Statistical matching:
a case study on EU-SILC and LFS

Aura Leulescu[1], Mihaela Agafitei[1] and Jean-Louis Mercy[1].

[1]Eurostat, European Commission, Luxembourg, Luxembourg, L-2920.

Abstract

One of the main actions foreseen by the current process of modernization of social statistics within the ESS is the streamlining of social surveys in order to enable their complementary use. In the frame of these new developments, model based techniques are explored with the aim of meeting new demands through a better exploitation of existing data sources. This paper focuses on the estimation of regional poverty indicators based on the integration of information from two social surveys: SILC and LFS. EU-SILC is the reference source for poverty indicators, but in several countries regional estimates are not of adequate precision due to the small sample size. In practice, this exercise aims to draw on the larger sample size of LFS for providing poverty estimates for areas where SILC, on its own, is not sufficient to provide a valid estimate.

Keywords: statistical matching, social surveys, regional estimates

1. INTRODUCTION

1.1. The need for better information at regional level

In the context of demographic and economic problems, policy makers put great emphasis on the development of detailed and reliable indicators on poverty and living conditions that capture regional disparities. In August 2009, the "GDP and beyond" Communication emphasised the importance of key distributional issues, including the "equitable sharing of benefits across regions". In June 2010, Europe 2020 makes explicit the linkage with the cohesion policy and highlights the strong diversity among EU regions (e.g. differences in characteristics, opportunities and needs) and the need for a strong role for regions, cities and local authorities in decision-making. The last cohesion report[1] emphasises that a key component of effectiveness for the cohesion policy is the alignment with Europe 2020, with a stronger focus on measurable results per region. Therefore, one critical need for policy makers is the provision of **reliable regional measures for poverty indicators** to be employed as benchmarks.

EU-SILC (European Union Statistics on Income and Living Conditions) provides the underlying data for the calculation of the headline indicator 'Population at risk of poverty or exclusion' and related indicators relevant to the headline target of reducing poverty of the Europe 2020 strategy. However, EU-SILC currently provides only partial information in terms of regional coverage, due to the relative small sample size in several countries. There are several countries for which direct regional estimates based on sample data are not of adequate precision due to large variances.

[1] http://ec.europa.eu/regional_policy/sources/docoffic/official/reports/cohesion5/index_en.cfm

1.2. A project for combining information from EU-SILC and LFS

The current process of modernization of social statistics within the ESS is focused on a better exploitation of existing data sources for meeting new demands. In the frame of these new developments, model based techniques (such as statistical matching and small area estimation) are explored within Eurostat in relation to specific practical needs in the field of social statistics: e.g. m*ultidimensional measures for quality of life; poverty/health estimates at regional level; joint information on income, consumption and wealth.*

Therefore, one specific stream investigated is the use of model-based methods for overcoming the problem of the small sample size for regional poverty indicators. These techniques are essentially based on statistically matching our sample with larger sample/auxiliary information in order to increase the precision of estimates.

This paper presents preliminary results on the estimation of regional poverty indicators based on the integration of EU-SILC with LFS. LFS can potentially be a good complement for this specific purpose as: it is accessible at Eurostat level and it covers all member states; it has an extensive coverage at regional level; it refers to the same population and contains a set of common variables at individual and household level. Practically, the exercise links poverty variables with covariates available in both surveys in order to impute poverty estimates for out-of-sample units (in LFS). The results illustrated in the paper refer to the integration of SILC-LFS data for only one country (Austria 2008). First results show that the integration process requires often specific solutions for different countries (different degrees of harmonization, different models, etc) and further work will need to explore the extent to which the current methodology can be applied at EU level.

The rest of the article is organized into three sections, following the main steps in the integration process. Section 2 summarises the process of coherence analysis and reconciliation between the two data sources both in terms of concepts and marginal/joint distributions. Section 3 presents the proposed methodology for building 'synthetic poverty estimates' that make use of related data from LFS. Section 4 concludes with a discussion of limitations and further methodological aspects which need to be tackled.

2. COHERENCE AND RECONCILIATION OF SOURCES

This first stage focused on assessing the existence of appropriate conditions for matching relative to the two sources involved: they should be independent samples of the same population and have the same unit of analysis; they share a common block of variables which are consistent in terms of definitions, scales, classifications, marginal and joint distributions. (D'Orazio et al, 2006)

In order to enable the integration of two or more datasets several harmonization actions needed to be undertaken so that the variables and their distributions could be made comparable. The harmonisation work required a careful consideration of both survey concepts and survey methods. Moreover, country-specific implementation aspects have to be considered. While efforts for harmonization across countries can foster a common integration approach at EU level, the exercise showed that the reconciliation of sources might require different solutions across countries.

2.1. Reference populations and units of analysis

The reference population in both surveys is the resident population living in private households. The statistical units for which information is provided are individuals and households. Same dwelling, sharing economic resources, common housekeeping and family ties are the main and mostly used criteria to identify a household. However, some methodological differences arise both between surveys and countries in terms of: (a) the application of 'economic interdependence of household members' concept, (b) the length of period of absence and (c) the treatment of specific groups (e.g. students).

For example, in both EU-SILC and LFS the recommended definition for the private household relies on the housekeeping unit concept. However, in the latter both housekeeping and dwelling concept are considered acceptable. In LFS Austria, the household concept used is the dwelling household, while for SILC AT uses the housekeeping concept. Further differences emerge for some countries in terms of the population covered and availability of household level information. Other differences emerge for persons temporary absent from the household dwelling (six months in SILC and one year in LFS for being excluded as household member) and particular groups (e.g students).

The preliminary data analysis for Austria seems to indicate that these differences do not have significant effects on comparability and we can therefore consider that the two populations overlap to a very large extent. However, this conclusion is based on data calibrated already at national level and therefore we might underestimate their impact. More in depth analysis needs to focus on specific aspects (e.g. particular categories, such as students).

2.2. Consistency definitions and scales of common variables

Both surveys provide individual and household level information. The starting point was the set of core social variables[2]. Most of them have consistent definitions with some exceptions: e.g. the activity status is optional for UK, DK; we have just the deciles for wage in LFS; for marital status and multiple citizenships there are some small differences in wording/guidelines for implementation; different typologies are applied for household composition variable(s).

In addition to the core social variables, the two sources share additional individual level labour and education variables. Data preparation and harmonization required several actions to enable the joint use and analysis as most variables need to be harmonized in terms of codification, level of aggregation, and/or format. Some variables are similar but cannot be harmonized: e.g. Years of work experience.

Both surveys provide also a great variety of additional information on the size and structure of households, number of children (dependent and non-dependent), and number of active/inactive individuals and so on. These are particularly relevant in the context of our objective as the poverty indicator is based on the household disposable income and therefore needs to be linked to household level covariates. Currently household variables

[2] http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-07-006/EN/KS-RA-07-006-EN.PDF

are often composed according to different criteria in the two sources and there are not clear standard outputs. Both enhanced harmonization and better documentation of the differences are required to foster the integration of information provided on household level. The existence of harmonized basic information allowed us to reproduce the same household variables in both surveys. These are essentially based on the combination of several socio-economic characteristics of the household members (see Table 2-1 in the annex). Thus, households are described in terms of several dimensions as follows:

- Household types in terms of size and socio-demographic characteristics of members
- Prevalence of employed/ retired/inactive persons,
- Prevalence of highly/low/medium educated,
- Prevalence of people in "high earnings" occupations/sectors.

2.3. Coherence of marginal distribution

Marginal and joint distributions were compared both for the individual and household level variables. There are three different methodologies for the analysis of distributions that were explored:

- The first and simplest one is to compute, for each potential common variable, the weighted frequency distributions for each category in the two surveys involved, and to calculate the differences. The maximum value of these differences can be taken as a criterion for comparison. Coherence of the variable in the two surveys will be rejected if this maximum difference is higher than 5 percentage points. Obviously, this is simply a rule of thumb without much theoretical background, and the threshold established is, on the other hand, arbitrary.
- Another possibility is to quantify similarity of two distributions so that we could give a relative measure of differences in the distributions of various common variables at different levels (national and regional level). We apply the Hellinger distance (HD) which lies between 0 and 1. A value of 0 indicates a perfect similarity between two probabilistic distributions, whereas a value of 1 indicates a total discrepancy. The Hellinger distance between a variable V in donor data source and the corresponding variable V' in the recipient data source is:

$$HD(V,V') = \sqrt{\frac{1}{2}\cdot\sum_{i=1}^{K}\left(\sqrt{p(V=i)}-\sqrt{p(V'=i)}\right)^2} = \sqrt{\frac{1}{2}\cdot\sum_{i=1}^{K}\left(\sqrt{\frac{n_{Di}}{N_D}}-\sqrt{\frac{n_{Ri}}{N_R}}\right)^2}$$

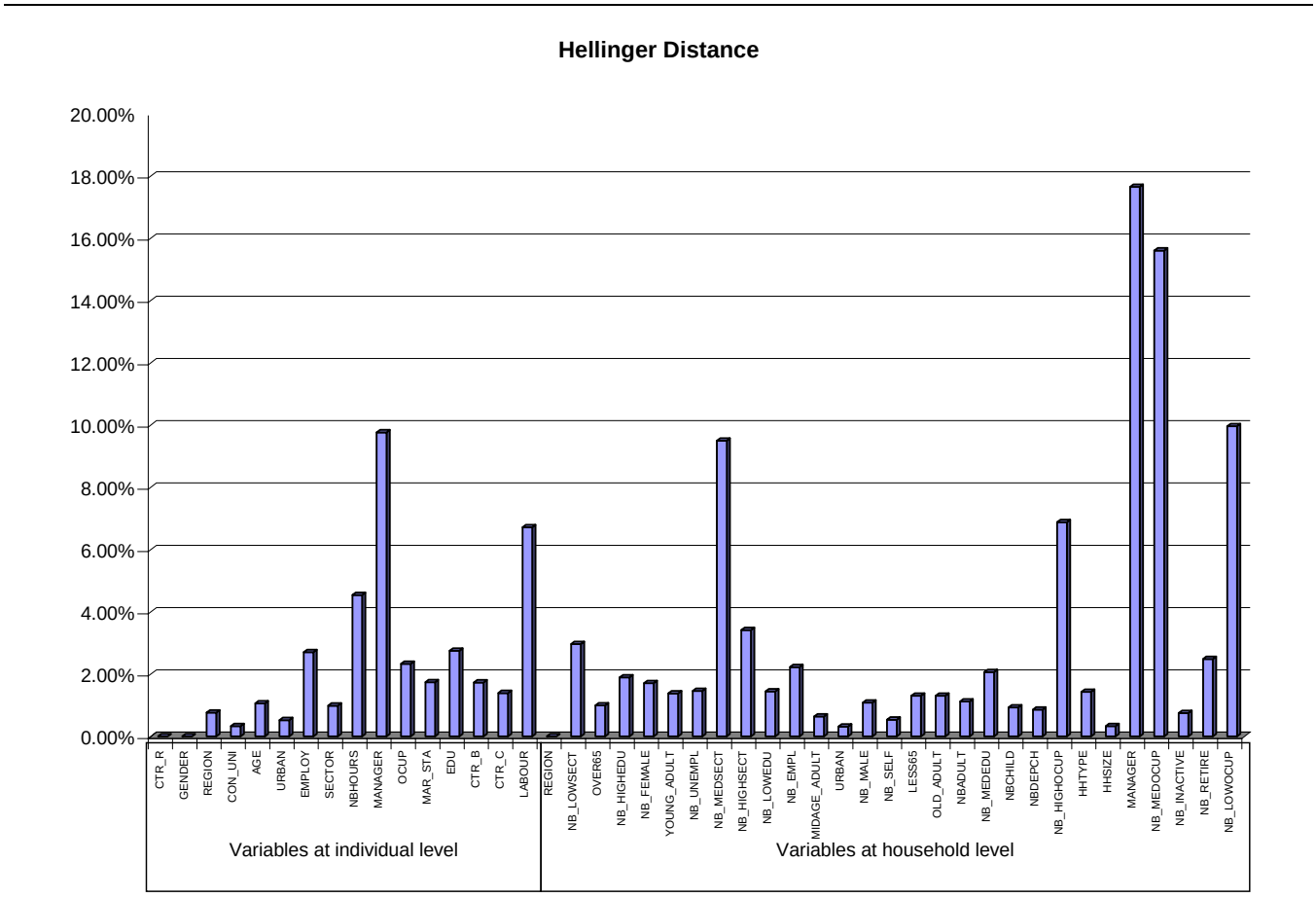
  where K is the total number of cells in the contingency table, n_{Di} is the frequency of cell i in the donor data D, n_{Ri} is the frequency of cell i in the recipient data R and N is the total size of the specific contingency table.

- The third group refers to statistical tests for the similarity of distributions (chi square; Kolmogorov Smirnov, Wald-Wolfowitz tests). These methods could give a stronger base to the conclusions on similarity/discrepancy between distributions coming from the two sources. However, both surveys have a complex design and this category of tests generally requires information on sampling design variables. LFS doesn't

provide this information at Eurostat level. Further work can investigate their application given the available data.

Our analysis was based on the first two methods for giving us a combined view on coherence of distributions. The Hellinger distance metric allows us to provide an easier to read comparative situation of the discrepancies in the data. Figure 2-1 provides an overview for both individual and household level variables. Inconsistencies at individual level translate in difficulties for the related household variables: number retired and number inactive. However, often inconsistencies come from "small cells". By aggregating these categories, the similarity of distributions improves. For instance, if we look at "self declared activity status" (LABOUR), by aggregating domestic tasks and other inactive into a single group, the Hellinger distance decreases from 6.72% to 1.41%. There are also some discrepancies for number of adults in the household working in low/medium earning occupations. In the annexes (table 5-2 in the annex), there are more detailed statistics on the coherence of marginal distributions at regional level.

Figure 2-1



These results and relevant inconsistencies need to take into account the weighting procedures that are applied at national level, weighting factors and benchmark files used for calibration, which are often different between sources.

In conclusion, ensuring coherence in terms of statistical output (marginal and joint distributions) needs both in-depth analysis and documentation of concepts and survey methods, as well as further methodological developments. Inconsistencies can emerge due to different concepts, due to operational differences, but also due to different survey

methods to treat missing information, weighting etc. Better coherence is essential for the complementary use of different data sources and it requires systematic checking of main distributions at MS and/or Eurostat level.

3. A METHOD FOR ESTIMATING REGIONAL POVERTY MEASURES

The validity of the exercise depends to a great extent on the selection of the model and the power of common variables to behave as good predictors. Our main target variable is the at-risk-of-poverty which is a binary index based on the relative position of the ***individual*** in the distribution of the income. Those bellow 60% of the median are considered income poor.

Several studies in the field of small area estimation for poverty, take income as the target variable and on the basis of income estimates they recompute other poverty measures, such as the at-risk-of-poverty rate (Molina and Rao, 2010), we decided to focus on modelling the income variable. However, a further issue emerges in relation to the level of analysis. Even if the at-risk-of-poverty index relates to the ranking in the distribution of individuals, its computations is done by assuming perfect intra-housing sharing of resources. The household disposable income, equivalised by the household size, is imputed to all individuals in the household no matter their actual contribution to the total resources of the household. Therefore, in the inference process we decided to focus our estimations on the household income. This is the income a household receives from wages and salaries, self-employment, benefits, pensions, plus any other sources of income. The household income is not normally distributed but positively skewed. Therefore, we use in the model the logarithm of the income so that this skewness is reduced and it can be assumed for the analysis that the transformed variable follows a normal distribution.

The whole estimation process is done at household level. The proposed method for providing model based regional estimates followed four main steps:

- fit a model at household level for the logarithm of household income based on EU-SILC
- multiply impute (L times) on the basis of the model "real donors" in LFS
- re-compute at-risk-of-poverty in LFS for each of the generated L vectors,
- estimate model-based regional at-risk-of-poverty rate (mean based on L imputations) and assess quality

3.1. Model specification

The analysis and techniques carried out aim at identifying the subset of common variables that best explain household disposable income. As several socio-economic factors contributing to poverty levels are at individual level we needed to translate individual characteristics in household typologies. As the reference person is defined differently in LFS and SILC and we cannot identify the "main income earner", we decided not to use the head of household characteristics. We used as predictors mainly the number/prevalence in the household of certain individual characteristics that determine the socio economic status of the household. For example, based on SILC we classified economic and occupation sectors in low, medium and high earning. We

therefore used as explanatory variables the percentage of household adults working in each of these categories.

In the first step, we correlated several variables with household income: the strongest positive correlations are for number of active people, number of highly educated people in the household, while the negative ones are for the number of unemployed, living alone, and single with children. Then, we have regressed the log of income on a subset of socio –demographic characteristics of the household. A stepwise regression was carried out in order to select the variables that best explain the household income. We used both alternatives with number and prevalence of adults with certain characteristics (e.g. employed, highly educated) in the household and they seem to give similar results.

The model seems to have a reasonable explanatory power. However, the shortcomings of the unit level area models are related to the non inclusion of location effects. If we ignore the structure and use a single-level model (e.g. individual effect) our analyses may be flawed because we have ignored the context in which processes may occur**.** One assumption of the single-level multiple regression model is that the measured individual units are independent while in reality the individuals in clusters (areas) have similar characteristics. We have missed important area level effects - this problem is often referred to as the atomistic fallacy**.** For example, this may occur, when we consider income as an outcome of interest and look at this with respect to household/individual characteristics. We might find that the individual income association with the household type might depend on the regional economic development.

Table 3-1 - Models: Dependent variable =log (household income) - AT

Variable	MODEL 1	MODEL 2	MODEL 3
Intercept	9.418***	9.25840***	9.409***
Household size	0.409***	0.152***	0.194***
No dependent children	-0.267***	-0.036***	-0.069***
Over 65	0.018***	-0.019***	0.129***
% female	-0.100***		
One adult,<65-male		-0.389**	0.026**
One adult,<65-female		-0.409***	-0.006*
One adult,>65-male		-0.243***	0.090**
One adult,>65-female		-0.396***	-0.116***
2adults,<65		0.027***	0.446***
Single parents		-0.246***	-0.292***
2adults, 1 dep child		0.036***	0.434***
Other hh, dep children		0.113**	0.513***
% unemployed adults	-0.426***	-0.313***	-0.388***
% employed adults	0.246***	0.256***	0.088***
% self employed adults	0.157***	0.333***	0.0417***
% inactive adults	-0.440***	-0.407***	-0.5171***
% retired		0.063***	0.059**
% Highly educated			0.167***
% Low educated			-0.171***
% adults -high earning occupations			0.231***
% adults -low earning occupations			-0.116***
% adults -high earning NACE			0.084***
% adults -low earning NACE			-0.106***
Manager			0.150***
	R2=0.45	**R2=0.51**	**R2=0.57**

One possibility to introduce this region-dependency is the stratification of the model. This means that we divide our sample in blocks, run the model and allow imputation just within blocks. Separate imputation allows the effects of covariates to vary between regions. This alternative assumes that our sample is informative at regional level and provides enough information to model income.

Another approach that accounts for space correlation is based on the use of **hierarchical/nested models** that include **two levels covariates**. By including both level 1 and level 2 predictors in the model, we can account for both individual characteristics as well as region characteristics. These account for between area variations beyond that explained by the variation in unit covariates. These models express relationships among variables within a given level, and specify how variables at one level influence relations occurring at another level. Both random and fixed effects can be used in the same model.

3.2. Matching with LFS

We relied in our exercise on mixed methods for the multiple imputation, and specifically on the "predictive mean matching method". This enables us to incorporate the robustness of regression based methods and in the same time to mitigate the typical "regression to the mean" effect inherent in predictions. The following steps are done through the imputation:

- Regress income on covariates
- Apply estimated coefficients also in LFS
- Find the shortest distance between estimates in SILC/LFS
- Impute (L times) the real value in LFS

We applied the model both globally and by region. In the latter case we allow different effects by region. A person with the same occupation might have different income according to the specific characteristics of the region. A further step will be to include hierarchical models in the multiple imputation procedure, in line with similar exercices in the small area estimation literature (Elbers et al., 2003, Molina and Rao 2010, Pratesi et al, 2011).

3.3. Quality evaluation

Some basic quality checks were implemented in order to check if distributions of imputed/original variables are consistent. Based on the imputed income in LFS, we re-computed the equivalised income and the at-risk-of-poverty at the individual level. We checked both marginal and joint distributions of the targeted variables, based on different models, stratification options and imputations. Table 3-1 to 3-3 highlight some of these results.

Table 3-1 - Distribution household income SILC/LFS (imputed)

Survey name (data source)	N Obs	Variable	Mean	Median
LFS	3565121	Household income	34072.06	28629.51
SILC	3561882	Household income	33097.88	29103.2

Table 3-2 - Distribution equivalised income and at-risk-of-poverty for SILC/LFS (imputed) –AT

Survey name (data source)	N Obs	Variable	Mean	Median
LFS	8144008	Eqvinc	20757.56	18822
		AROP60	0.1227793	0
SILC	8234551	Eqvinc	21383.51	19010.52
		AROP60	0.1235803	0

Table 3-3 – Differences in joint distribution of AROP with common variables (EU-SILC versus LFS)

VARIABLE	HD	WITHOUT STRATA			WITH STRATA		
		Joint distribution of		**HD**	Joint distribution of		**HD**
GENDER	0.00%	AROP60	GENDER	0.04%	AROP60	GENDER	1.23%
AGE	1.06%	AROP60	AGE	1.79%	AROP60	AGE	1.98%
CTR_B	1.74%	AROP60	CTR_B	3.29%	AROP60	CTR_B	3.07%
CTR_C	1.40%	AROP60	CTR_C	2.94%	AROP60	CTR_C	2.79%
MAR_STA	1.74%	AROP60	MAR_STA	1.83%	AROP60	MAR_STA	2.40%
CON_UNI	0.33%	AROP60	CON_UNI	0.51%	AROP60	CON_UNI	1.32%
CTR_R	0.00%	AROP60	CTR_R	0.04%	AROP60	CTR_R	1.23%
URBAN	0.53%	AROP60	URBAN	1.38%	AROP60	URBAN	1.53%
LABOUR	6.72%	AROP60	LABOUR	7.28%	AROP60	LABOUR	7.30%
LABOUR2	1.41%	AROP60	LABOUR2	2.37%	AROP60	LABOUR2	2.52%
EMPLOY	2.71%	AROP60	EMPLOY	3.19%	AROP60	EMPLOY	4.13%
OCUP	2.33%	AROP60	OCUP	3.46%	AROP60	OCUP	4.16%
SECTOR	0.99%	AROP60	SECTOR	2.26%	AROP60	SECTOR	2.56%
EDU	2.76%	AROP60	EDU	4.77%	AROP60	EDU	4.43%
NBHOURS	4.54%	AROP60	NBHOURS	4.73%	AROP60	NBHOURS	5.20%
MANAGER	9.77%	AROP60	MANAGER	9.77%	AROP60	MANAGER	9.90%

Based on the estimated AROP we computed synthetic estimates at regional level. For each region, we calculate the at-risk-of-poverty rate as the mean over L=100 imputations.

$$\hat{Y}_{reg} = \frac{\sum_{l}^{L} \hat{Yreg}^{l}}{L}$$

Bellow we present some preliminary results comparing the direct and indirect regional estimates for the mean income and at-risk-of-poverty. The results show an artificial reduction of poverty differentials between regions when we apply the same model for the whole sample (figure 3-1). An important factor is certainly the lack in the model of **location effects**. In fact when we apply **strata by region**, allowing for imputation just within regions, the indirect (model based) estimates follow the same variability patterns as direct estimates (figure 3-2). Practically, stratified imputation allows having different coefficients by region in the model, and for example the effect of household type will depend on the specific region. However, for certain regions the discrepancies between SILC and LFS become more pronounced. Further work will need to include hierarchical models that account for both household and area level effects.

Figure 3-1 – Regional AROP – Austria – Imputation with NO stratification

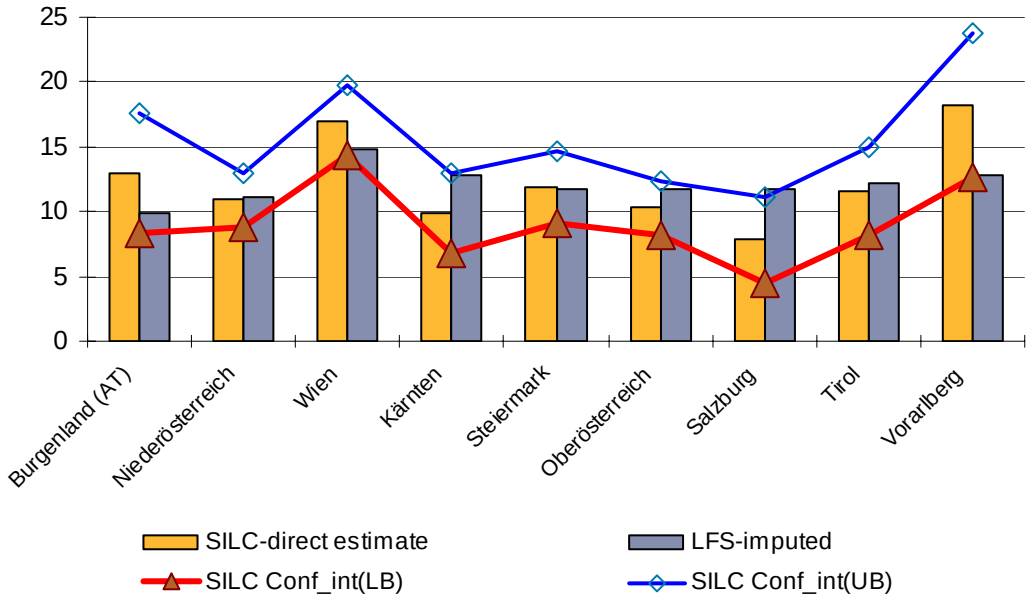
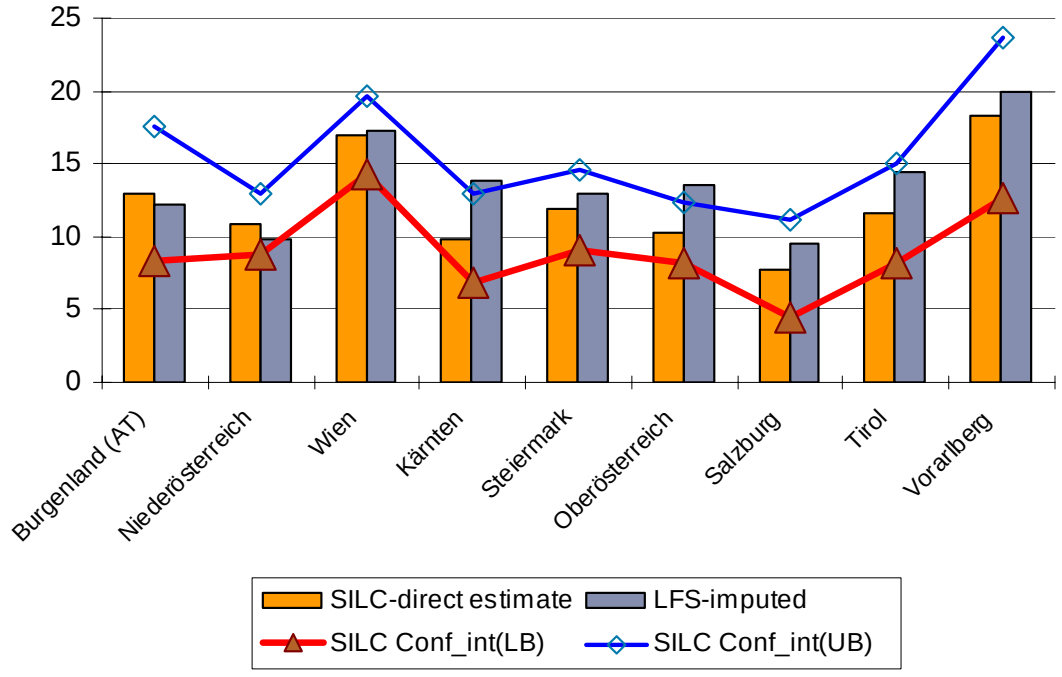


Figure 3-2 – Regional AROP – Austria – imputation WITH stratification by region

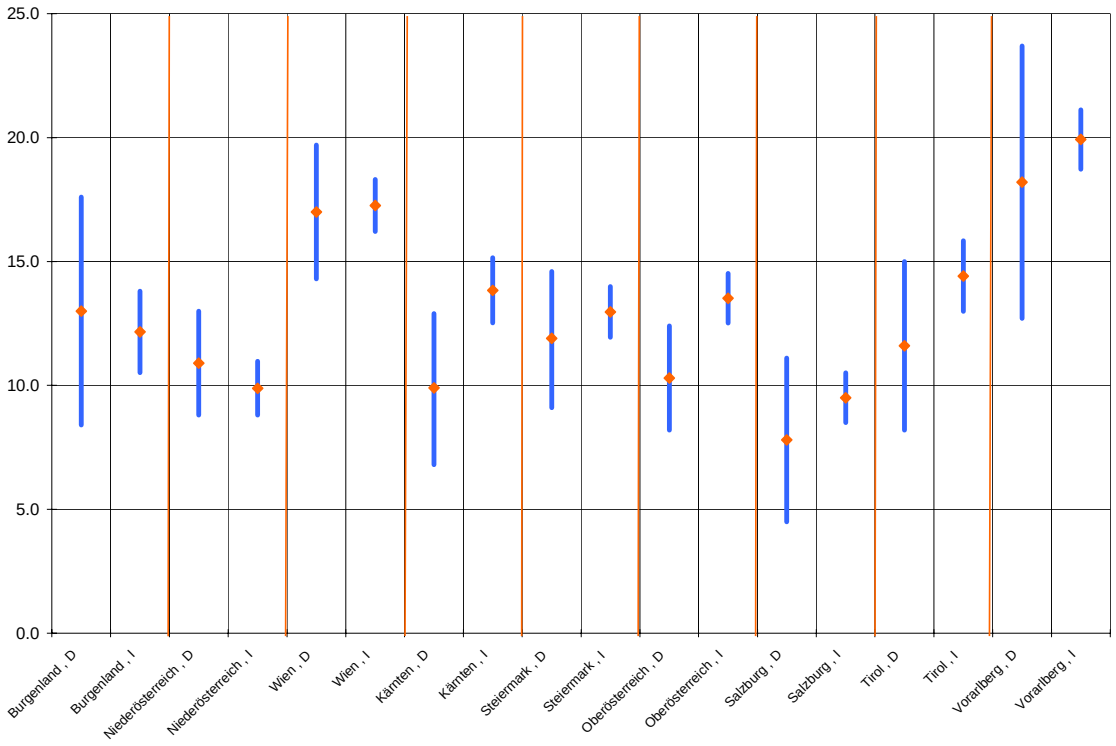


An exercise was also done for checking the value added of the model based regional estimates. We estimate the mean square error (MSE) by the average of the sum of squares of the replicates estimates around their mean:

$$MSE(\hat{Y}_d) = \frac{1}{L}\sum_{l}^{L} (\hat{Y}_d - \hat{Y}_d^{\,l})^2 , \quad \text{where } \hat{Y}_d = \frac{\sum_{l}^{L} \hat{Y}_d^{\,l}}{L}$$

The standard deviation over all the replicates is the standard error of the estimation. On the basis of these simulations we can compare the original confidence intervals for the direct estimates with the synthetic intervals computed on the basis of the estimated standard error. Even if these first results indicate an improvement in the 'precision' of estimates we need to interpret them with caution as further work needs to develop the methodology for estimating the MSE based on a larger number of replicates, using bootstrap methods. Moreover, the overlap of intervals is sometimes very small and therefore we need to further investigate the root of these inconsistencies.

Figure 3-3 – Overlap between intervals for *direct* estimates (based on SILC data) and *indirect* estimates (based on 100 imputations)



4. CONCLUSIONS AND FURTHER STEPS

The application of model based estimates for regional poverty indicators it is still research domain and there are still several open issues. This exercise explored one potential approach for improving the precision of SILC regional poverty estimates. Further work will need to focus on specification of multilevel models that incorporate location effects.

For quality assessment, further work will draw on the literature of small area estimation that uses methods on the line of bootstrap and simulation studies for estimating the MSE. This will allow comparing direct and synthetic estimates, in order to assess the potential value added of model based estimates. In some cases the two estimates are combined based on criteria such as the sample size at regional level.

5. ANNEXES

Table 5-1 – Household dimension

Household derived variable	*Description*
HHTYPE	*Household type* *'01' = 'One adult younger than 65 years - male'* *'02' = 'One adult younger than 65 years - female'* *'03' = 'One adult older or equal than 65 years - male'* *'04' = 'One adult older or equal than 65 years - female'* *'06' = '2 adults, both < 65 years'* *'07' = '2 adults, at least one 65+ years'* *'08' = 'Other no dependent children'* *'09' = 'Single parent, at least 1 dependent child'* *'10' = '2 adults, 1 dependent child'* *'11' = '2 adults, 2 dependent children'* *'12' = '2 adults, 3+ dependent children'* *'13' = 'Other households with dependent children'* *'16' = 'Other';*
HHSIZE	*Household size*
NBADULT	*Number of adults living in a household*
NBCHILD	*Number of children under 18 living in a household*
NBDEPCH	*Number of dependent children living in a household*
OVER65	*Number of adults aged 65 or less living in a household*
OVER65	*Number of adults aged over 65 living in a household*
YOUNG_ADULT	*Number of young adults (less than 35) living in a household*
MIDAGE_ADULT	*Number of mid-age adults (35-65) living in a household*
OLD_ADULT	*Number of elder adults (over 65) living in a household*
NB_MALE	*Number of male adults living in a household*
NB_FEMALE	*Number of female adults living in a household*
NB_UNEMPL	*Number of unemployed adults living in a household*
NB_EMPL	*Number of employees adults living in a household*
NB_SELF	*Number of self-employees adults living in a household*
NB_RETIRE	*Number of retired adults living in a household*
NB_INACTIVE	*Number of other inactive adults living in a household*
NB_HIGHOCUP	*Number of adults living in a household and involved in a high-paid occupation*
NB_MEDOCUP	*Number of adults involved in a medium-paid occupation*
NB_LOWOCUP	*Number of adults living in a household and involved in a low-paid occupation*
NB_HIGHSECT	*Number of adults involved in a high-paid sector*
NB_LOWSECT	*Number of adults living in a household and involved in a medium-paid sector*
NB_LOWSECT	*Number of adults living in a household and involved in a low-paid sector*
NB_HIGHEDU	*Number of high-educated adults living in a household*
NB_MEDEDU	*Number of medium-educated adults living in a household*
NB_LOWEDU	*Number of low-educated adults living in a household*
MANAGER	*Number of adults with managerial position living in a household*

Table 5-2 – Marginal distributions for each region (AT)

REGION	VARIABLE	HD
11	region*URBAN	1.14%
12	region*URBAN	0.71%
13	region*URBAN	0.00%
21	region*URBAN	3.66%
22	region*URBAN	2.11%
31	region*URBAN	1.83%
32	region*URBAN	1.97%
33	region*URBAN	2.71%
34	region*URBAN	4.97%
11	region*HHSIZE	7.87%
12	region*HHSIZE	0.95%
13	region*HHSIZE	1.49%
21	region*HHSIZE	3.88%
22	region*HHSIZE	2.46%
31	region*HHSIZE	2.32%
32	region*HHSIZE	3.70%
33	region*HHSIZE	3.67%
34	region*HHSIZE	4.01%
11	region*HHTYPE	9.54%
12	region*HHTYPE	4.67%
13	region*HHTYPE	5.43%
21	region*HHTYPE	6.01%
22	region*HHTYPE	5.09%
31	region*HHTYPE	3.62%
32	region*HHTYPE	5.75%
33	region*HHTYPE	6.87%
34	region*HHTYPE	8.57%
11	region*NBADULT	6.08%
12	region*NBADULT	3.01%
13	region*NBADULT	2.67%
21	region*NBADULT	4.74%
22	region*NBADULT	1.80%
31	region*NBADULT	2.80%
32	region*NBADULT	4.59%
33	region*NBADULT	7.60%
34	region*NBADULT	5.17%
11	region*NBCHILD	4.97%
12	region*NBCHILD	1.37%
13	region*NBCHILD	2.90%
21	region*NBCHILD	4.12%
22	region*NBCHILD	2.49%
31	region*NBCHILD	2.69%
32	region*NBCHILD	2.72%
33	region*NBCHILD	1.43%
34	region*NBCHILD	5.59%
11	region*NBDEPCH	5.22%
12	region*NBDEPCH	1.21%
13	region*NBDEPCH	2.23%
21	region*NBDEPCH	3.46%
22	region*NBDEPCH	1.72%
31	region*NBDEPCH	2.35%
32	region*NBDEPCH	2.20%
33	region*NBDEPCH	1.28%
34	region*NBDEPCH	4.36%
11	region*NB_CHILD15	2.35%
12	region*NB_CHILD15	0.93%
13	region*NB_CHILD15	2.18%
21	region*NB_CHILD15	1.11%
22	region*NB_CHILD15	2.02%
31	region*NB_CHILD15	2.38%
32	region*NB_CHILD15	2.13%
33	region*NB_CHILD15	1.65%
34	region*NB_CHILD15	3.07%
11	region*NB_UNEMPL	5.92%
12	region*NB_UNEMPL	1.45%
13	region*NB_UNEMPL	1.39%
21	region*NB_UNEMPL	2.65%
22	region*NB_UNEMPL	2.34%
31	region*NB_UNEMPL	3.05%
32	region*NB_UNEMPL	3.06%
33	region*NB_UNEMPL	3.27%
34	region*NB_UNEMPL	9.42%
11	region*NB_EMPL	8.08%
12	region*NB_EMPL	3.93%
13	region*NB_EMPL	3.87%
21	region*NB_EMPL	5.10%
22	region*NB_EMPL	6.07%
31	region*NB_EMPL	3.07%
32	region*NB_EMPL	6.70%
33	region*NB_EMPL	6.86%
34	region*NB_EMPL	3.93%
11	region*NB_SELF	2.36%
12	region*NB_SELF	1.23%
13	region*NB_SELF	2.53%
21	region*NB_SELF	6.22%
22	region*NB_SELF	0.23%
31	region*NB_SELF	1.78%
32	region*NB_SELF	1.22%
33	region*NB_SELF	2.25%
34	region*NB_SELF	0.96%
11	region*NB_RETIRE	4.02%
12	region*NB_RETIRE	3.83%
13	region*NB_RETIRE	2.56%
21	region*NB_RETIRE	3.84%
22	region*NB_RETIRE	3.35%
31	region*NB_RETIRE	3.53%
32	region*NB_RETIRE	1.09%
33	region*NB_RETIRE	6.97%
34	region*NB_RETIRE	1.83%
11	region*NB_INACTIVE	3.45%

REGION	VARIABLE	HD
12	region*NB_INACTIVE	2.01%
13	region*NB_INACTIVE	3.28%
21	region*NB_INACTIVE	1.92%
22	region*NB_INACTIVE	2.93%

REGION	VARIABLE	HD
31	region*NB_INACTIVE	2.02%
32	region*NB_INACTIVE	6.53%
33	region*NB_INACTIVE	1.81%
34	region*NB_INACTIVE	4.96%

6. REFERENCES

- ESSnet on Data integration materials: http://www.essnet-portal.eu/di/data-integrationColi, A., Tartamella, F., Sacco, G., Faiella, I., Scanu, M., D'Orazio, M., Di Zio, M., Siciliani, I., Colombini, S. and Masi, A. (2005) La costruzione di un archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine ISTAT sui consumi del
- Conti P. L., Di Zio M., Marella D., Scanu M. (2009) Uncertainty analysis in statistical matching, *First Italian Conference on Survey Methodology* (ITACOSM09), Siena 10-12 June 2009.
- Conti P.L., Marella D., Scanu M. (2008) Evaluation of matching noise for imputation techniques based on the local linear regression estimator. *Computational Statistics and Data Analysis*, **53**, 354-365.
- D'Orazio M., Di Zio M., Scanu M. (2006) *Statistical Matching, Theory and Practice*. Wiley, Chichester.
- D'Orazio, M., Di Zio, M. and Scanu, M. (2006) Statistical matching for categorical data: displaying uncertainty and using logical constraints. *Journal of Official Statistics*, **22**, 137-157
- Elbers, C., Lanjouw J.O., Lanjouw P. (2003). *Micro-Level Estimation of Poverty and Inequality*. Econometrica 71(1): 355–364
- Gilula, Z, McCulloch, R.E., Rossi, P.E. (2006). A direct approach to data fusion, *Journal of Marketing Research*, 43, 73-83.
- Kadane, J.B. (1978) Some statistical problems in merging data files. In Department of Treasury, *Compendium of Tax Research*, pp. 159–179. Washington, DC: US Government Printing Office.
- Lanjouw, P., Mathernova K., de Laat J.. *World Bank Poverty Maps to Improve Targeting and to Design Better Poverty Reduction and Social Inclusion Policies*. Presentation, 24 March 2011.
- Marella D., Scanu M., Conti P.L. (2008). On the matching noise of some nonparametric imputation procedures, *Statistics and Probability Letters*, **78**, 1593-1600.
- Molina, I., Rao, J. N. K. (2010). Small area estimation of poverty indicators. Canadian Journal of Statistics, 38: 369–385.
- Moriarity C. (2009) *Statistical Properties of Statistical Matching*, VDM Verlag
- Moriarity, C. and Scheuren, F. (2001) Statistical matching: a paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics* **17**, 407–422.
- Moriarity, C. and Scheuren, F. (2003) A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputation. *Journal of Business and Economic Statistics* **21**, 65–73.
- Office for National Statistics. *Small Area Model-Based Income Estimates, 2007/2008*. http://neighbourhood.statistics.gov.uk/dissemination/Info.do?page=analysisandguidance/analysisarticles/income-small-area-model-based-estimates-200708.htm
- Paass, G. (1986) Statistical match: evaluation of existing procedures and improvements by using additional information. In G.H. Orcutt, J. Merz and H. Quinke (eds) *Microanalytic Simulation Models to Support Social and Financial Policy*, pp. 401–422. Amsterdam: Elsevier Science.

- Pratesi, M., Marchetti, S., Giusti, C., Salvati N. (2011). *Robust Small Area Estimation for Poverty Indicators*. Department of Statistics and Mathematics Applied to Economics, University of Pisa, ITACOSM 2011, Presentation Pisa, 27-29 June 2011.
- Raessler, S. (2002) *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. New York: Springer-Verlag.
- Raessler, S., Kiesl, H. (2009). How useful are uncertainty bounds? Some recent theory with an application to Rubin's causal model. 57th Session of the International Statistical Institute, Durban (South Africa), 16-22 August 2009.
- Rodgers, W.L. (1984) An evaluation of statistical matching. *Journal of Business and Economic Statistics* **2**, 91–102.
- Rubin, D.B. (1974) Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association* **69**, 467–474.
- Rubin, D.B. (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics* **4**, 87–94.
- Ruggles, N. (1999) The development of integrated data bases for social, economic and demographic statistics. In N. Ruggles and R. Ruggles (eds) *Macro- and Microdata Analyses and Their Integration*, pp. 410–478. Cheltenham: Edward Elgar.
- Singh, A.C., Mantel, H., Kinack, M. and Rowe, G. (1990) On methods of statistical matching with and without auxiliary information. Technical Report SSMD-90-016E, Methodology Branch, Statistics Canada.
- Singh, A.C., Mantel, H., Kinack, M. and Rowe, G. (1993) Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology* **19**, 59–79.