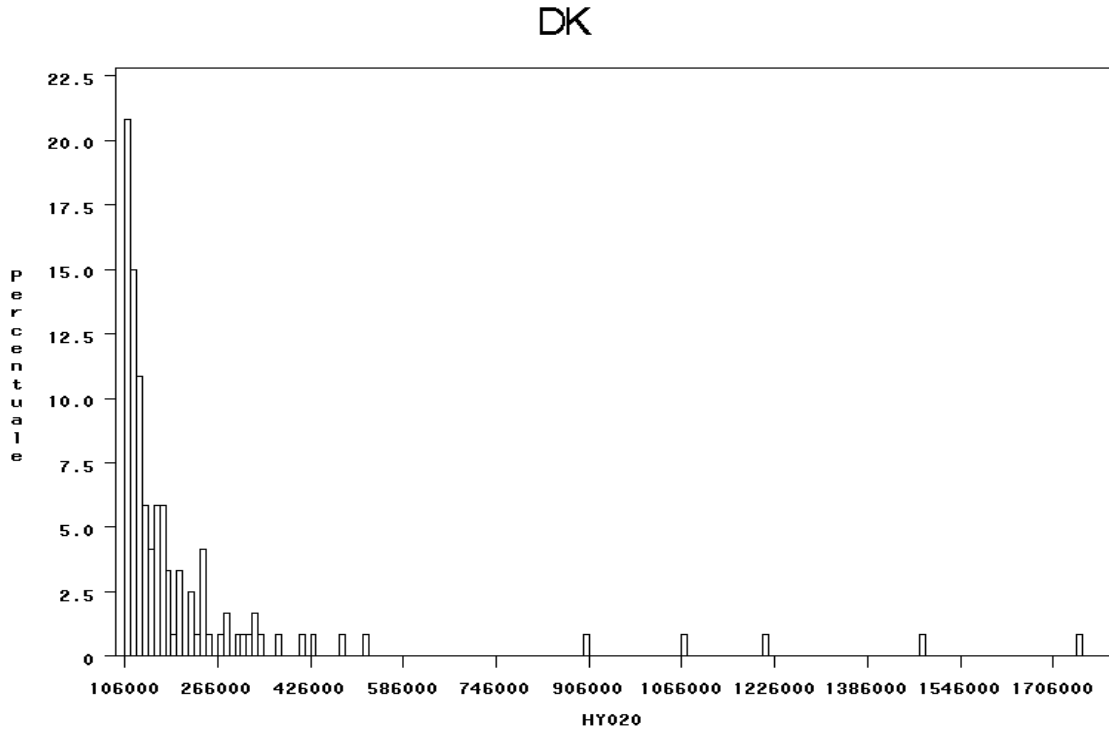
Option	(i)	(ii)	(iii)	(iv)	(ii)/(i)	(iii)/(i)	(iv)/(i)
Countries	Original	TRIM II	TRIM III	TRIM IV	ratio_2vs1	ratio_3vs1	ratio_4vs1
DK	25.46	24.72	25.03	23.56	0.9708	0.9832	0.9255
FR	26.69	26.55	26.58	25.91	0.9946	0.9959	0.9705
IE	32.36	31.41	31.82	30.90	0.9707	0.9833	0.9547
LU	28.05	27.82	27.92	27.11	0.9919	0.9952	0.9665
PT	38.18	37.71	37.96	37.16	0.9877	0.9944	0.9734
UK	33.53	32.86	33.18	32.25	0.9800	0.9896	0.9619

Table 6b The impact of trimming on computed S80/S20 ratio

Option	(i)	(ii)	(iii)	(iv)	(ii)/(i)	(iii)/(i)	(iv)/(i)
Countries	Original	TRIM II	TRIM III	TRIM IV	ratio_2vs1	ratio_3vs1	ratio_4vs1
DK	3.6472	3.5399	3.5876	3.3853	0.9706	0.9837	0.9282
FR	3.8674	3.8466	3.8514	3.7519	0.9946	0.9959	0.9702
IE	4.9742	4.7870	4.8757	4.7071	0.9624	0.9802	0.9463
LU	4.1098	4.0898	4.0892	3.9661	0.9951	0.9950	0.9650
PT	6.8750	6.7600	6.8228	6.6249	0.9833	0.9924	0.9636
UK	5.6037	5.4665	5.5323	5.3423	0.9755	0.9873	0.9534

**Annex A. Close view of very large values at upper end of the household income distribution.
Illustration from six countries (EU-SILC, 2007)**

Graph A.1 Denmark



From Graph A.1 it can be observed that for Denmark the number of outliers (7, 0.12%). In percentage terms this is a high value compared to three other (France, Portugal and the United Kingdom) among the 6 countries analysed, but very similar to the values reported below for Ireland and Luxembourg. From the ratio between the net and the gross income, HY020/HY010, in Table A.1, the tax amount seems to be heterogeneous among these households. Observing the equivalised income (HX090), it seems that all these high values are highly likely to be erroneous.

The income component mainly responsible of the outliers values is interest, dividends, profit from capital investments in unincorporated business (HY090), in three cases is also employee cash or near cash income (HPY010) and in only one case is cash benefits from self-employment (HPY050).

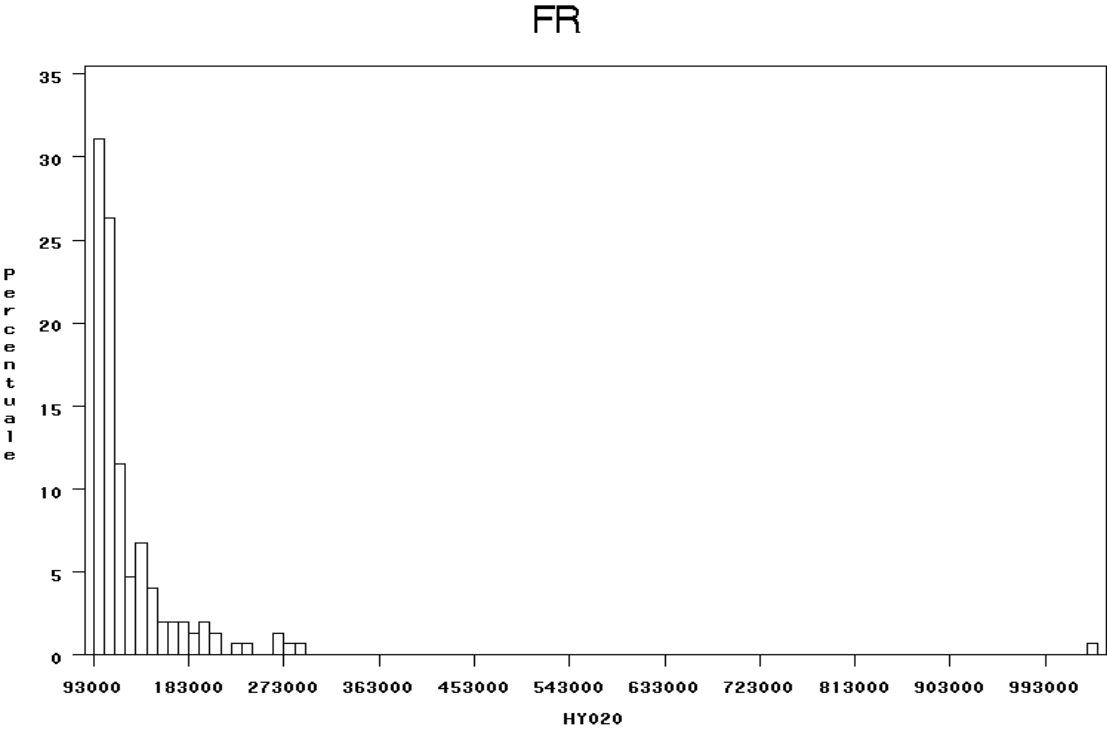
Table A.1 Denmark

	HY010	HY020	ratio (hy020/hy010)	HY040G	HY050G	HY060G	HY070G	HY080G	HY090G	HY110G	HX090
1	1,077,483	1,069,860	0.9929	0	0	0	0	0	1,048,710	0	534,930
2	1,971,544	1,485,404	0.7534	0	0	0	0	0	-3,872	0	990,270
3	885,995	522,664	0.5899	0	0	0	0	0	709,908	0	348,443
4	504,758	482,015	0.9549	0	1,294	0	0	0	396,454	-916	267,786
5	3,968,571	1,755,973	0.4425	0	2,719	0	0	1,670	2,016,033	6,953	627,133
6	2,189,362	1,211,117	0.5532	0	1,294	0	0	0	2,040,025	0	605,558
7	917,035	897,112	0.9783	0	2,589	0	0	0	885,405	-38	345,043

	hpy010g	hpy020g	hpy035g	hpy050g	hpy070g	hpy080g	hpy090g	hpy100g	hpy110g	hpy120g	hpy130g	hpy140g
1	15,240	0	0	0	0	5,934	0	0	0	0	0	7,600
2	25,751	0	316,110	1,949,665	0	0	0	0	0	0	0	0
3	77,651	34	0	86,429	0	0	0	0	0	0	12,007	0
4	101,109	5,698	17,415	1,118	0	0	0	0	0	0	0	0
5	1,941,196	402	1,287	0	0	0	0	0	0	0	0	0
6	119,028	0	5,631	29,015	0	0	0	0	0	0	0	0
7	29,080	0	11,261	0	0	0	0	0	0	0	0	0

Legend is reported below in table A.2 France.

Graph A.2 France



From Graph A.2 it can be observed that in France only one real outlier is present. The income component responsible of it is employee cash or near cash income (HPY010).

Table A.2 France

	HY010	HY020	ratio (hy020/hy010)	HY040G	HY050G	HY060G	HY070G	HY080G	HY090G	HY110G	HX090
1	1,755,254	1,041,917	0.5936	0	0	0	0	0	748	0	520,959

	hpy010g	hpy020g	hpy035g	hpy050g	hpy070g	hpy080g	hpy090g	hpy100g	hpy110g	hpy120g	hpy130g	hpy140g
1	1,754,506	0	0	0	0	0	0	0	0	0	0	0

HY010: Total household gross income

HY020: Total disposable household income

HX090: Equivalised disposable income

HY040: Income from rental of a property or land

HY090G: Interest, dividends, profit from capital investments in unincorporated business

HY050G: Family/Children related allowances

HY060G: Social exclusion not elsewhere classified

HY070G: Housing allowances

HY080G: Regular inter-household cash transfer received

HY110G: Income received by people aged under 16

Aggregation of individual variables to the household level:

HPY010G: Employee cash or near cash income

H PY020G: Non-Cash employee income

H PY035G: Contributions to individual private pension plans

H PY050G: Cash benefits or losses from self-employment

H PY070G: Value of goods produced by own-consumption

H PY080G: Pension from individual private plans

H PY090G: Unemployment benefits

H PY100G: Old-age benefits

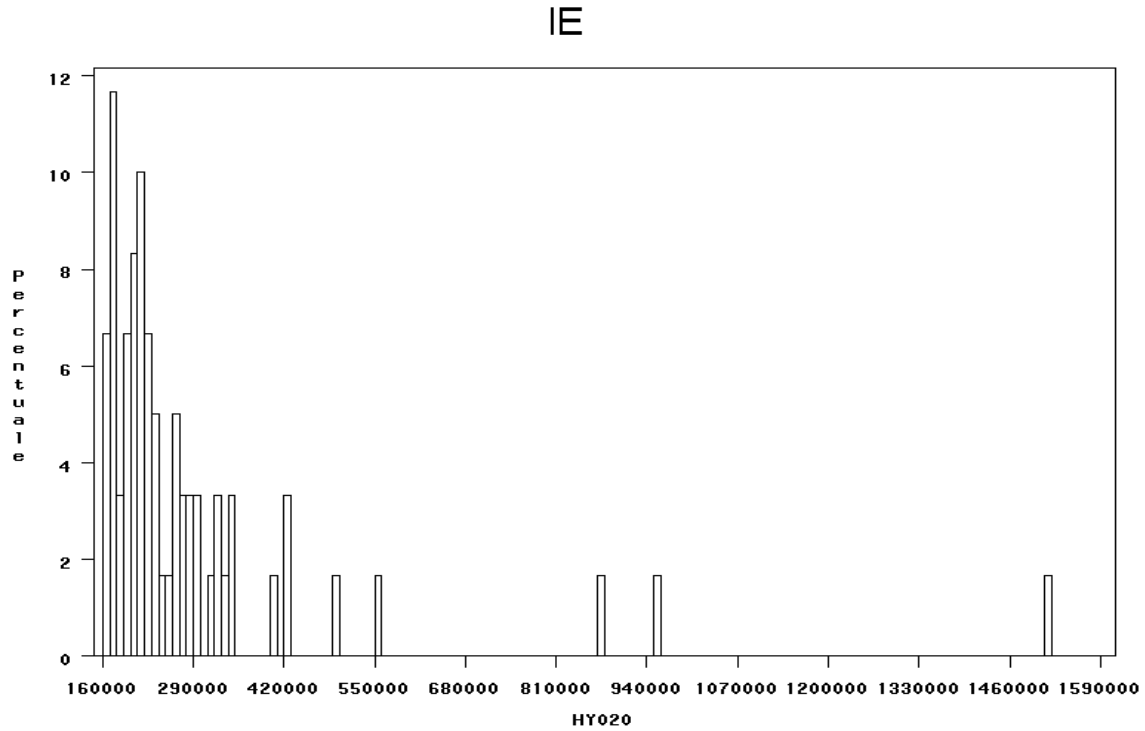
H PY110G: Survivor' benefits

H PY120G: Sickness benefits

H PY130G: Disability benefits

H PY140G: Education-related allowances

Graph A.3 Ireland



From Graph A.3 it can be observed that for Ireland the number of outliers (8, 0.14%), in percentage terms is very similar to that for Denmark and Luxembourg, and much higher than the remaining three countries analysed. From the ratio between the net and the gross income, HY020/HY010, in Table A.3, the tax amount seems to be quite uniform among these households, with the exception of few cases. Observing the equivalised income (HX090), it seems that all the outliers are likely to be genuine though, very large, values.

The income component mainly responsible of the outliers values is cash benefits from self-employment (HPY050). Only in one cases the components responsible of the outlier value is employee cash or near cash income (HPY010) combined with interest, dividends and profit from capital investments in unincorporated business (HY090).

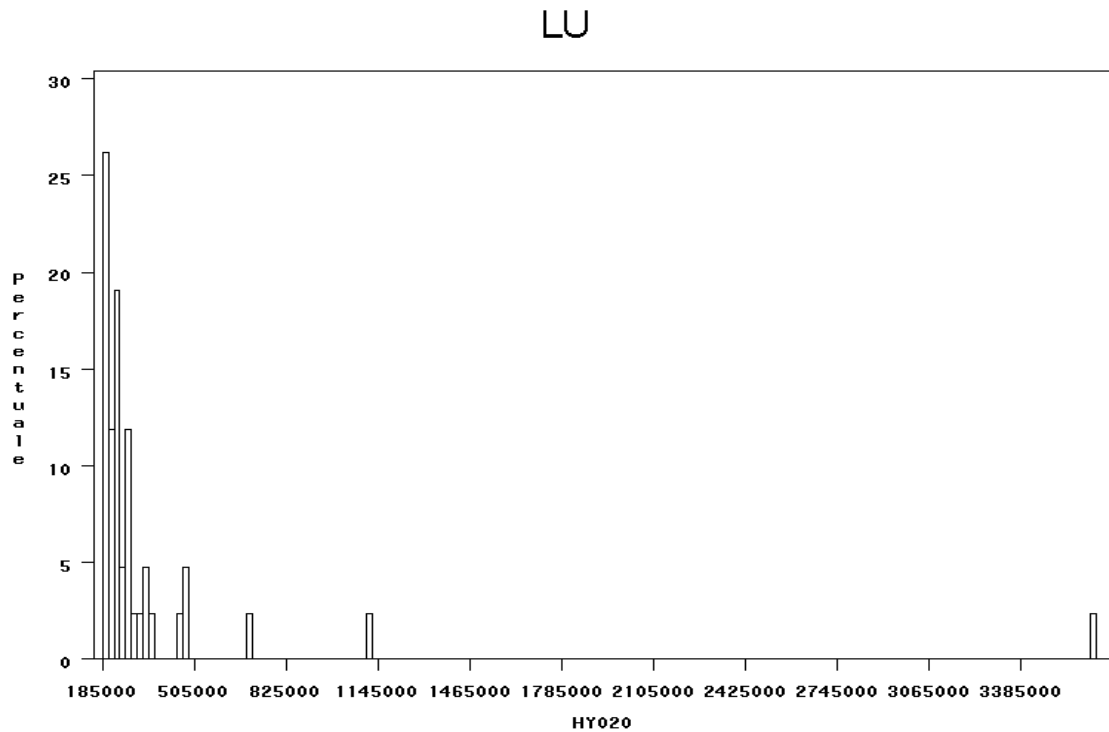
Table A.3 Ireland

	HY010	HY020	ratio (hy020/hy010)	HY040G	HY050G	HY060G	HY070G	HY080G	HY090G	HY110G	HX090
1	536,334	420,891	0.7848	0	0	0	0	0	349,840	0	168,356
2	550,935	406,336	0.7375	200,000	8,160	0	0	0	83,577	0	140,116
3	1,716,160	1,510,760	0.8803	190,000	8,160	0	0	0	12,500	0	559,541
4	571,480	422,225	0.7388	0	5,880	0	0	0	11,897	0	140,742
5	1,101,164	870,587	0.7906	279,082	0	0	0	0	0	0	580,392
6	606,855	497,489	0.8198	0	5,720	0	0	0	12,232	0	236,900
7	1,295,800	951,410	0.7342	800	0	0	0	0	525,000	0	634,273
8	673,475	558,642	0.8295	12,000	10,970	0	0	0	36,767	0	164,306

	hpy010g	hpy020g	hpy035g	hpy050g	hpy070g	hpy080g	hpy090g	hpy100g	hpy110g	hpy120g	hpy130g	hpy140g
1	186,494	0	0	0	0	0	0	0	0	0	0	0
2	9,343	0	0	249,855	0	0	0	0	0	0	0	0
3	0	0	3,120	1,505,500	0	0	0	0	0	0	0	0
4	1,044	0	31,982	249,642	0	0	303,017	0	0	0	0	0
5	0	0	5,880	822,082	0	0	0	0	0	0	0	0
6	27,403	0	3,120	560,000	0	0	0	0	0	0	0	1,500
7	150,000	0	126,000	620,000	0	0	0	0	0	0	0	0
8	50,000	0	4,200	563,738	0	0	0	0	0	0	0	0

Legend is reported above in table A.2 France.

Graph A.4 Luxembourg



From Graph A.4 it can be observed that for Luxembourg the number of outliers (6, 0.15%), in percentage terms is again at the highest end among the highest among the 6 countries analysed. From the ratio between the net and the gross income, HY020/HY010, in Table A.4, the tax amount seems to be rather heterogeneous among these households. Observing the equivalised income (HX090), all these values seem to be erroneous.

The income components mainly responsible of the outliers values is employee cash or near cash income (HPY010). In one case the component responsible of the outlier value is interest, dividends, profit from capital investments in unincorporated business (HY090) combined with cash benefits from self-employment (HPY050). In one case one of the component mainly responsible of the outliers values is disability benefits (HPY130).

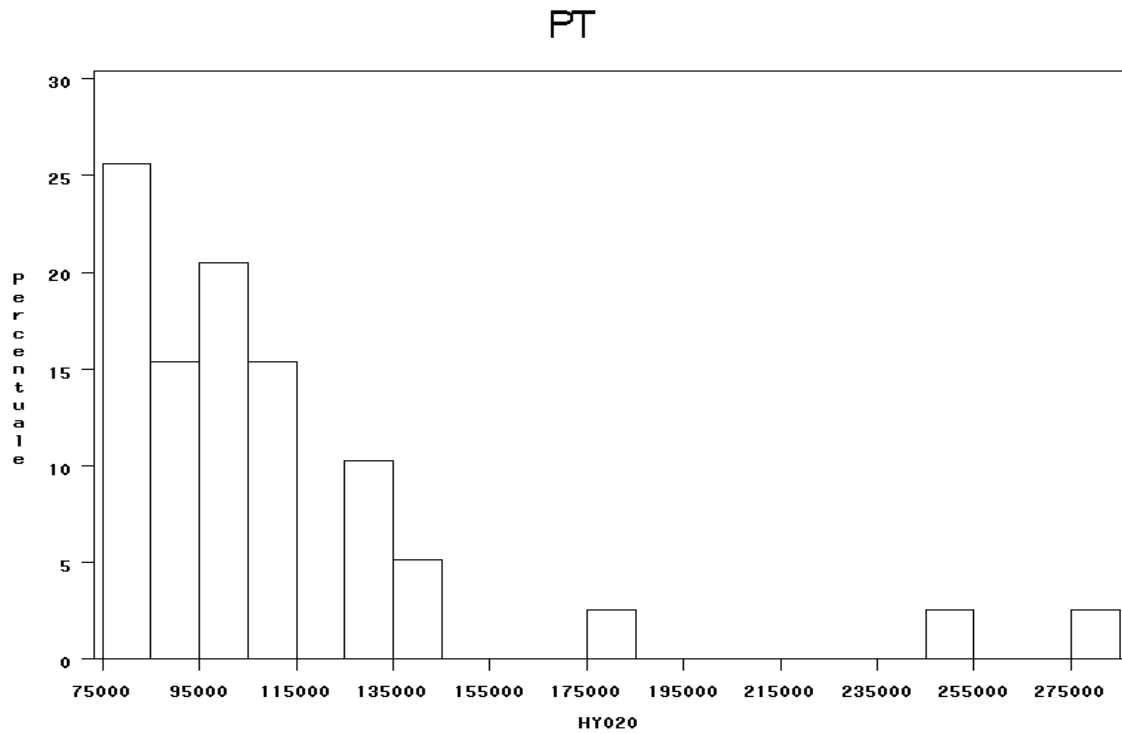
Table A.4 Luxembourg

	HY010	HY020 ratio (hy020/hy010)	HY040G	HY050G	HY060G	HY070G	HY080G	HY090G	HY110G	HX090
1	749,678	477,257	0.6366	0	5,678	0	0	0	0	227,265
2	2,058,916	1,105,755	0.5371	0	0	0	0	1,000,000	0	737,170
3	1,137,941	699,877	0.6150	40,000	0	0	0	250,000	0	699,877
4	519,706	452,187	0.8701	76,800	10,517	0	0	2,850	0	173,918
5	532,059	478,348	0.8991	0	9,100	0	0	500	0	207,977
6	5,194,609	3,629,892	0.6988	0	0	0	0	95	0	2,419,928

	hpy010g	hpy020g	hpy035g	hpy050g	hpy070g	hpy080g	hpy090g	hpy100g	hpy110g	hpy120g	hpy130g	hpy140g
1	744,000	0	36,000	0	0	0	0	0	0	0	0	0
2	43,916	0	1,200,000	1,015,000	0	0	0	0	0	0	0	0
3	840,000	7,941	1,500	0	0	0	0	0	0	0	0	0
4	213,460	0	4,260	216,080	0	0	0	0	0	0	0	0
5	222,228	0	0	0	0	0	0	0	0	0	300,230	0
6	5,164,226	7,788	0	0	0	0	0	22,500	0	0	0	0

Legend is reported above in table A.2 France.

Graph A.5 Portugal



From Graph A.5 it can be observed that for Portugal the number of outliers, only 3, is very low. Observing the equivalised income (HX090), it seems that all of these households may be erroneous. The income components mainly responsible of the outliers values are employee cash or near cash income (HPY010) and/or cash benefits from self-employment (HPY050).

Table A.5 Portugal

	HY010	HY020 ratio (hy020/hy010)	HY040G	HY050G	HY060G	HY070G	HY080G	HY090G	HY110G	HX090
1	344,034	249,625	0.7256	0	0	0	0	0	0	124,813
2	228,085	181,484	0.7957	0	0	0	0	6,250	0	120,989
3	456,959	275,900	0.6038	0	0	0	0	0	0	137,950

	hpy010g	hpy020g	hpy035g	hpy050g	hpy070g	hpy080g	hpy090g	hpy100g	hpy110g	hpy120g	hpy130g	hpy140g
1	122,689	3,625	1,772	217,720	0	0	0	0	0	0	0	0
2	0	0	3,000	175,000	0	0	0	46,835	0	0	0	0
3	447,059	0	0	0	0	0	0	9,900	0	0	0	0

Legend is reported above in table A.2 France.

Table A.6 United Kingdom

	HY010	HY020	ratio (hy020/hy010)	HY040G	HY050G	HY060G	HY070G	HY080G	HY090G	HY110G	HX090
1	803,759	560,873	0.6978	17,602	1,381	0	0	0	6,326	0	186,958
2	707,510	459,758	0.6498	70,409	0	0	0	0	554,547	0	306,505
3	778,698	487,250	0.6257	35,938	0	0	0	0	458	0	324,834
4	865,603	583,102	0.6736	0	1,331	0	0	0	0	0	253,523
5	639,001	477,874	0.7478	0	0	0	0	0	14,221	0	477,874
6	2,582,268	1,835,132	0.7107	132,017	1,407	0	0	0	9,168	0	797,884

	hpy010g	hpy020g	hpy035g	hpy050g	hpy070g	hpy080g	hpy090g	hpy100g	hpy110g	hpy120g	hpy130g	hpy140g
1	745,079	33,371	0	0	0	0	0	0	0	0	0	0
2	12,146	0	0	70,409	0	0	0	0	0	0	0	0
3	0	0	30,804	742,301	0	0	0	0	0	0	0	0
4	864,272	0	3,520	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	624,780	0	0	0	0
6	10,561	0	0	0	0	0	0	2,429,114	0	0	0	0

Legend is reported above in table A.2 France.

References

- Cowell F.A., Victoria-Feser M.P. (1996), Poverty measurement with contaminated data: a robust approach. *European Economic Review*, **40**, 1761-1771.
- European Commission (2003), Laeken indicators. Detailed calculation methodology. DOC. E2/IPSE/2003.
- Eurostat (2006), Treatment of negative income: empirical assessment of the impact of methods used. Report N. ISR I.04, Project EU-SILC (Community statistics on income and living conditions) 2005/S 116-114302 – Lot 1 (Methodological studies to estimate the impact on comparability of the national methods used).
- Eurostat (2007a), Proposal on constructing the equivalised income variable for standardised computation of Laeken indicators. Report N. ISR I.11, Project EU-SILC (Community statistics on income and living conditions) 2005/S 116-114302 – Lot 1 (Methodological studies to estimate the impact on comparability of the national methods used).
- Eurostat (2007b), An examination of outliers at upper end of income distribution. Report N. ISR I.12, Project EU-SILC (Community statistics on income and living conditions) 2005/S 116-114302 – Lot 1 (Methodological studies to estimate the impact on comparability of the national methods used).
- Eurostat (2007c), An examination of outliers at upper end of income distribution (EU-SILC 2005). Report N. ISR I.16, Project EU-SILC (Community statistics on income and living conditions) 2005/S 116-114302 – Lot 1 (Methodological studies to estimate the impact on comparability of the national methods used).
- Van Kerm P. (2006), “Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC” Helsinki 6-7 November 2006.

Annex B: SAS codes with comments

```
/* SAS Program for an examination of outliers at upper end of income
distribution in EU-SILC 2007
Authors: Francesca Gagliardi, Giulia Ciampalini and Gianni Betti*/

/* Section 1: Construction of Tables 1-4 of the Report - WP DMQ 86 */

/* Construction of a household database with cross-sectional weight */
data h07; set silc07r2.h07;run;* silc07r2 is the library contained 2007
database;
data d07 (keep= hb020 hb030 DB090); set silc07r2.d07;
rename db020=hb020;
rename db030=hb030;
run;
proc sort data=h07;by HB020 hb030;run;
proc sort data=d07;by HB020 hb030;run;
data dh07; merge d07 h07;
by hb020 hb030;
run;

/* Construction of an output (Output0) containing:
sample size minimum mean median maximum and all percentile from 90 to 100 and
P_80 */
proc sort data=dh07;by HB020;run;
proc univariate data=dh07 noprint;
var HY020;
by HB020;
output out=output0 N=num MIN=min MEDIAN=median MEAN=mean MAX=max
pctlpre=P_ pctlpts=80,90 to 99 by 1 pctlpre=P_ pctlpts=99.1 to 100 by 0.1;
weight db090;
run;

/* Construction of the ratio between percentiles, mean, maximum and national
median (output1)*/
data output1; set output0;
R_P_80=(P_80/median);
R_P_90=(P_90/median);
R_P_91=(P_91/median);
R_P_92=(P_92/median);
R_P_93=(P_93/median);
R_P_94=(P_94/median);
R_P_95=(P_95/median);
R_P_96=(P_96/median);
R_P_97=(P_97/median);
R_P_98=(P_98/median);
R_P_99=(P_99/median);
R_P_99_1=(P_99_1/median);
R_P_99_2=(P_99_2/median);
R_P_99_3=(P_99_3/median);
R_P_99_4=(P_99_4/median);
R_P_99_5=(P_99_5/median);
R_P_99_6=(P_99_6/median);
R_P_99_7=(P_99_7/median);
R_P_99_8=(P_99_8/median);
R_P_99_9=(P_99_9/median);
R_P_100=(P_100/median);
R_mean= (mean/median);
R_max= (max/median);
run;
```

```

/* Construction of Tables 1-3 for Report */
data table1 (keep= HB020 num min median mean max R_mean R_P_80 R_P_90 R_P_95
R_P_99 R_P_99);
set output1; run;
PROC EXPORT DATA=work.table1
OUTFILE="C:\Documents and Settings\silc\Desktop\table1.xls"
DBMS=EXCEL
replace;run;
data table2 (drop= num -- P_100 R_P_80 R_P_99_1 -- R_P_99_9 R_mean R_max);
set output1;run;
PROC EXPORT DATA=work.table2
OUTFILE="C:\Documents and Settings\silc\Desktop\table2.xls"
DBMS=EXCEL
replace;run;
data table3a (drop= num -- P_100 R_P_80 -- R_P_98 R_mean R_max);
set output1; run;
PROC EXPORT DATA=work.table3a
OUTFILE="C:\Documents and Settings\silc\Desktop\table3a.xls"
DBMS=EXCEL
replace;run;
data table3b (keep= hb020 P_99 P_99_1 P_99_2 P_99_3 P_99_4 P_99_5 P_99_6 P_99_7
P_99_8 P_99_9 P_100 );
set output1; run;
PROC EXPORT DATA=work.table3b
OUTFILE="C:\Documents and Settings\silc\Desktop\table3b.xls"
DBMS=EXCEL
replace;run;

/* Construction of a file (output4) containing the number of households with
HY020 greater than specified multiples of national median and P_99 */
data output2 (keep= HB020 median num P_99); set output1; run;
proc sort data=output2; by hb020; run;
proc sort data=dh07; by hb020; run;
data output3; merge output2 dh07;
by hb020; run;
data output4; set output3;
if HY020 gt (2*median) then d2=1; else d2=0;
if HY020 gt (3*median) then d3=1; else d3=0;
if HY020 gt (4*median) then d4=1; else d4=0;
if HY020 gt (5*median) then d5=1; else d5=0;
if HY020 gt (6*median) then d6=1; else d6=0;
if HY020 gt (7*median) then d7=1; else d7=0;
if HY020 gt (8*median) then d8=1; else d8=0;
if HY020 gt (9*median) then d9=1; else d9=0;
if HY020 gt (10*median) then d10=1; else d10=0;
if HY020 gt (12*median) then d12=1; else d12=0;
if HY020 gt P_99 then d13=1; else d13=0;
run;

/* Construction of Table 4 for the Report */
proc univariate data=output4 noprint;
var d2 d3 d4 d13 d5 d6 d7 d8 d9 d10 d12;
by HB020;
output out=table4 N=num sum=HY020_2m sum=HY020_3m sum=HY020_4m sum=HY020_P99
sum=HY020_5m
sum=HY020_6m sum=HY020_7m sum=HY020_8m sum=HY020_9m sum=HY020_10m sum=HY020_12m
;
run;
PROC EXPORT DATA=work.table4
OUTFILE="C:\Documents and Settings\silc\Desktop\table4.xls"
DBMS=EXCEL
replace;run;

```

```
/* Section 2: Construction of Histogram, Table B, input for Table 5-6 of the
Report - WP DMQ 86 */
```

```
/* Merging h file with d file (weights) and p file (personal incomes
summed up at household level)*/
data h; set silc07r2.h07; where hb020 eq 'LU'; run; ****Change country's name;
proc sort data=h; by hb030; run;
data weight; set silc07r2.d07 (keep= db020 db030 db090); where db020 eq
'LU';****Change country's name;
rename db020=hb020; rename db030=hb030; run;
proc sort data=weight; by hb030; run;
data p; set silc07r2.p07; where pb020 eq 'LU'; ****Change country's name;
run;
proc sort data=p; by px030; run;
proc univariate data=p noprint;
var py010g py020g py035g py050g py070g py080g py090g py100g py110g py120g
py130g py140g;
output out=sum_p_income
sum=hpy010g hpy020g hpy035g hpy050g hpy070g hpy080g hpy090g hpy100g hpy110g
hpy120g hpy130g hpy140g;
by pb020 px030;
run;
data p_income; set sum_p_income;
rename pb020=hb020;
rename px030=hb030; run;
proc sort data=p_income; by hb020 hb030; run;
data h_file; merge h weight p_income; by hb020 hb030; run;
proc univariate data=h_file noprint;var hy020;
output out=percentile P99=hy020_P99 median=median_hy020 n=num_hhs; by hb020;
weight db090; run;
data perc; merge h_file percentile;by hb020; run;
data file_p99; set perc; where hy020 ge hy020_P99; run;

/* Histogram P99 (very large data) */
proc univariate data=file_p99 noprint;var hy020;
histogram / endpoints= 185000 to 3700000 by 20000;
where hb020 eq 'LU';****Change country's name;
title 'LU';****Change country's name;
run;

/* Identification of the 'outliers' and table with
all components of income (Table B)*/
data silc07r2.TableB_lu; ****Change data set's name;
set file_p99 (keep=hb020 hy010 hy020 hy040g hy090g hy050g hy060g hy070g hy080g
hy110g
hx090 hpy010g hpy020g hpy035g hpy050g hpy070g hpy080g hpy090g hpy100g hpy110g
hpy120g hpy130g hpy140g median_hy020 num_hhs);
ratiohy020_hy010=hy020/hy010 ;
where hb020 eq 'LU' and hy020 gt 450000; ****Change country's name and value;
run;
PROC EXPORT DBMS=excel2002 DATA= silc07r2.tableB_lu /*Change data set's name*/
OUTFILE= "D:\new eu-silc project 2009\Reports\output task
2\tableB_lu.xls" REPLACE;****Change directory's name;
RUN;

/* TRIMMING
(i) Leaving all extreme (outliers and very large) values unchanged
(ii) Removing outliers from the dataset
(iii) Trimming the outliers to the value immediately below the smallest outlier
(iv) Trimming all the very large values to the value immediately below the lower
boundary */
```

```

data trim_i; set perc; dispinc=hy020; eqinc_or=dispinc/hx050; run;
data trim_ii; set perc;where (hb020 eq 'LU' and hy020 lt 450000); ****Change
country's name and value;
dispinc=hy020; eqinc_or=dispinc/hx050; run;
data trim_iii; set perc;dispinc=hy020;
if hb020 eq 'LU' and hy020 gt 450000 then dispinc=355503;****Change country's
name and values;
eqinc_or=dispinc/hx050; run;
data trim_iv; set perc; dispinc=hy020;
if hb020 eq 'LU' and hy020 gt 187637 then dispinc=186398; ****Change country's
name and values;
eqinc_or=dispinc/hx050; run;

*****6. S80/S20: ratio of income shares of the percentiles;
%macro stat6;
data working;set working; wj=db090; run;
data input_bound;set working;run;
%perc_bound(80);
data output_bound_80;set output_bound;rename y_perc=y_perc_80;run;
%perc_bound(20);
proc means data=working noprint; output out=media mean=media;var eqinc;weight
wj;by country;run;
data working1;merge working output_bound output_bound_80 media;by country;
alfai_80=sign(max(0,(eqinc-y_perc_80)));alfai_20=1-sign(max(0,(eqinc-y_perc)));
z_80=(eqinc/media)*alfai_80;z_20=(eqinc/media)*alfai_20;run;
proc univariate data=working1 noprint; output out=est mean=est_80 mean=est_20;
var z_80 z_20;weight wj;by country;run;
data est_s80s20 (drop=est_80 est_20); set est;s80s20_&j=est_80/est_20;run;
title 'stat6';
%mend;

*****12. Gini;
%macro stat12;
data working; set working; wj=db090; run;
proc means data=working noprint; output out=media mean=media;var eqinc;weight
wj;by country;run;
data media;set media (keep = media country);run;
proc sort data=working;by eqinc;run;
proc iml;use working;read all var {hb030} into hb030;read all var {wj} into
weig;read all var {eqinc} into eqinc;
num=nrow(hb030);
share_w=repeat(0,num);share_inc=repeat(0,num);eqinc_w=eqinc#weig;
tot_w=sum(weig);tot_inc=sum(eqinc_w);
i=1; do while (i<num+1);
share_inc[i]=sum(eqinc_w[1:i])/tot_inc;
share_w[i]=sum(weig[1:i])/tot_w;
i=i+1; end;
create share var {hb030 share_inc share_w};append;close share;quit;run;
proc sort data=working;by hb030;run;proc sort data=share;by hb030;run;
data working1;merge working share;by hb030;wj=db090;run;
data working2;merge working1 media;by country;z=2*((eqinc-
media)/media)*share_w;run;
proc univariate data=working2 noprint;output out=est_gini mean=gini&j;var
z;weight wj;by country;run;
title 'stat12';
%mend;

*****Perc bound;
%macro perc_bound (perc);
data input;set input_bound;percent=&perc;run;
proc sort data=input;by eqinc;run;
proc iml;

```

```

use input;read all var {hb030} into hb030;read all var {wj} into wj;
read all var {eqinc} into eqinc;read all var {percent} into percent;
read all var {country} into country_vec;
num=nrow(hb030); share=repeat(0,num);ratio=percent[1]/100;
tot=sum(wj);
  i=2; do while (i<num+ 1);
    share[i]=sum(wj[1:i])/tot;
    if (share[i-1] < ratio) & (share[i] >= ratio)
      then
        y_perc=eqinc[i-1]+(eqinc[i]-eqinc[i-1])*((ratio-share[i-
1])/((share[i]-share[i-1])));
        i=i+1; end;
country=country_vec[1];
create output_bound var {y_perc country};append ;close output_bound;quit;run;
%mend;

/* Estimation of s80/s20 and Gini for the four kinds of trimming and bottom
coding
of the equivalised disposable income distribution at 15% of the median */
%MACRO CICLO;
  %do J=1 %to 4;
%if &j=1 %then %do;
proc univariate data=trim_i noprint; var eqinc_or;
output out=median_i median=median_eqinc; weight db090; by hb020; run;
data working; merge trim_i median_i; by hb020; country=1;
med_15=0.15*median_eqinc; eqinc=eqinc_or; if eqinc_or lt med_15 then
eqinc=med_15;run;
%stat6; data est_s80s20_&j; set est_s80s20; run;
%stat12; data est_gini_&j; set est_gini; run;
%end;
%if &j=2 %then %do; proc univariate data=trim_ii noprint; var eqinc_or;
output out=median_ii median=median_eqinc; weight db090; by hb020; run;
data working; merge trim_ii median_ii; by hb020; country=1;
med_15=0.15*median_eqinc;eqinc=eqinc_or; if eqinc_or lt med_15 then
eqinc=med_15;run;
%stat6; data est_s80s20_&j; set est_s80s20; run;
%stat12; data est_gini_&j; set est_gini;run;
%end;
%if &j=3 %then %do; proc univariate data=trim_iii noprint; var eqinc_or;
output out=median_iii median=median_eqinc; weight db090; by hb020; run;
data working; merge trim_iii median_iii; by hb020; country=1;
med_15=0.15*median_eqinc;eqinc=eqinc_or; if eqinc_or lt med_15 then
eqinc=med_15;run;
%stat6; data est_s80s20_&j; set est_s80s20; run;
%stat12; data est_gini_&j; set est_gini; run;
%end;
%if &j=4 %then %do; proc univariate data=trim_iv noprint; var eqinc_or;
output out=median_iv median=median_eqinc; weight db090; by hb020; run;
data working; merge trim_iv median_iv; by hb020; country=1;
med_15=0.15*median_eqinc; eqinc=eqinc_or; if eqinc_or lt med_15 then
eqinc=med_15;run;
%stat6; data est_s80s20_&j; set est_s80s20;run;
%stat12; data est_gini_&j; set est_gini;run;
%end;
%end;
%mend;
%ciclo;
data silc07r2.s80s20_lu ;****Change data set's name;
merge est_s80s20_1 est_s80s20_2 est_s80s20_3 est_s80s20_4; by country;
hb020='LU';****Change country's name;run;
data silc07r2.gini_lu ;****Change data set's name;
merge est_gini_1 est_gini_2 est_gini_3 est_gini_4; by country;
hb020='LU'; run;****Change country's name;

```

```

/* Section 3: Construction of Tables 5-6 of the Report - WP DMQ 86 */

*****table 5;
data table; set silc07r2.tableb_dk silc07r2.tableb_fr silc07r2.tableb_ie
silc07r2.tableb_lu
silc07r2.tableb_pt silc07r2.tableb_uk; run;
proc sort data=table; by hb020;
proc univariate data=table noprint; var hy020;
output out=values min=hy020_min max=hy020_max n=num_hhs_out; by hb020; run;
data med (keep=hb020 num_hhs median_hy020); set table;
if first.hb020; by hb020; run;
data silc07r2.table_5; merge med values; by hb020;
boundary_outliers=hy020_min/median_hy020;
percent_hhs_outliers=(num_hhs_out/num_hhs)*100;
run;
PROC EXPORT DBMS=excel2002 DATA= silc07r2.table_5
OUTFILE= "D:\new eu-silc project 2009\Reports\output task 2\table_5
.xls" REPLACE;****Change directory's name;
RUN;
*****table 6a;
data silc07r2.table_6a (drop=country); set silc07r2.gini_dk silc07r2.gini_fr
silc07r2.gini_ie silc07r2.gini_lu
silc07r2.gini_pt silc07r2.gini_uk;
ratio_2vsl=gini2/gini1;
ratio_3vsl=gini3/gini1;
ratio_4vsl=gini4/gini1;
run;
PROC EXPORT DBMS=excel2002 DATA= silc07r2.table_6a
OUTFILE= "D:\new eu-silc project 2009\Reports\output task
2\table_6a.xls" REPLACE;****Change directory's name;
RUN;
*****table 6b;
data silc07r2.table_6b (drop=country); set silc07r2.s80s20_dk silc07r2.s80s20_fr
silc07r2.s80s20_ie
silc07r2.s80s20_lu silc07r2.s80s20_pt silc07r2.s80s20_uk;
ratio_2vsl=s80s20_2/s80s20_1;
ratio_3vsl=s80s20_3/s80s20_1;
ratio_4vsl=s80s20_4/s80s20_1;
run;
PROC EXPORT DBMS=excel2002 DATA= silc07r2.table_6b
OUTFILE= "D:\new eu-silc project 2009\Reports\output task
2\table_6b.xls" REPLACE;****Change directory's name;
RUN;

```