WP. 14
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21-23 April 2008)

Topic (ii): Editing administrative data and combined source

COMBINING SURVEY AND ADMINISTRATIVE DATA IN THE ITALIAN EU-SILC EXPERIENCE: POSITIVE AND CRITICAL ASPECTS

Supporting Paper

Submitted by National Institute of Statistics, Italy[1]

I. INTRODUCTION

1. The European Survey on Income and Living Conditions (EU-SILC) is yearly carried out in different EU countries. It aims at providing data for indicators of social cohesion and social exclusion, through collecting a large set of qualitative and quantitative information at individual and household level. In addition, it provides both cross-sectional and longitudinal data, for measuring the degree of persistency for the listed indicators. For this purpose, Italy, like most EU countries adopted a rotational sample design, composed of four rotational groups, each to be followed-up during 4 years. EU-SILC is the natural successor of the European Community Household Panel (ECHP) project. Based on the experience gained from the latter and namely to overcome its quality problems, regulations for EU-SILC are characterized by a preference for output harmonization (a common framework) rather than for an input one (common survey). These regulations configure a flexible environment for National Statistical Offices (NSOs) that carry out this survey whilst settling common guidelines and procedures aiming at minimising cross-country non-comparability.

2. In this context, the Italian experience is innovative for national statistical offices (NSOs) as well as other national household surveys, for the wide and complex use of administrative registers in different phases of the survey. In this paper we discuss how the use of administrative data allows for improving data quality by (i) ensuring the correct identification and tracing of sample units in order to reduce impact of editing procedures, (ii) editing item non-response, and (iii) reducing under-reporting, memory effect and telescoping. To this purpose we consider the steps of the editing procedures involving the *population register* and *tax registers*.

3. Through *population register*, we yearly select a nationally representative probability sample of the population residing in private households within the country. Particularly, the population register provides information about the address of residence of the sampled household as well as some demographic information about household members (i.e. name, surname, sex, date of birth, municipality or state of birth, and citizenship). These information allow to correctly identify the household members, and consequently eases their follow up during the period of observation. In a longitudinal survey, households and individuals run the risk to drop out simply because they get "lost": i.e. no information about them is achieved during the fieldwork. To reduce the sample attrition due to this kind of reason, the Italian EU-SILC survey strongly relies on population register to know if "lost" household members have become out of scope after the last wave (i.e. they died, moved abroad or did not reside any longer in a private household), or have moved to a new address in the country where they can be contacted.

[1] Prepared by Claudio Ceccarelli clceccar@istat.it (corresponding author), Lucia Coppola lcoppola@istat.it, Andrea Cutillo cutillo@istat.it and Davide Di Laurea dilaurea@istat.it.

4. *Tax registers* for employees, self-employees and retirees have been used, as well as smaller groups of percipients of social and unemployment transfers and educational allowances. These data "enter" at the micro level through an exact matching. They serve the scope to: (i) reduce the impact of item non-response for quantitative amounts in survey data, and (ii) minimise phenomena like voluntary under-reporting (particularly relevant for self-employment income), memory effect and telescoping. Moreover, since different sources might use dissimilar concepts, definitions and classifications, data integration process has also to aim at avoiding income misclassification or – even worse – double counting.

5. The use of information from population and administrative registers allows for a more complete, coherent and accurate measure of households and individuals transitions in eligibility status, and moves over the national territory, as well as of income components at individual level. However, the undeniable advantages for the data quality due to combining data from different sources are not exempt from troubles. Some critical aspects mainly affect the timeliness and comparability of the released data.

6. In this paper we first discuss the use of population register for improving the quality of data referring to the tracing of sample individuals and households (Section II), then the use of tax register for improving the quality of data referring to income components (Section III), and eventually the use of both kind of registers for improving the quality of estimators (Section IV). Final remarks follow (Section V).

II. TRACING RULES AND POPULATION REGISTER

7. Longitudinal surveys are becoming increasingly attractive in socio-economic analyses for providing precious information about individual and household dynamics and their interaction over life courses. However, this kind of survey shows some drawbacks researchers have to deal with. Attrition is considered as one of the most relevant, because if produced by a non-random mechanism might lead to biased results (Lillard and Panis, 1998). Once researchers have access to survey data, they might adopt several strategies to control for the existence and the effects of possible biasing attrition (Vandecasteele and Debels, 2007). However, as data producer, the ONSs have the responsibility to adopt any feasible strategy in order to reduce attrition while carrying out the survey. Thus, for instance, EUROSTAT suggested EU countries taking part to the SILC venture, to prefer a rotational panel sample design to a pure panel one, for showing among others the advantage to lessen the impact of sample size reduction due to attrition on cross-sectional estimates (Eurostat, 2004). With the same purpose, ISTAT has developed a strategy that integrating survey data with the population register, allows for reducing sample attrition by (i) clearly identifying whether survey non-participation is due to ineligibility or non-response, and (ii) easing the follow-up of sample households and individuals moving in the national territory. We particularly focus on these two aspects for their relevance in assessing panel data quality.

8. The distinction between survey non-participation due to ineligibility or unit non-response is relevant for inference, because changes in eligibility reproduce the dynamics of the target population (deaths, moves abroad..), while changes in the response status might create problems of self-selection (Nicoletti and Peracchi, 2005). Thus, when the interviewer does not succeed in contacting the household/individual, becomes crucial knowing the reason for non-contact. As we will discuss in the following, population register is useful to integrate survey information about the reason why a household or individual is non-contacted.

9. There is evidence that response rates strongly depend on whether households move during the period of observation (Berh et al., 2005). Indeed, a household or individual moves implies that finding information about the new address might be problematic, and in case of long distance moves, the interviewer is more likely to change and consequently the household is less likely to co-operate (Nicoletti and Peracchi, 2005). As we will discuss in the following, population register is useful to integrate survey information about the new address of the household/individual to interview.

10. In order to establish the mechanism leading to unit ineligibility or follow-up, we need to define the target population as well as the tracing rules of sample units. According to the Commission Regulation (European Commission, 2003a; 2003b), the EU-SILC target population is represented by all private

households and their current members residing in the territory of the Member State at the time of data collection. Persons living in collective households and in institutions are generally excluded. The objective of the tracing rules is to reflect in the sample any changes in the target population and to follow-up individuals over time. Ideally, households and individuals should be followed wherever they move, but since tracing households and individuals is particularly time and resource consuming, some restrictions are convenient to clearly define under which conditions households and individuals have to be followed wave by wave, or have to be considered as ineligible.

11. ISTAT has adopted and adapted to the Italian survey specific characteristics the tracing rules defined by EUROSTAT (EUROSTAT, 2003; European Commission, 2003a; 2003b). Broadly, individuals surveyed at the first wave who are aged 14 and over, are defined as *sample individuals.* Any household containing at least one sample individual is defined as *sample household*. From the second wave onwards, whenever a sample individual moves to a new private household, or a sample household moves to a new address in the national territory, has to be followed and interviewed at the new address. In contrast, if a sample individual or household moves abroad, to a collective household or institution, or dies, becomes ineligible, and is not traced any longer. If an individual younger than 14 years at the first wave, moves to a new address in the following waves, is not traced or followed, because is not defined as a sample individual. Similarly, an individual who joins a sample household at the second or following waves, irrespectively from the age, is not considered as sample, and is not followed when moves. Similarly, a household that does not contain a sample individual becomes ineligible and is not interviewed any longer.

12. The EU-SILC rotational sample panel adopted by ISTAT is composed by four independent rotational groups. Every year the sample belonging to one group is renewed, and surveyed during four years. The use of population register is integrated with several stages of the survey: (i) before the fieldwork takes place, population register is used to draw the initial sample of the renewed rotational group, and to collect information about events (i.e. moves in the national territory to a private or collective household, abroad or deaths), possibly experienced by sample individuals who have to be contacted for the second or following time; (ii) during the fieldwork, population register is used for integrating survey information about sample households and individuals; (iii) eventually, after the fieldwork completion, population register is further used for imputing and correcting information about non-response reason of households and individuals surveyed, and for whom information might be incomplete or incorrect for item-non response or data-entry errors. In the following we describe these phases of the survey, for showing the relevance of the use of population register in the Italian EU-SILC.

13. Every year, municipalities joining the survey, draw from the population register a nationally representative probability sample of the population residing in private households within the country, to define the theoretical sample of the renewed rotational group,. Then municipalities provide ISTAT with information extracted from the population register not only about the address of residence of the households, but also some relevant information about household members: name, surname, sex, date of birth, municipality or state of birth, and citizenship. These data are collected in an informative system called SIGIF (*SI*stema di *G*estione delle *I*ndagini sulle *F*amiglie) that among others has the task to collect, integrate and control for the coherence among some individual and household characteristics during the longitudinal survey of the sample.

14. In this first stage, through SIGIF, an identification number (ID) is assigned to each household (HHID) and household member (PID), and is fixed during the four years of observation. During the fieldwork of the second and following waves, individuals might move from a household to another. If some household members move to a new private address, while others remain at the original address, then a so called *split* household is formed and a new HHID has to be generated by the informative system. Consequently, household members of the original and the newly formed household have to be updated according to individuals moves. Generating and managing ID codes through SIGIF, before and during the fieldwork, prevent from assigning the same ID to different individuals or households at the same wave or at different waves.

15. During the fieldwork, interviewers are provided with an auxiliary model showing the HHID and the PID associated with the Name and Surname of each household member. Household and individual

questionnaires are associated with individuals through this identification codes. This strategy allows for an accurate identification of sample households and individuals, and significantly reduces the risk of associating the information collected through questionnaires to the wrong household/individual. This aspect is particularly relevant in a longitudinal survey, because the information surveyed at four different points in time necessarily have to be coherent. Obviously, the first step to prevent incoherence and impossible transitions in one's life trajectories is guaranteeing the correct identification of individuals, and the correct generation of ID.

16. Once sample individual and household identification is assured, the challenge is represented by the tracing of individuals on the national territory, and the consequent dynamics characterising the transformation of sample household composition, the formation of new sample households and the change in the eligibility of household and individuals. Crucial is achieving information about all sample individuals who have to be contacted for the second or following time. Since acquiring these information is resource demanding, an since the fieldwork has to be carried out in a relatively short period of time, municipalities have to extract from the population register information about the events of interest possibly experienced by any sample individual, and provide ISTAT with it (events drawn from the population register are integrated in SIGIF, and ready to use to support the interviewer during the fieldwork).

17. During the fieldwork, interviewers are asked to obtain information about events experienced by sample individuals (i.e. new address of residence or death) through the other household members or neighbours. In some cases these sources allow the correct tracing of households and individuals. But in many cases, if the interviewer is not able to contact the household is also unlikely to know whether the household moved in the national territory, abroad, or died. Furthermore, the interviewer is not able to know if all the household components are not contactable for the same reason. As an instance, we might consider a household composed by three sample individuals: one moves abroad, the second dies and the third moves to a new private household. In this case the interviewer runs the risk to be unable to get information about the whole set of events, and especially on the new address of the third household member, and consequently to declare that the whole household become ineligible, although the third individual should be re-interviewed at the new address. In order to prevent this kind of misinformation that might produce a wrong classification of non-response reason, or hamper the follow-up of sample individuals, the interviewer is systematically supported during the fieldwork, by ISTAT that using population register information previously provided by municipalities, guaranteeing a better tracing of sample units. Particularly, if an individual has moved to a new address in the same municipality, the interviewer can contact him/her at the new address. If the individual has moved to another municipality, the interviewer can decide whether contacting the individual at the new address, or leave the interview to another colleague, who works closer to the new address of residence of the individual to be followed. However, in some cases the population register might not solve all situations. Sometimes, households are not contactable during the fieldwork, although they have not change their residence, for temporary reasons (e.g. illness, holidays, etc.). In these cases, households are not considered as out of scope, but a further effort to contact them is done at the following wave. We found that the wide use of population register during the fieldwork significantly improves data quality, by reducing the risk to "lose" individuals or households who have moved in the national territory, or to consider as "lost" individuals who have become ineligible because moved abroad, moved to a collective household or institution, or dead.

18. After the fieldwork completion, a further control is made between survey data and population register. Indeed, since ISTAT uses a PAPI technique to carry out EU-SILC, some errors might occur when filling in the questionnaires or during the data-entry step. Furthermore, there might be missing data about contact status, and non-response reasons. In these cases, the information drawn from the population register before and during the fieldwork, and updated in SIGIF, is used to edit survey data.

19. Summarizing, the use of information from population register allows for (i) a clear identification of sample household and individuals; (ii) an easier tracing of sample households and individuals; (iii) a more accurate classification of non-response reasons and consequently the identification of sample unit who have become ineligible.

20. However, the undeniable advantages for the data quality due to combining survey data with population register are contrasted by drawbacks worth of concerns. Particularly, drawing information form the population register before the fieldwork takes place is particularly resource consuming for the municipalities joining the survey. They have to collect information about all sample individuals although just a small proportion of them would experience the events of interest. For this reason, the collection of information about sample individuals was carried out in 2005 and 2006 but has been suspended in 2007. Consequences of this choice are currently under evaluation. A further drawback is that asking the interviewer to contact ISTAT to achieve information about sample households and individuals who cannot be contacted at their last known address, significantly increase interviewer burdens, and fieldwork timing. Nevertheless, these last aspects are considered negligible if compared with the advantages represented by the correct registration and update of household and individual dynamics.

III. MICRO-INTEGRATION OF TAX REGISTERS AND SURVEY DATA

21. It is a well established fact that household surveys aiming to collect data on income are affected by non-random total non-responses. Item non-responses for income variables are likely not to satisfy the missing-at-random hypothesis too. Survey data on income may suffer other problems than selective non-response, such as memory effect and/or telescoping. In order to reduce, or even to entirely remove, these potential biases the Italian Statistical Institute have experimented and implemented a massive recourse to tax data in the EU-SILC statistical production process.

22. The Italian tax registers have been made available for the EU-SILC project since its first year. The relevant tax forms used as sources of income micro-data are: i) the "CUD"; ii) the "UNICO persone fisiche"; iii) the "730" tax returns. They have been extensively used for "integrating" survey data on employment income and social security monetary transfers – for pensioners as well as for other percipients. These forms make also available data on capital income and rents that have been disregarded at this stage.

23. The integration among these different sources and with survey data is performed at a micro-level through the exact matching technique: all the relevant available information is combined in a integrated framework. The sequence of micro-integration requires the implementation of the following steps: i) each sample eligible person has been assigned his/her tax code as an individual identifier (matching-key); they are checked and, if necessary, corrected on the basis of auxiliary information made available from Population Registers. ii) The whole set of corrected tax codes is matched with those in the Personal Tax Annual Register; all the relevant information in the listed tax forms pertaining to the sample persons' tax codes is combined in a unique data set. iii) In turn, this data set is merged with survey records; complex statistical data editing using the fully combined data is performed, resulting in the final data set made available in the public domain. (Consolini et al., 2006).

24. Let us focus on the third step in order to illustrate the salient characteristics of the data editing process. Plenty of qualitative and quantitative information, all together with the number of routes in EU-SILC questionnaire, are high demanding in terms of cross-checks to be settled. In this context, the availability of tax data exponentially increases the degree of complexity to cope with. Relatively simpler cases are those in which one of the sources has no income information. When it is the record in the survey to lack information, tax data fill the missing item. The specular strategy applies in the opposite case, except than for records for which the interviewers report an unreliability evaluation. In the vast majority of cases, however, the two sources contain income information. Given that none of the sources is a priori considered as the most reliable, it has been implemented a complex framework of cross-checks in order to determine which income "profile" is more likely to be true. By "profile" it is to be meant the complete set of qualitative and quantitative characteristics concerning individual income: for each record his/her own profile determines each type of perceived income and the relative amount.

25. We have already noticed the potential problems coming from individual interviews when collecting data on income. However, tax data are usually far to be prompt for statistical use in household surveys at a micro-level. They are often based on concepts, definitions and classifications different from those adopted in EU-SILC. An example concerning a different classification may help to clarify the issue: employment income for cooperatives members is treated as a dependent work income following the Italian fiscal rules

whilst may be classified as a self-employment income according to EU-SILC regulations, given that their remuneration is a function of receipts or profits from the sale of the cooperatives' products or services. If so, the simple "combination" of the sources would result in a double counting: the same income is correctly reported as from self-employment in the survey and would "enter" as a dependent work one via fiscal data integration.

26. A more sophisticated case is the following: a "not so uncommon" tax-elusion practice for firms consists in remunerating their own partners' working contribution as dependent employment in order to reduce business income. In such a case the individual record referring to the partner in the tax registers would contain both dependent and independent employment income, whereas it might have been fully declared in one form (dependent employment) or, more often and more appropriately, in the other (self-employment) when interviewed. Once again, the simple combination of the two sources would over-estimate income. It is important to notice that the classification system for each of the two sources is internally consistent: in both cases income is correctly classified with respect to the original scope of the source. The problem of misclassification, which leads to partial or total double counting, arises when using tax registers in a somewhat different context from its proper one.

27. A harmonization process is essential to minimize the impact of misclassified income components. It is possible to undertake this process only in the fully combined data set. In fact, most of the auxiliary information used to detect whether an income component reported in tax data may be misclassified comes from the questionnaire: more precisely we refer to the status in employment. The Italian EU-SILC questionnaire allows for collecting more information on this aspect with respect to EUROSTAT minimum requirements. In addition to the self-defined economic status, which is the definition officially adopted by EU-SILC, the Italian version also permits to determine the status in employment according to ILO and provides other detailed employment information. While augmenting the room for potential inconsistencies – between self-defined and "objective" status in employment-, this choice proved to be crucial in exploiting tax registers at best.

28. Income definitions are also domain-specific: for the sake of simplicity it is possible to claim that a tax register collects data on taxable income whereas EU-SILC main interest is in disposable income. As argued in Di Marco (2006) the issue is less dramatic than it may appear at first glance. Survey data may be affected by underreporting, were it voluntary or not. In turn, tax evasion prevents to observe the "real" taxable income. Even disregarding it, legal tax avoidance may reduce significantly the taxable income. The general rule applied to Italian EU-SILC is the following: in presence of the same kind of employment income in both sources –administrative and survey-, and after having convincingly excluded any potential misclassification, the "final" value is set to be equal to the maximum between the two amounts. The rationale behind this decision rule rests on the consideration that it allows for minimizing the distance between the "true" value of employment disposable income and the available information. If the underreported survey income is less than the one reported in tax registers, the latter prevails and reduce the impact of sampling underreporting; in the opposite case, the underreported amount prevails, diminishing the potential negative impact of tax evasion/tax avoidance on final estimates. The validity of this decision rule crucially depends on the hypothesis that no overreporting could systematically occurs in survey data.

29. Summing up, the preference for a micro-integration between tax and survey data on income comes from the need to improve the accuracy of the data in such a way that alternative technique could not ensure. However this implies that the statistical production process had to be widely redesigned in order to allow the proper insertion of tax data resulting in a coherent final data set. The complexity of the process, in terms of aspects to be considered as well as with respect to the wideness of information to check and, if necessary, to make reciprocally consistent, does not allow for simply considering the micro-integration as a modular set of procedures to be added to a traditional production process. Much research is still to be devoted to automating editing process of the fully combined data.

30. In the Italian EU-SILC project, the integration presents also some disadvantages: (i) as any new development, it is still a "trial and error" process; (ii) It is time-expensive and, at the same time, excessively restricts the final part of the data production process, considering that tax registers are made accessible 15-18 months after the income reference period; (iii) It exposes the NSO to exogenous and unforeseeable risks:

changes in the availability of tax data or a significant variation in the relevant fiscal rules might make the process not reproducible causing a break in the time-series. Even a variation in the tax-compliance could have the same effect; (iv) The way in which it is implemented make it not possible to use tax registers to reduce the response burden.

IV. ADMINISTRATIVE DATA IN THE NON-RESPONSE CORRECTION

31. The weighting procedure for the Italian Eu-Silc accounts for three usual steps: determination of the design weight; correction for non-response; calibration to obtain final weights. The design weight for each household *j* (d_j) is given by the inverse of its inclusion probability by a stratified design, with stratification by region and demographic size of the municipality; the design weights are then transformed into a set of intermediate weights (p_j) in order to compensate for non-response; eventually, the final weights (w_j) are obtained applying a calibration (Deville and Sarndal, 1992) of the household weights to demographic data sources.

32. Using administrative data allows a better correction for non-response bias. If there are not available information on the extracted sample, the design weights are adjusted by a coefficient corrector equal to the inverse of the observed non-response rate, generally calculated at a sub-national territorial domain, such as the stratum. The drawback of this method is the underlying assumption that the behaviour of non-respondent households is the same of the respondent ones: this is a fairly strong assumption, given that behaviour in population surveys can be different between different sub-groups. If information related with the non-response behaviour are available for the extracted sample, a different coefficient corrector can be calculated. The literature contains several methods to compensate for non-response (refer, for instance, to Kalton and Kasprzyk, 1986): one of these involves adjusting weights in accordance with the inverse of the predicted probability of response obtained through a logistic regression; a second one involves constructing subgroups (weighting cells) designed so that each one comprises units having similar probability of non-response.

33. Let us first analyze the non-response rates based on the auxiliary variables. The population register provides information on the municipality size, region of residence, household size and nationality of the household head; moreover, by the use of fiscal data, we can obtain information on the household income type and amount. The difficulties to obtain the interview arises with the demographic size of the municipality and with the decreasing of the household size, because it is most often difficult to contact the household[2]. Another important determinant of the participation is the household head nationality: non-national households have a great mobility on the territory, with consequences on contacting the household; moreover, due to diffidence or difficulties with the language, they often refuse to participate. Differences are also consistent splitting the sample by the region of residence. Concerning the use of fiscal data, while we did not encounter significant differences between the sub-groups defined by the type of income declared, we encountered differences due to the level of income: the response rate arises with the declared income, showing that in the low declared income groups there are households that, due to different reasons (such as social exclusion, tax avoidance…), refuse to participate. The use of one of the above depicted methods allows to consider different response rates even considering the interactions between the auxiliary variables.

34. Using information related with non-response assures a better accuracy of the estimates, but can introduce a greater variability of the final weights in respect of the design set. By the use of a decomposition method (Dufur et al., 2001) we compared the logistic method and the segmentation method, the ones that accounts for the administrative data. Moreover, we compared these methods with the stratum method, to study if we are really introducing a greater variability of the final weights. The segmentation method is based on the CHAID (Chi-square Automatic Interaction Detection) algorithm (Kass, 1980), and divides the sample into subgroups according to the response rate of the explanatory variables by using a Chi-square test. The segmentation continues until a significant explanatory variable is found. Within each group an adjustment factor equal to the inverse of the response weight is calculated. According to the authors, the total measure of change between design and final weights, D, in the respondent group R, can be broken into four components.

[2] Refuse and non-contact are not distinguished, because both cases introduce a bias in respect of the sample design.

$$D = \left\{ \left[\sum_{i\in R} d_i \left({w_i}\Big/{d_i} - 1 \right)^2 \right] \Big/ \sum_{i\in R} d_i \right\} = R_{01} + R_{12} + R_{\mathrm{int}02} + G$$

where R_{01} measures the individual weight changes which result from going from the design to the corrected for non-response set of weights; R_{12} measures the individual weight changes which result from going from the corrected for non-response to final set of weights. R_{int} measures the interaction between the two types of change and G measures the change in average weight between the initial and the final weight. G, due to the non-response level, assumes always the same value, even concerning the different methods. Three are the main topics to observe: the amount of the distance D, that should be as little as possible; the percentage contribution of R_{01} and R_{12} to D: according to the quoted authors, R_{01} is associated with the quality of the non-response model and a larger contribution of R_{01} should be preferred, being equal D; the sign of R_{int}, that shows if the two types of change are moving in the same direction or in opposite directions: a negative sign implies that the final calibration is somehow annulling the non-response adjustment.

Table 1 – Average value of D for each component and their contribution (%) to the measure of change

Method	D	R_{01}	*R_{01}/D (%)*	R_{12}	*R_{12}/D (%)*	R_{int02}	*R_{int}/D (%)*	G	*G/D (%)*
Stratum	0,387	0,096	*24,9*	0,205	*53,0*	-0,008	*-2,1*	0,094	*24,3*
Logistic regression	0,412	0,068	*16,4*	0,268	*65,1*	-0,018	*-4,3*	0,094	*22,8*
Segmentation	0,365	0,076	*20,7*	0,195	*53,5*	0,0002	*0,1*	0,094	*25,7*

35. Comparing the methodologies that use administrative variables, the logistic and the weighting cells methods, we observe for the second one in respect of the first a lower value of D and a greater percentage of R_{01} to D (20.7% against 16.4%). Moreover, the value of R_{int} is lower and with the positive sign, indicating that the calibration step does not move against the non-response adjustment. These results leaded us to the use of the weighting cells method that appears to be a better model for non-response adjustment. Comparing the weighting cells method with the stratum method, it is possible to study the effect of the use of available auxiliary information on the entire process: we observe a lower value of D for the weighting cells but also a lower contribution of R_{01}, even if for the stratum method the percentage contribution of R_{01} is partly compensated by the negative effect of R_{int}. The results show that the use of auxiliary administrative data, besides ensuring a better accuracy of the estimates, does not imply a grater variability in the final weights.

36. Analyzing the results obtained by the CHAID algorithm, the first partition encountered is based on the demographic size of the municipality: in the largest towns it is more often probably to not find the extracted household. Successively, the different sub-groups encountered different variables explaining the non-response. The successive partition for the household sampled in the 12 biggest municipalities is based on the nationality of the household head; for the middle size municipalities the partition is based on the region of residence; for the other municipalities (under 2.000 or over 50.000 inhabitants) the partition is based on the household size. The segmentation continued until a significant explanatory variable was found. The decision tree is a particularly flexible method: when we encountered a small size cell or a high tax of non response we could “prune” the tree, till obtaining a desirable partition. In this way, we obtained 104 sub-groups, the biggest of 2.072 households and the smallest of 19 household, with a non-response rate varying from 0% to 61.9%. For every cell *z* we estimated the response rate (η_z) as the ratio between respondent households and extracted sample, both groups weighted by the design weight.

37. The results show that the most important variables explaining the non-response are the municipality size, the territorial domain, the household size and the nationality of the household head. While the first two are accounted even by the stratum method, the second two could not be considered without the use of population register linked with the extracted sample. Moreover, by the use of fiscal data, we could account for different non-response behaviours related with the type and the amount of income declared to the fiscal agency. This is a particularly important issue in Italy where there is a well established amount of tax avoidance that is particularly crowded in particular sub-groups, such as self-employed workers.

V. DISCUSSION AND FINAL REMARKS

38. The integration of survey data with information from registers is particularly important for improving data quality in a complex survey as EU-SILC. However the relevant advantages have to be discussed together with non-negligible drawbacks.

39. As argued, the use of population register for identifying sample units, and trace them over the period of observation, the use of tax registers for editing income components, and the use of both administrative sources for weighting procedures, improve different dimensions of quality. For instance, concerning income, the NSO could not rely on the exclusive use of tax registers in order to achieve EU-SILC goals, for Italy witnessing tax avoidance. However, the use administrative sources for editing survey data allows for achieving higher level of *accuracy* and *completeness*. Tax register might help not only for imputing missing items, but also for defining the respondent's income profile. Accuracy is also increased by using population register for tracing sample households and individuals moving over the national territory, for editing non-contact and/or non-response reasons, and for updating changes in the eligibility status of the sample units. Similarly, tax register and demographic register are used not only to produce calibration estimators, but also to decrease bias due to total non-response and, by this way too, improving accuracy. *Comparability* over time is improved by the use of tax registers that reflecting changes in the annual budget law, allow for properly representing trends due to legal aspects and not easily reflected by the survey. Drawing and updating sample using the population register, and integrating income information with tax registers, increase the *coherence* of the survey data with the information provided by administrative sources, both under a cross-sectional and longitudinal perspective.

40. In contrast, integration between survey and administrative data introduce some problems due to the longitudinal characteristics of EU-SILC. Developing the longitudinal editing procedures, we noticed that integrating sources we do not prevent longitudinal incoherence in the variables of interest. For instance, tax registers yearly provide coherent information for each individual. But such coherence is not guaranteed when tax register information for two consecutive years are taken into account. That is, coherence under a cross-sectional perspective does not necessarily imply longitudinal consistency. Using directly administrative data, implausible or even impossible transitions might be imputed, although original survey data were longitudinally coherent. This kind of problem would not be necessarily due to errors in the administrative data, coming instead from procedures following their collection and successive combination with survey data.

41. Concerning the use of population register, we relied on it before the fieldwork took place to collect information about events (i.e. moves in the national territory to a private or collective household, abroad or deaths), possibly experienced by sample individuals who have to be contacted for the second or following time. This practice was useful also for correcting information about individuals that could be affected by error: misspell of name, surname, or address etc.. This kind of errors would certainly not be due to errors in the register, but to mistakes in the transmission procedure that took place the previous year, for the sample drawing. As we argued, this practice has been suspended in 2007 to reduce municipalities' burden. Consequently, these kind of correction necessarily take place during the fieldwork, and strongly rely on the interviewer and his/her ability in controlling for the correctness of demographic information at household and individual level. Timeliness risks to decrease, for fieldwork timing increasing.

42. Timing for accessibility of administrative data, especially for tax data, is not under the strict control of the NSO. As a consequence it has an exogenous and deep impact on the management of the whole process, once the fieldwork is completed. Information from the tax register are available about 15-18 months after the income reference period. Current EU-SILC timetable allows for carrying out data sources integration in time with EUROSTAT deadlines. But constraint due to tax register timing does not allow for advancing data integration and editing procedures, and as a consequence hampers any chance to improve timeliness. Furthermore, any attempt to carry out the fieldwork just after the income reference period is not worth, because in any case we should wait for the tax registers accessibility in order to get the information needed to complete the data production. Among the exogenous effects, we have to consider that any non-negligible modification of fiscal rules, and consequently of the information provided by the administrative

registers, could cause a non-repeatability of the current process of data production, and a lack of temporal and cross-countries comparability.

43. The empirical evidence and the experience so far gained, suggest the relevance of preliminary analyses of the longitudinal coherence of administrative data, provided by both tax and population registers. This would allow disregarding information provided by administrative sources whenever it would induce incoherent transition, favouring the use of survey data.

REFERENCES

BEHR A., BELLGARDT E. AND RENDTEL U. (2005), Extent and determinants of panel attrition in the European Community Household Panel, *European Sociological Review*, 21(5), 489-512.

CONSOLINI P., DI MARCO M., RICCI R. AND VITALETTI V. (2006), Administrative and Survey Microdata on Self-Employment: the Italian Experience with the EU SILC project, IARIW 29th General Conference, Joensuu, 20-26 August 2006.

DEVILLE J.C., SARNDAL C.E. (1992), Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, vol. 87, pp. 367-382.

DI MARCO M (2006), Self-Employment Incomes in the Italian EU SILC: Measurement and International Comparability, Eurostat and Statistics Finland International Conference on "Comparative EU Statistics on Income and Living Conditions: Issues and Challenges", Helsinki, 6-7 November 2006.

DUFOUR J., GAGNON F., MORIN Y., RENAUD M., AND SARNDAL C.E. (2001), A Better Understanding of Weight Transformation Through a Measure of Change, *Survey Methodology*, June 2001, vol.27, N.1.

EUROPEAN COMMISSION, 2003a, COMMISSION REGULATION (EC) No 1981/2003 of 21 October 2003 implementing Regulation (EC) No 1177/2003 of the European Parliament and of the Council concerning Community statistics on income and living conditions (EU-SILC) as regards the fieldwork aspects and the imputation procedures, *Official Journal of European Union*, 46, 23-28.

EUROPEAN COMMISSION, 2003b, COMMISSION REGULATION (EC) No 1982/2003 of 21 October 2003 implementing Regulation (EC) No 1177/2003 of the European Parliament and of the Council concerning Community statistics on income and living conditions (EU-SILC) as regards the sampling and tracing rules, *Official Journal of European Union*, 46, 29-33.

EUROSTAT, (2004), Description of Target Variables: Cross-sectional and Longitudinal, Doc. EU-SILC 065/2004

KALTON G. KASPRZYK, D. (1986), The treatment of missing survey data, *Survey Methodology*, 12, 1-16.

KASS G.V. (1980), An explanatory technique for investigating large quantities of categorical data, *Applied Statistics*, 29, 119-127.

LILLARD L.A., PANIS C.W.A. (1998), Panel attrition from the Panel Study of Income Dynamics, *The Journal of Human Resources*, 33, 437-457

NETHERLANDS OFFICIAL STATISTICS (2000), Special Issue: Integrating administrative registers and households surveys, Vol. 15, Summer 2000.

NICOLETTI C., PERACCHI F. (2005), Survey response and survey characteristics: microlevel evidence from the European Community Household Panel, *Journal of the Royal Statistical Society*, 168(4), 763-781.

VANDECASTEELE, L., DEBELS, A. (2007), Attrition in Panel Data: The Effectiveness of Weighting, *European Sociological Review*, 23(1), 81-97.