Accurate Income Measurement for the Assessment of Public Policies (AIM-AP)
Contract no 028412.

Workpackage 3.2:
Harmonization of the budget and income surveys

André Decoster[*], Kris De Swerdt[*], Jason Loughrey[†], Cathal O'Donoghue[†], Dirk Verwerft[*]

Abstract: This paper constitutes the second deliverable of project 3 of the AIM-AP project (Contract no 028412). The content of this paper describes the datasets that will be used for the actual imputation of expenditures into the EUROMOD income datasets in Workpackage 3.5. In particular, a comparison is made between the matching variables in the budget and income dataset for each country in order to ensure that these common variables have the same meaning in both datasets. To that end, a distributional analysis will be performed. Finally, the comparability between countries, especially concerning the disposable income concept and the harmonisation of expenditure variables, is discussed.

I. INTRODUCTION

In order to make imputations from one dataset into another, it is important that they both cover the same population and that the common characteristics are defined in the same way. The information drawn from the source (budget) dataset has to be as relevant and adequate as possible for the target (income) dataset observations. That is why the first part of this Workpackage provides a comparison of populations, sampling frames and variables of the source and target surveys, structured by country. The countries studied here are Belgium, Hungary, Ireland and the UK. Simulations will also take place for Greece, but

[*] University of Leuven

[†] Rural Economy Research Centre, Teagasc

since the Greek dataset contains both the necessary budget and income data, no matching procedure is necessary. It can therefore be used as a benchmark case to evaluate and interpret differences in simulation results.

For each country, a description of the populations in both datasets is given, with special consideration for the comparability of geographical coverage and sample years in both datasets. A description of the non-response, together with the correlation between non-response and socio-demographic characteristics of the sample units, would also be very useful in this respect. However, the registration and provision of this kind of information is far from standard in socio-economic surveys. In most cases, the inclusion of a weight variable correcting for sample bias on the basis of external, aggregate information on properties like socio-economic status, age and sex of the household head, is the best one can hope for. The identification of such a weight variable is the third component of the population descriptions in the next paragraph. All this leads to question marks as to how one should proceed whenever the two sample populations are not the same. Since the income datasets are already integrated in EUROMOD and have to be treated as given, it is the budget datasets that have to be adjusted. Leaving out observations in regions that are not included in the income datasets, using a deflator to rescale monetary variables to the year of the income surveys and making use of weight variables depending on socio-economic and demographic characteristics are the main strategies to counter sampling errors.

A second important topic is the harmonisation of common variables. Common variables are the ones that are present in both the source and target dataset and will be used to match observations between the datasets. Since the matching will occur on the basis of a measure of resemblance and "closeness" in these variables, it is essential that the variables used measure the same characteristic. For some variables, like the number of persons in a household, this is not expected to be very problematic.[1] Nevertheless, they will be included in the analysis as an additional test for the resemblance of the two datasets. The method adopted here is to present first a synthetic test of distributional equality like the Kolmogorov-Smirnov and the Kruskal-Wallis tests for quantitative variables and the chi-square test for qualitative variables. However, these tests are often too strict and will too easily reject the null hypothesis of equality or independence (of the sample). Moreover, they do not tell how strong a possible dependence on the sample is and where (in which intervals or categories) the problem is situated. Therefore another analysis of equalities and differences is performed by looking at QQ-plots and the comparison of decile means for quantitative variables and the chi square contributions for the qualitative ones. This provides

[1] Although a significant difference in the number of persons in a household can point to a divergence in the definition of what a household is.

clues for solving problems that present themselves or draws attention to categories or intervals where the matching procedure will encounter difficulties.

As for the selection of the common variables, the presence of closely defined variables in both datasets is the only criterion used. No use is being made of selection procedures like the stepwise regression method. The rationale behind this choice is that in the context of matching data, the addition of new adequate common variables provide additional percentages of explained variance which can help to make the matching more accurate. The standard regression trade off between this issue and building in efficiency and simplicity into the model is not relevant here because the aim is not to construct a model that is easy to interpret.

Some imputation methods are based on clustering the population into groups based upon some common variables.[2] Since this clustering depends on theoretical considerations as well as practical issues, such as the fact that the number of observations in each cluster should be high enough, these clusters have been determined a priori.[3] Because this can vary from country to country and according to the interests of the researcher, no automatic procedure for creating clusters has been installed. The advice towards an implementation of matching techniques using clusters into EUROMOD is that the definition of clusters should be made flexible and dependent on the input of EUROMOD users.

Having analysed the common variables, the next step is to offer some cross-national comparisons. The disposable income concept is especially relevant here, since this will serve as input for the indirect tax module in EUROMOD and as such has to be standardised across countries. But also some general problems common to all or most of the countries will be summed up. In the end, the definition of expenditure aggregates is discussed shortly, but the main discussion of this topic takes place in Workpackage 3.3.

The structure of this Workpackage is as follows. First the tests used for the analysis of common variables are described and discussed. Afterwards sample designs and common variables are compared between the datasets for each country separately. Finally, some remarks are given concerning cross-national comparability.

---

[2] See Workpackage 3.4 for an overview and description of all the methods tested.

[3] In Workpackage 3.4, the population is divided on the basis of age of the household head (younger than 35, between 35 and 55, older than 55), whether or not there are children present in the household and whether or not the household head is economically active.

II. METHODOLOGY

This section discusses the statistical tests used to compare distributions across different datasets. Firstly the tests for equality of quantitative variables (in the case of disposable income) are described, and secondly some measures for the equality of proportions for qualitative variables are proposed.

Disposable income is compared by means of the nonparametric Kolmogorov-Smirnov and Kruskal-Wallis tests. The choice for nonparametric methods can be justified by the fact that income is not normally distributed,[4] and moreover, by the presence of a lot of influential outliers in the data, which tend to provoke violations of the normality assumption. The Kolmogorov-Smirnov test is based upon the cumulative distribution functions of the source and target groups. It measures the largest vertical distance between these functions, for which an asymptotical distribution has been derived. Relying on the distribution function reduces the impact of outliers compared to for instance the standard t-test. It must be noted that the distribution function will not be affected too much in the presence of few outliers. The same is true for the Kruskal Wallis test, which ranks all observations of both groups according to the selected variable, and then compares the mean rank in the source dataset with that in the target dataset. Counting only the rank of an outlier rather than its absolute value also reduces its influence in the final measure. The main difference between the two tests is that the former is a distributional test, whereas the latter is a location test: not a means but a mean ranks test.

As already mentioned, these tests do not give an indication of how strong the inequality of common variable values is between the datasets and whether the problems are situated in particular intervals. A graphical analysis by means of quantile-quantile-plots, henceforth referred to as QQ-plots, is added to provide a closer indication about the degree of inequality. In this kind of graphs, the quantiles of disposable income in the budget dataset are plotted against the quantiles of disposable income in the income dataset. Perfect equality between distributions would result in a 45-degree line, whereas deviations from equality appear as deviations from this line. In this manner, problematic intervals can be identified. The same patterns can be expressed in a tabular form by comparing the quantile values or the decile means between the two populations.

For qualitative variables, the Pearson chi-square test for the equality of proportions can be used as a synthetic measure. The great advantage of this test is the easy extension to a more analytical tool by including the cells' contributions to the chi-square statistic by means

[4] Often the logarithm of the income is assumed to be normal, but this transformation does not consider negative and zero incomes and does not deal with the outlier argument.

of the Pearson residuals. In this workpackage, the standardized Pearson residuals will be used:

$$e_{ij} = \frac{n_{ij} - \mu_{ij}}{\sqrt{\mu_{ij}(1 - p_{i+})(1 - p_{+j})}} \qquad (2.1)$$

In this expression, n_{ij} is the observed frequency in cell ij and μ_{ij} the expected frequency under the null hypothesis of independency. Furthermore, p_{i+} is the proportion of observations in row i and p_{+j} the proportion of observations in column j. These standardized residuals have an asymptotically standard normal distribution. Hence, values between -2 and 2 indicate that the cell frequency is in line with the null hypothesis, while cells with an absolute value larger than two can be considered problem categories. Agresti[5] relaxes the rejection level to an absolute value of between 2 and 3. A significantly negative residual marks an underrepresentation of the group (row) in the dataset (column), while a significantly positive residual marks an overrepresentation. This allows for identifying problem cells in the contingency table.

With these statistical tools, an analysis for each country separately will be carried out. This is the subject of the next paragraph. The common variables included are disposable income, number of persons, actives and children in the household, age, gender and professional status of the household head and region.

III. COMPARISONS WITHIN COUNTRIES

1. Belgium

For Belgium, the PSBH 2001 is being used as the income dataset, whereas the National Budget Survey of 2004 serves as the budget dataset. Both surveys should be representative for the entire Belgian population. All three regions in the country were present in the sample. However, the difference in sample year means that monetary variables have to be adjusted. Since the PSBH2001 is part of EUROMOD, the budget disposable income variable will be deflated with the value 0.948575 to be expressed in 2001

[5] Agresti, A. (2002) *Categorical Data Analysis. Second Edition.* John Wiley & Sons, Inc.: Hoboken, NJ, USA, p. 81.

prices. This value was derived from inflation data provided by the National Bank of Belgium.[6]

Disposable income in the budget dataset includes imputed rents for home owners, and the latter also appear as housing expenditures to make the balance right. For reasons of international comparability of the income concept in EUROMOD the decision has been made to subtract imputed rents from the disposable income variable (and of course also leave it out of the expenditures).[7] Moreover, a weight variable was identified in both datasets.

Table 1.1 gives the Kolmogorov-Smirnov and Kruskal-Wallis tests for disposable income and the chi-squared tests for the qualitative variables. The p-values in brackets behind the test statistics give the probability of the given statistic under the null hypothesis that both samples are drawn from the same underlying population. In other words, a low p-value means that the distribution of the variables is dependent on the sample, so there is a difference in distribution between the two samples. Except for gender of household head, independence of the sample is rejected everywhere on a significance level of 0.01. This does not mean necessarily that matching both datasets will be problematic, firstly since the matching between observations will occur conditional upon these variables and hence e.g. the over- or underrepresentation of certain groups will have no influence on the process. Only procedures where the matching is constrained, in the sense that observations in the budget dataset are assigned to observations in the income dataset without replacement (so one can use each observation only once), can be biased. In WP 3.4, the latter methods will be implemented in such a way that matching can only take place within the same category of households. These categories will be defined upon one or more of the variables described here. As a consequence, some conditionality will also be included in these methods. Secondly, the non-parametric tests used are known to be very – and perhaps too – strict, as stated before, so their outcome should be interpreted as an indication rather than a sound proof.

[6] The consumer price indices were used here: 114.93 for 2004 and 109.02 for 2001 with 1996 as base year.

[7] See WP 3.3.

Table 1.1: Common Variables for Belgium

Common Variable Name	χ^2 / Kolmogorov-Smirnov test value (p-value)	Kruskal-Wallis test value (p-value)
Disposable Income	0.1073 (0.000)	84.738 (0.0001)
Number of persons	135.0990 (0.000)	
Number of active persons in household	48.8982 (0.000)	
Number of children	394.0977 (0.000)	
Age of household head	135.6497 (0.000)	
Gender of household head	6.2702 (0.012)	
Region	85.7877 (0.000)	
Socio-economic status of household head	131.8116 (0.000)	

Tables 1.2 and 1.3 allow a closer look at the disposable income comparison, unequivalised and equivalised respectively. Equivalised income is calculated here as household disposable income divided by the square root of the number of household members. Imputed rents were subtracted, as it was decided that they should be left out of consideration. In the tables, the mean unequivalised and equivalised disposable incomes for each equivalised disposable income decile are compared across datasets. It is clear that the values presented are not very far apart, especially the equivalised ones[8], and this more or less contradicts the statistical tests used earlier. The reason for the divergent results is revealed in figure 1.1, which shows the QQ-plot for disposable income in both datasets and as such is a generalisation of table 1.3. Most of the quantiles lie on the 45-degrees line. About 98.7% of the data lie between 1000 and 100000 Euros and are thus represented by this straight- line component. The problems arise for the low and the very high incomes. These will in turn be discussed below.

[8] When the mean unequivalised disposable incomes are calculated per unequivalised disposable income decile, the differences are also minor, comparable to the results in table 1.3.

Table 1.2: Mean yearly disposable income per equivalised disposable income decile for Belgium (EUR)

Disposable Income decile	Mean disposable income in budget survey	Mean disposable income in income survey
1	8.388	7.866
2	13.630	16.550
3	16.724	20.115
4	19.598	23.860
5	22.637	26.281
6	26.468	31.636
7	31.766	34.034
8	35.651	38.740
9	41.616	45.709
10	63.962	68.523

Table 1.3: Mean yearly equivalised disposable income per equivalised disposable income decile for Belgium (EUR)

Disposable Income decile	Mean disposable income in budget survey	Mean disposable income in income survey
1	6.322	6.208
2	10.349	11.484
3	12.295	13.695
4	14.162	15.616
5	16.177	17.576
6	18.196	19.597
7	20.624	21.692
8	23.715	24.295
9	28.233	28.009
10	43.704	42.859

Figure 1.1: QQ-plot for disposable income quantiles in budget dataset (dib) versus quantiles in income dataset (dii), Belgium (Euros per year)

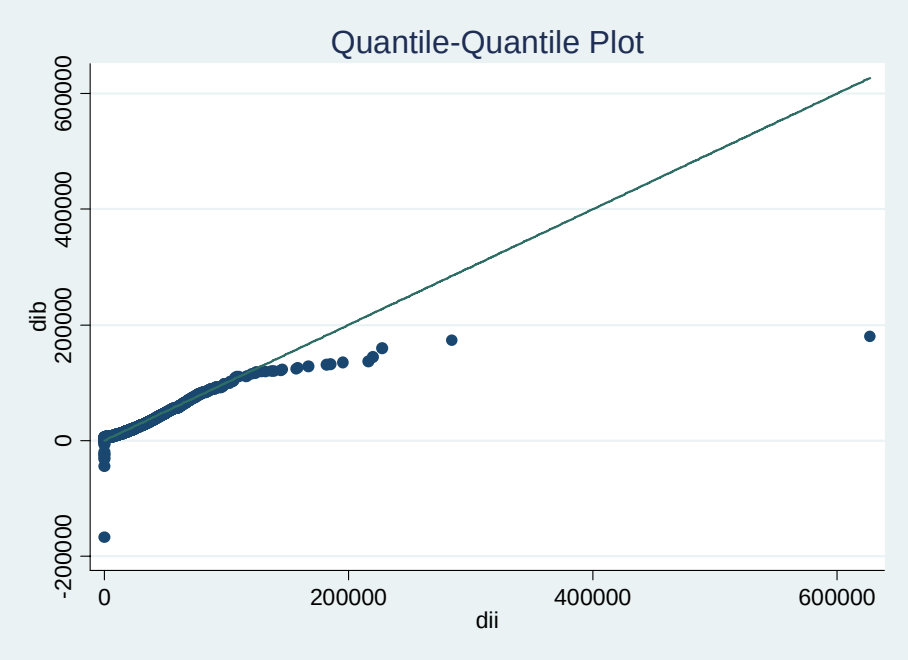
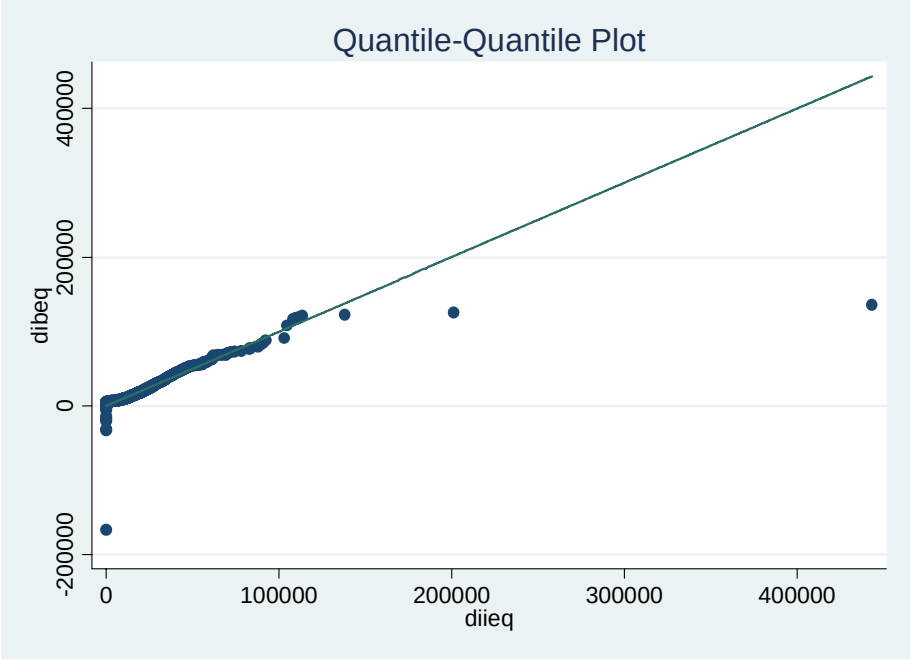


Figure 1.2: QQ-plot for equivalised disposable income quantiles in budget dataset (dibeq) versus quantiles in income dataset (diieq), Belgium (Euros per year)



In the budget dataset, some (large) negative incomes are present, while in the PSBH 2001 negative incomes are absent. This is probably a consequence of the fact that in the

budget survey incomes are registered for only one month, whereas this period is one year for the PSBH. The budget data are thus expected to show a higher degree of volatility due to e.g. temporary losses of the self- employed. The question arises about what can be done to counter this effect. The approach adopted in the later steps of the project is just to drop the negative incomes, for three reasons. Firstly, there are no negative incomes in the income dataset. It seems a strange idea to match a household with a negative income to one with a positive income, no matter how alike they are in other respects. Secondly, because of the interpretation above, negative incomes can be considered to be due to temporary circumstances. It is a reasonable expectation that of each "household type"[9] in the population there will be one household present in the budget dataset that - more or less - earns the mean income over time of that type.[10] For unconstrained matching, dropping the negative incomes means that only (some of) the extremes over time are neglected, whereas no information on household types is lost. Of course, the distribution of income for a household type is changed in the budget dataset, and especially parametric methods that estimate a model on the dataset would be influenced. But this effect is judged to be rather small because of the small number of observations involved. Thirdly, most parametric methods, that use Engel curves for the matching, will perform better with logarithmic specifications. So negative and zero incomes have to be dropped. Thus, there is a trade off between keeping the information in negative incomes and gaining accuracy by using a better model.

For the highest quantiles, the income dataset contains higher incomes than the budget dataset. An explanation might be an overrepresentation in the income versus the budget dataset of household types with higher incomes, such as large households. For instance, equivalising disposable income by means of the OECD scale, gives figure 1.2 where the QQ- plot coincides much better with the 45-degrees line. This gives some additional evidence that using the number of persons in the household as a common matching variable is important.

The results for some qualitative variables are summarised in tables 1.4 – 1.8. The first value in each cell is the number of observations in the cell, whereas the second value

---

[9]Household type refers here to an ideal typical group of households that are in some way "alike". It is thus used more as a thought construct than as a concrete specification of relevant characteristics. Mean income over time, however, will be an important component if consumer behaviour of the household types is to be simulated or modelled.

[10] The length of the time is here one year, which is the registration period in the PSBH. Although during an entire lifetime, one cannot have negative average income, this is still possible during one year. The assumption made before dropping the negative incomes is thus that the time volatility of the income of one type considered per year is much less than the volatility on a monthly basis.

gives the standardized Pearson residual. The column and row totals are also given. Looking at the magnitude of the Pearson residuals, it is clear that single-person households have the largest contribution to the chi- square test statistic. There seems to be an overrepresentation of those households in the budget dataset, of course with respect to the income dataset rather than to the real population. This is in line with the findings in the QQ- plots above. As for the region variable (table 1.5), Brussels is overrepresented and Flanders underrepresented in the budget dataset. The gender of the household head can be considered to be independent of the sample (table 1.6), whereas the budget dataset contains relatively more households with older household heads (table 1.7) and less household heads in the rest category of professional status (table 1.8).

Table 1.4: Frequencies and Pearson residuals for number of household members, BE

Np	Income dataset	Budget dataset	Total
1	1,040	1,537	2,577
	-11.0	11.0	
2	1,114	1,034	2,148
	3.0	-3.0	
3	590	487	1,077
	4.0	-4.0	
4	603	464	1,067
	5.2	-5.2	
5	236	182	418
	3.1	-3.1	
6	53	64	117
	-0.8	0.8	
7	8	11	19
	-0.6	0.6	
8	6	5	11
	0.4	0.4	
9	3	1	4
	1.0	-1.0	
10	1	0	1
	1.0	-1.0	
Total	3,654	3,785	7,439

Table 1.5: Frequencies and Pearson residuals for region, BE

region	Income dataset	Budget dataset	Total
Brussels	393 -9.1	691 9.1	1,084
Flanders	1,961 5.0	1,810 -5.0	3,771
Wallonia	1,300 1.5	1,284 -1.5	2,584
Total	3,654	3,785	7,439

Table 1.6: Frequencies and Pearson residuals for gender of household head, BE

Sexr	Income dataset	Budget dataset	Total
Male	2,546 2.5	2,535 -2.5	5,081
Female	1,108 -2.5	1,250 2.5	2,358
Total	3,654	3,785	7,439

Table 1.7: Frequencies and Pearson residuals for age groups, BE

Agegroup	Income dataset	Budget dataset	Total
< 35	711 6.0	540 -6.0	1,251
35-55	1,594 7.1	1,345 -7.1	2,939
> 55	1,349 -11.5	1,900 11.5	3,249
Total	3,654	3,785	7,439

Table 1.8: Frequencies and Pearson residuals for professional status, BE

socstat	Income dataset	Budget dataset	Total
Other	355 9.1	165 -9.1	520
Self-employed	243 3.4	181 -3.4	424
Employed	1,828 -5.6	2,140 5.6	3,968
Unemployed	194 -4.4	298 4.4	492
Retired	999 1.2	987 -1.2	1,986
Student	35 3.1	14 -3.1	49
Total	3,654	3,785	7,439

As stated before, all the variables in this list will be used in the matching process and hence an overrepresentation of one group does not necessarily constitute a real problem because conditionality will be built in. This is clearly so for the regional variable which has a straightforward meaning so that households cannot be classified differently in both datasets. An overrepresentation of a small region like Brussels is common practice to ensure that significant conclusions can be drawn for the specific subpopulation. The same can be said for the age group of the household head. There are, however, two caveats. First, when it comes to household definition a larger number of single-person households may point to the fact that some forms of cohabitation are classified differently across the samples. But note that this can also be a consequence of the higher proportion of older household heads in the National Budget Survey, since older households have a higher probability of being single-person due to the death of a partner. Secondly, the "other" category of professional status is relatively larger in the National Budget Survey, which may be caused by diverging interpretations about this status for some households.

2. Hungary

For Hungary, the EU-SILC 2004 is used as the income dataset and the Household Budget Survey of 2004 as the budget dataset. No deflator is necessary in this case. The

regional variable for Hungary was excluded since the NUTS1 classification does not coincide with socio-economic divisions in the country. Weight variables were identified in both datasets.

The picture of the common variables is more or less the same as for Belgium. The absence of sample effects is rejected by the Kolmogorov-Smirnov, Kruskal-Wallis and chi-squared tests (table 2.1). But for disposable income, the decile mean tables (2.2 and 2.3) and QQ- plots (figure 2.1) show the reverse. Of the entire population, 99.8% is on the straight line component (between 0 and 40000 Euros). There is no problem with lower quantiles, although the budget data was measured on a monthly basis and the income data on a yearly basis as in the Belgian case. However, the income of the highest quantiles is also underestimated in the HBS with respect to the EU-SILC. Contrary to the Belgian case, equivalisation does not help to make this effect disappear (figure 2.2).

Table 2.1: Common Variables for Hungary

Common Variable Name	χ^2 / Kolmogorov-Smirnov test value (p-value)	Kruskal-Wallis test value (p-value)
Disposable Income	0.0995 (0.000)	165.833 (0.0001)
Number of persons	55.2658 (0.000)	
Number of active persons in household	88.9373 (0.000)	
Number of children	851.8378 (0.000)	
Age of household head	247.3058 (0.000)	
Gender of household head	27.4207 (0.000)	
Education level of household head	700.7824 (0.000)	
Socio-economic status of household head	154.0963 (0.000)	

Table 2.2: Mean yearly disposable income per equivalised disposable income decile for Hungary (EUR)

Disposable Income decile	Mean disposable income in budget survey	Mean disposable income in income survey
1	2.576	2.317
2	3.713	3.403
3	4.506	3.999
4	5.262	4.661
5	6.178	5.380
6	6.990	6.192
7	8.038	6.938
8	9.212	8.093
9	10.974	9.792
10	16.988	15.897

Table 2.3: Mean yearly equivalised disposable income per equivalised disposable income decile for Hungary (EUR)

Disposable Income decile	Mean disposable income in budget survey	Mean disposable income in income survey
1	1.780	1.572
2	2.572	2.330
3	3.012	2.774
4	3.410	3.174
5	3.834	3.536
6	4.294	3.915
7	4.819	4.354
8	5.468	4.940
9	6.476	5.821
10	10.310	9.526

Figure 2.1: QQ-plot for disposable income quantiles in budget dataset (dib) versus quantiles in income dataset (dii), Hungary (Euros per year)

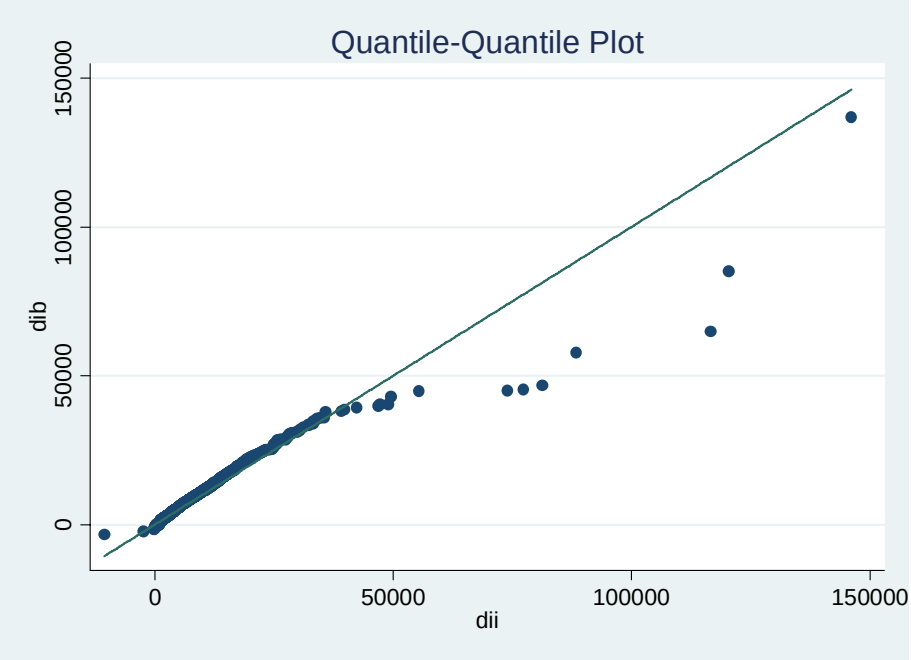
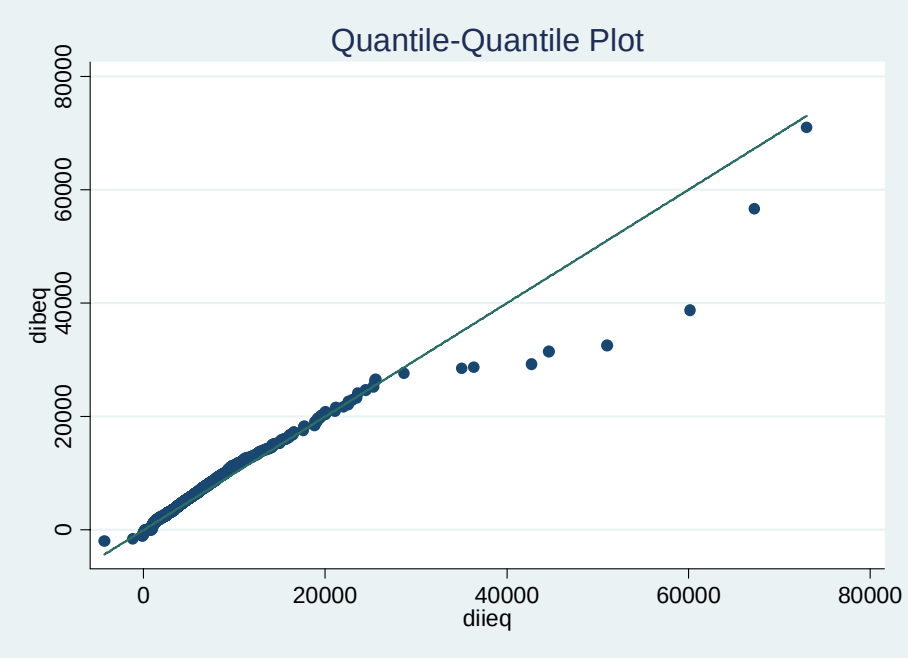


Figure 2.2: QQ-plot for equivalised disposable income quantiles in budget dataset (dibeq) versus quantiles in income dataset (diieq), Hungary (Euros per year)



Relatively more singles and couples are present in the EU-SILC, whereas the opposite is true for households with 3 or 4 members (table 2.4). The EU-SILC has

significantly more households with older and female household heads than the HBS, (tables 2.5 and 2.6), and probably correlated with this age effect also relatively more retired household heads to the expense of employed and self-employed heads in particular.

Table 2.4: Frequencies and Pearson residuals for number of household members, HU

np	Income dataset	Budget dataset	Total
1	1,721 4.2	1,918 -4.2	3,639
2	2,137 3.3	2,474 -3.3	4,611
3	1,303 -3.4	1,832 3.4	3,135
4	1,123 -4.5	1,652 4.5	2,775
5	427 -1.5	587 1.5	1,014
6	139 0.1	172 -0.1	311
7	40 0.8	42 -0.8	82
8	22 1.1	20 -1.1	42
9	9 0.7	8 -0.7	17
10	3 -0.4	5 0.4	8
11	1 1.1	0 -1.1	1
12	1 1.1	0 -1.1	1
14	1 1.1	0 -1.1	1
Total	6,927	8,710	15,637

Table 2.5: Frequencies and Pearson residuals for gender of household head, HU

sexr	Income dataset	Budget dataset	Total
Male	4,680 -5.2	6,222 5.2	8,469
Female	2,247 5.2	2,488 -5.2	7,168
Total	6,927	8,710	15,637

Table 2.6: Frequencies and Pearson residuals for agegroups, HU

agegroup	Income dataset	Budget dataset	Total
< 35	1,160 0.9	1,413 -0.9	2,573
35-55	2,568 -14.8	4,258 14.8	6,826
> 55	3,199 14.3	3,039 -14.3	6,238
Total	6,927	8,710	15,637

Table 2.7: Frequencies and Pearson residuals for professional status, HU

socstat	Income dataset	Budget dataset	Total
Other	265 -1.5	376 1.5	641
Self-employed	385 -5.1	662 5.1	1,047
Employed	2,783 -9.0	4,128 9.0	6,911
Unemployed	358 2.1	386 -2.1	744
Retired	3,082 11.2	3,104 -11.2	6,186
Student	54 1.2	54 -1.2	108
Total	6,927	8,710	15,637

3. Ireland

For Ireland, the Living in Ireland Survey 2000 is used as the income dataset and the Household Budget Survey of 1999/2000 as the budget dataset. No deflator is necessary as the reference period for income is similar for both surveys. Amounts are converted from Irish pounds to euros at the rate of €1 = IR£0.787564 since the survey period is before the introduction of the Euro.

The overlapping variables appear to have a distribution of values similar in both datasets as shown by the chi-squared tests (table 3.1). Gender is the only variable for which independence of the sample is not rejected based upon a significance value of 0.01. Tables 3.4-3.8 show the frequencies and Pearson residuals for each variable. Table 3.4 shows that there are more one person households in the budget dataset than the income dataset than one would have anticipated given the relative sample sizes. This could well be due to the high proportion of households in the budget survey headed by people over 79 years old and therefore more likely to be widowed and living alone. Table 3.8 shows that with regard to social status, it appears that there are more self employed in the budget dataset and more employed in the income dataset than one would expect from the relative sample sizes. There

is no one question in both the budget and income surveys that fully accounts for the information about socio-economic status in this table. Therefore some aggregation is necessary.

Table 3.1: Common Variables for Ireland

Common Variable Name	χ^2 / Kolmogorow-Smirnov test value (p-value)	Kruskal-Wallis test value (p-value)
Disposable Income	0.0565 (0.000)	1781.088 (0.0001)
Number of persons	45.3614 (0.000)	
Number of children	116.5923 (0.000)	
Age of household head	110.7650 (0.000)	
Gender of household head	4.7732 (0.029)	
Region	10.2574 (0.006)	
Socio-economic status of household head	91.9810 (0.000)	

Table 3.2: Mean yearly disposable income per equivalised disposable income decile for Ireland (EUR)

Disposable Income decile	Mean disposable income in budget survey	Mean disposable income in income survey
1	7,042.80	5,834.57
2	11,618.83	10,257.01
3	16,277.61	14,433.83
4	21,644.12	19,967.83
5	27,364.82	25,713.03
6	33,442.36	31,433.07
7	40,158.11	37,508.62
8	48,390.59	44,024.43
9	59,949.84	53,744.52
10	88,514.28	80,050.08

Table 3.3: Mean yearly equivalised disposable income per equivalised disposable income decile for Ireland (EUR)

Disposable Income decile	Mean disposable income in budget survey	Mean disposable income in income survey
1	5,966.29	5,140.03
2	8,064.78	6,905.35
3	9,705.50	8,577.92
4	11,899.92	10,879.39
5	14,540.54	13,084.50
6	17,267.97	15,290.16
7	20,231.14	17,771.77
8	24,086.99	20,780.87
9	29,242.88	24,904.46
10	44,525.70	38,444.44

Table 3.4: Frequencies and Pearson residuals for number of household members, IRE

np	Income dataset	Budget dataset	Total
1	548 -5.0	1513 5.0	2061
2	824 -1.4	1911 1.4	2735
3	573 1.6	1170 -1.6	1743
4	658 0.6	1417 -0.6	2075
5	495 2.1	980 -2.1	1475
6	230 2.7	408 -2.7	638
7	97 2.7	150 -2.7	247
8	28 -0.3	66 0.3	94
9	10 0.5	18 -0.5	28
10+	3 -0.8	11 0.8	14
Total	3466	7644	11,110

Table 3.5: Frequencies and Pearson residuals for region, IRE

region	Income dataset	Budget dataset	Total
Rural	1527 -1.3	3468 1.3	4995
Urban-Non Dublin	1185 3.1	2387 -3.1	3572
Dublin	754 -1.9	1789 1.9	2543
Total	3466	7644	11,110

Table 3.6: Frequencies and Pearson residuals for gender of household head, IRE

sexr	Income dataset	Budget dataset	Total
Male	2383 -2.1	5412 2.1	7795
Female	1083 2.1	2232 -2.1	3315
Total	3466	7644	11,110

Table 3.7: Frequencies and Pearson residuals for age groups, IRE

agehoh	Income dataset	Budget dataset	Total
2	190 7.3	208 -7.3	398
3	539 4.1	969 -4.1	1508
4	751 -2.0	1784 2.0	2535
5	738 0.1	1622 -0.1	2360
6	580 0.8	1231 -0.8	1811
7	469 -1.3	1103 1.3	1572
8	199 -6.7	727 6.7	926
Total	3466	7644	11,110

Table 3.8: Frequencies and Pearson residuals for professional status, IRE

socstatr	Income dataset	Budget dataset	Total
Self-employed	474 -2.5	1189 2.5	1663
Employed	1869 9.4	3387 -9.4	5256
Unemployed	82 -3.2	267 3.2	349
Other	1041 -6.8	2801 6.8	3842
Total	3466	7644	11,110

Table 3.1 shows that the assumption of equal variance of disposable income in both datasets (Kolmogorov-Smirnov statistic) cannot be rejected. Tables 3.3-3.4 show that there is some discrepancy between the surveys in terms of disposable income values. Much of this discrepancy can be attributed to the treatment of state transfers. Data on social transfers in the income dataset are taken from administrative sources whereas the budget data relies entirely on the calculations of the household interviewee. The budget data results are therefore subject to greater error because of memory loss or failure to provide complete information. Therefore income survey values are more reliable.

The QQ-plots (Figures 3.1 and 3.2) show that most incomes lie on the 45-degree line which is promising for the statistical match. There is however some discrepancy at the top of the distribution perhaps because of top coding which exists in both surveys in particular the budget survey. This may have implications for matching households in the top two per cent or so of the income distribution. Despite the greater relevance of top-coding in the budget survey, market incomes are higher in the budget than the income survey because the reference period is shorter and no account is taken of fluctuations in employment activity over the previous year.

Figure 3.1: QQ-plot for disposable income quantiles in budget dataset (trl498) versus quantiles in income dataset (zhincn), Ireland (Euros per week)

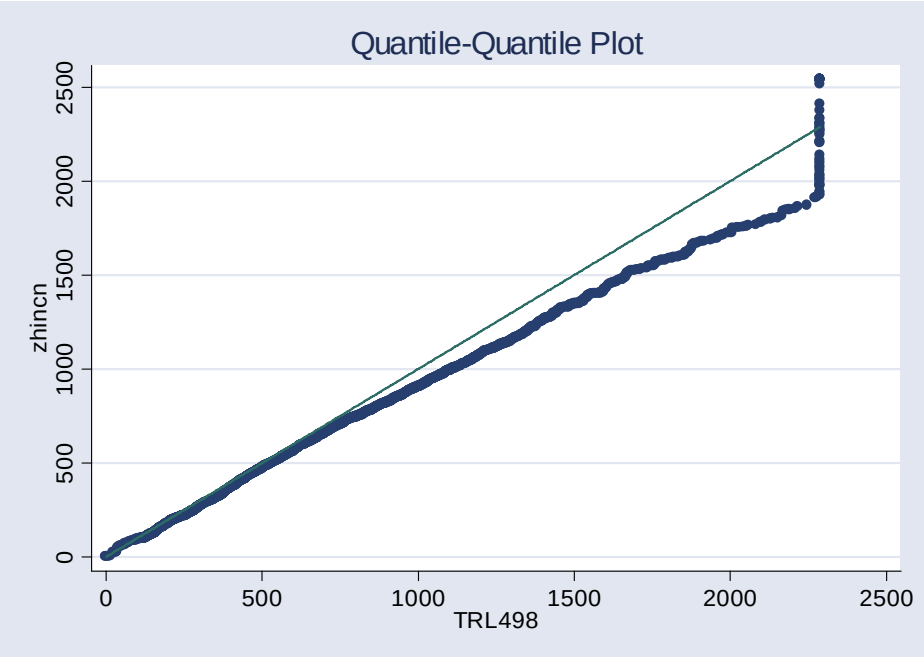
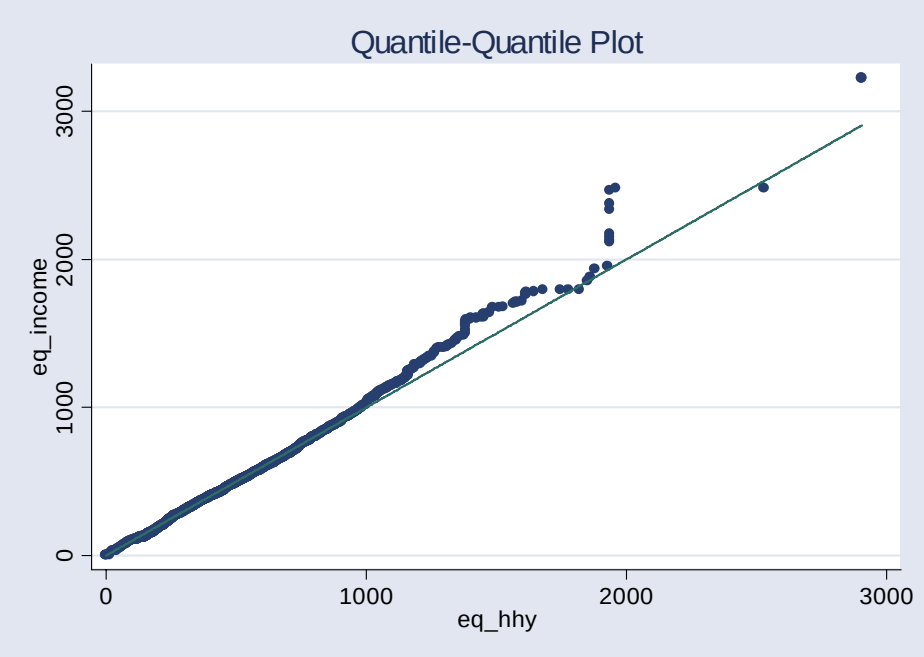


Figure 3.2: QQ-plot for equivalised disposable income quantiles in budget dataset (eq_hhy) versus quantiles in income dataset (eq_income), Ireland (Euros per week)



4. UK

For the UK, the FES 2003-04 and the FRS 2003-04 are used as budget and income datasets respectively. No deflator is necessary. Table 4.5 shows that there are no observations of the Merseyside region in the FRS. This might be a reason to exclude this category out of the budget dataset during the imputation, whenever there is reason to believe that inhabitants of this region ceteris paribus differ significantly from others in consumer behaviour. Another possible solution is to aggregate Merseyside and North West observations since Merseyside is situated in the North West.

The synthetic measures point out that, with the exception of the number of children in the family and the gender of the household head, all variables are significantly influenced by the sample. For disposable income, however, tables 4.2- 4.3 and figures 4.1 and 4.2 show again that quantile values are not so far apart. The straight line segment in figure 4.1 contains 99% of the observations (bounds chosen are 0 and 3000 GBP). Only the higher quantiles have a bending shape resulting from a relative overrepresentation of higher incomes in the FES, but equivalisation can again correct for this (figure 4.2).

Table 4.1: Common Variables for the UK

Common Variable Name	χ^2 / Kolmogorov-Smirnov test value (p-value)	Kruskal-Wallis test value (p-value)
Disposable Income	0.0438 (0.000)	10.125 (0.0015)
Number of persons	4.3e+03 (0.000)	
Number of active persons in household	80.8301 (0.000)	
Number of children	11.2063 (0.190)	
Age of household head	22.3010 (0.000)	
Gender of household head	1.1959 (0.274)	
Region	974.7608 (0.000)	
Socio-economic status of household head	67.2241 (0.000)	

Table 4.2: Mean weekly disposable income per equivalised disposable income decile for the UK (GBP)

Disposable income decile	Mean disposable income in budget survey	Mean disposable income in income survey
1	99	91
2	164	179
3	226	220
4	291	268
5	359	320
6	425	389
7	491	465
8	588	557
9	705	690
10	1.220	1.197

Table 4.3: Mean weekly equivalised disposable income per equivalised disposable income decile for the UK (GBP)

Disposable income decile	Mean disposable income in budget survey	Mean disposable income in income survey
1	70	78
2	122	150
3	159	181
4	194	213
5	232	250
6	270	291
7	313	340
8	370	401
9	454	494
10	812	859

Figure 4.1: QQ-plot for disposable income quantiles in budget dataset (dib) versus quantiles in income dataset (dii), UK (pounds per week)

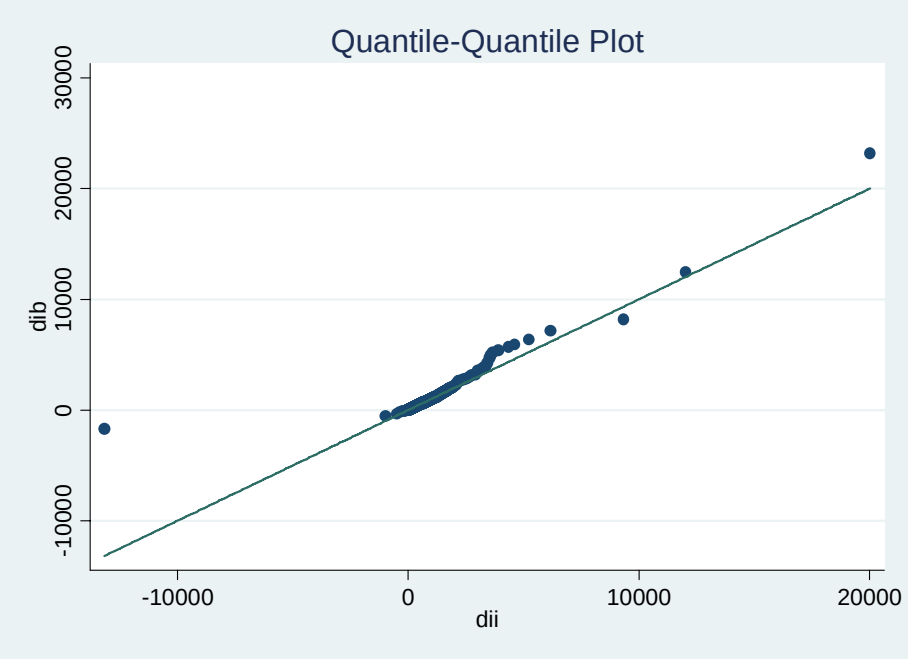
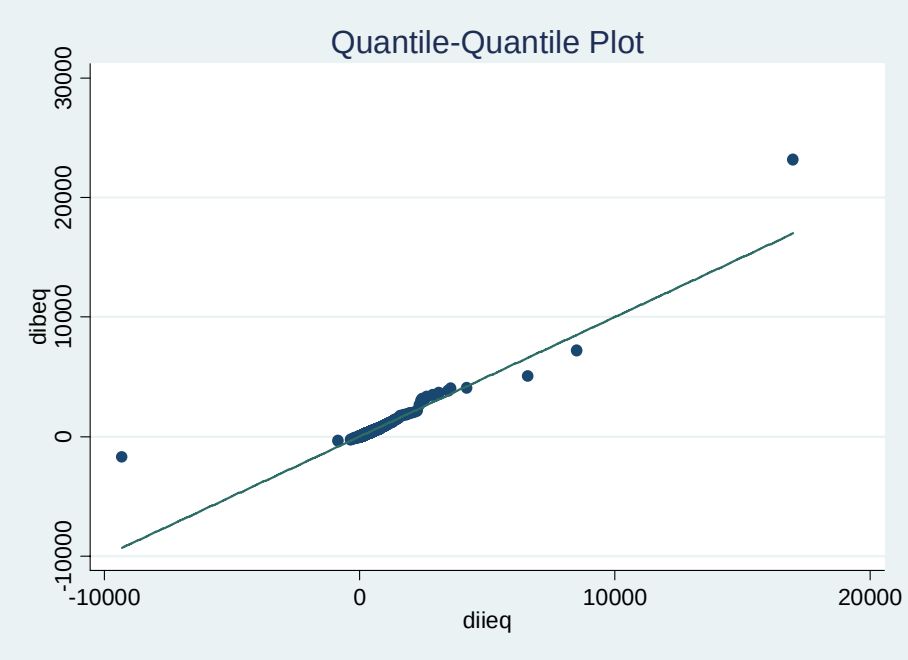


Figure 4.2: QQ-plot for equivalised disposable income quantiles in budget dataset (dibeq) versus quantiles in income dataset (diieq), UK (pounds per week)



From table 4.4, it becomes clear that differences in household sizes between the two datasets are large. The FRS contains relatively more single- person households and couples and less households of larger size than the FES. The regional differences (table 4.5) are rather modest, with the exception of the Merseyside region (no observations in the FRS), Scotland and the North West (overrepresented in the FRS) and The South West and Northern Ireland (overrepresented in the FES). The gender of the household head is perfectly balanced across the samples (table 4.6), while the FRS tends to have proportionally more elderly household heads than people in the middle category (table 4.7). For professional status (table 4.8), the FRS has relatively more household heads in the rest category and less in the employed category. Again, this can be the most important problem here since the differences between the samples may be related to a different classification of similar households.

Table 4.4: Frequencies and Pearson residuals for number of household members, UK

np	Income dataset	Budget dataset	Total
1	10,673 14.6	1,953 14.6	12,626
2	15,360 26.1	2,531 26.1	17,891
3	2,123 -19.2	1,028 19.2	3,151
4	586 -45.3	1,019 45.3	1,605
5	95 -32.6	367 32.6	462
6	16 -18.7	106 18.7	122
7	6 -10.6	35 10.6	41
8	1 -5.2	8 5.2	9
10	0 -2.0	1 2.0	1
Total	28,860	7,048	35,908

Table 4.5: frequencies and Pearson residuals for region, UK

region	Income dataset	Budget dataset	Total
north east	1,215 -0.9	313 0.9	1,528
north west	2,935 4.1	603 -4.1	3,538
merseyside	0 -24.0	140 24.0	140
yorkshire and the hum	2,201 -2.3	596 2.3	2,797
east midlands	1,893 -1.5	497 1.5	2,39
west midlands	2,201 -1.3	571 1.3	2,772
eastern	2,253 -2.1	605 2.1	2,858
london	2,682 0.6	639 -0.6	3,321
south east	3,359 -2.6	898 2.6	4,257
south west	2,129 -5.3	652 5.3	2,781
wales	1,280 -2.9	370 2.9	1,65
scotland	4,795 18.7	548 -18.7	5,343
northern ireland	1,917 -6.2	616 6.2	2,533
Total	28,860	7,048	35,908

Table 4.6: Frequencies and Pearson residuals for gender of household head, UK

sexr	Income dataset	Budget dataset	Total
Male	17,838 -1.1	4,406 1.1	22,244
Female	11,022 1.1	2,642 -1.1	2,358
Total	28,860	7,048	35,908

Table 4.7: Frequencies and Pearson residuals for agegroups, UK

agegroup	Income dataset	Budget dataset	Total
< 35	5,869 -0.8	1,462 0.8	7,331
35-55	10,963 -4.0	2,858 4.0	13,821
> 55	12,028 4.5	2,728 -4.5	14,756
Total	28,860	7,048	35,908

Table 4.8: Frequencies and Pearson residuals for professional status, UK

socstat	Income dataset	Budget dataset	Total
Other	4,324 6.4	844 -6.4	5,168
Self-employed	2,201 -0.4	548 0.4	2,749
Employed	13,685 -6.8	3,660 6.8	17,345
Unemployed	643 1.8	132 -1.8	775
Retired	7,841 2.3	1,819 -2.3	9,66
Student	166 -0.6	45 0.6	211
Total	28,860	7,048	35,908

IV. COMPARISON BETWEEN COUNTRIES

Some of the problems discussed appear common to most of the countries. The problem of different household definitions in budget and income datasets cannot be dismissed here because of high Pearson residuals in the contingency tables for household size. Because this is part of the sample setup, there is not a lot that can be done about this before the simulations. The comparison with the Greek data, where budget and target dataset are the same, can be very instructive for the interpretation of the simulation results. The same is true for the rest category of professional status. The divergence of other common variables across samples, however, may not be a very large problem, since the bias seems not to be related to diverging definitions of the variables. The only real interpretative issue here is the identification of household head, because the EUROMOD datasets do not contain an identifying variable. But the fact that the gender of the household head has in most cases very low Pearson residuals can be an indication that this is not very problematic.

As for disposable income, the guiding definition across all countries has been the income that is really available for spending. This means summing up all incomes of all different sources, subtracting direct taxes and social security contributions, and adding social

security benefits. For Belgium, this also means removing imputed rents for home owners from the calculations. The QQ-plots show clearly that incomes in budget and income datasets are comparable for the largest group of the populations, always more than 95%. There is for instance no vertical shift with respect to the 45-degrees line, which is an indication that the location of the disposable income is more or less the same in both datasets. A persistent feature for all countries, however, is a divergence across datasets for the highest disposable income quantiles. Sometimes this can be partly adjusted by equivalising for household size. But the fact remains that the imputation for higher incomes may be a lot more difficult for those higher incomes. This is partly influenced by top coding in the case of Ireland but not in many of the other countries. The ceilings for Ireland are €1803 per week in the case of the budget survey and €2000 per week in the case of the income survey.

One last issue in the cross-country comparability is the harmonisation of expenditure items. In the context of this project, the strategy is to use the COICOP aggregation system, where aggregate indirect taxes will be calculated for each category. This issue, however, is discussed in WP 3.3.

V. CONCLUSION

This workpackage contains a comparison between the populations and common variable distributions of budget and income datasets for Belgium, Hungary, Ireland and the UK. For Greece, no matching between income and expenditure data is necessary, so the Greek data can act as a benchmark during the evaluation of simulation results.

The results presented here indicate that there is an important sample bias in all countries. The distributions of age and professional status of the household head and the number of persons in a household clearly show dependence on the sample: there is a considerable difference between the budget and income surveys. In itself, this need not cause many problems for the matching process as a whole, since both regression methods and direct matching of observations with replacement[11] will match conditional upon these variables. However, for matching methods without replacement this is not the case. Two methods can be recommended in order to correct for this: 1) to build in conditionality in the latter methods, which can be done by only allowing the matching to take place within certain categories of observations, and 2) to use weight variables (from external sources) to adjust for the sample bias. The last option is clearly the most problematic one, since the

[11] This means that an observation from the budget set already matched with an observation from the income set can be used again for other income observations.

decision about which characteristics to use for the construction of weight variables boils down to relying on new assumptions and leaving the probabilistic design of the representative survey.

Special attention is given to the disposable income variable. To ensure international comparability, it was necessary to take out imputed rents from the Belgian datasets. Furthermore, the Kolmogorov-Smirnov and Kruskal-Wallis tests rejected independence across samples of the disposable income distribution and the mean rank respectively. However, it was argued that these nonparametric tests are often too strict. More intuitive measures like the QQ-plot show on the contrary that for every country the distributions are comparable across samples, and more specifically that the only problems arise at the far tails of the distribution. The part of the population not on the 45-degree line in the QQ-plot never exceeds 5%. The conclusion here is that concerning disposable income, the distribution of the majority is not different between the samples, but that for the extreme high and low (often negative) incomes, problems in matching may arise. When it comes to the disposable income concept, these findings are rather reassuring since they point to the fact that more or less the same thing is being measured by income and budget surveys. The presence of outliers can account for the divergence at the extreme value parts of the distribution. A more thorough investigation of these outliers and addressing the question whether or not they should be included in the matching process will occur while performing the concrete matching in workpackage 3.5.