SAMPLE DELIVERABLE 9

INTEGRATING DATA MODEL – FIRST RELEASE

Grant agreement No:	SSH - CT - 2007 – 217565
Project Acronym:	SAMPLE
Project Full title:	Small Area Methods for Poverty and Living Conditions Estimates
Funding Scheme:	Collaborative Project - Small or medium scale focused research project
Deliverable n.	9
Deliverable name:	Integrating data model – First release
WP no.:	3.3
Lead beneficiary:	1
Nature:	Report
Dissemination level:	PU
Due delivery date from Annex I:	31/10/2009
Due delivery date from Annex VI of Periodic Report	30/11/2009
Actual delivery date:	07/01/2010
Project co-ordinator name:	Mrs. Monica Pratesi
Title:	Associate Professor of Statistics - University of Pisa
Organization:	Department of Statistics and Mathematics Applied to Economics of the University of Pisa (UNIPI-DSMAE)
Tel:	+39-050-2216252, +39-050-2216492
Fax:	+39-050-2216375
E-mail:	coordinator@sample-project.eu
Project website address:	www.sample-project.eu

INTEGRATING DATA MODEL – FIRST RELEASE

Project Acronym:	SAMPLE
Project Full title:	Small Area Methods for Poverty and Living Conditions Estimates
Grant agreement No:	SSH - CT - 2007 – 217565
WP no:	3.3
Deliverable n.	9
Document title:	Integrating data model – First release
Status:	Final
Editors:	C.Rognini, M.Casarosa
Authors:	Monica Pratesi (m.pratesi@ec.unipi.it), UNIPI - DSMAE
	Alessandra Coli (a.coli@ec.unipi.it), UNIPI - DSMAE
	Moreno Toigo (simurg@simurgricerche.it) SR
	Claudio Rognini (c.rognini@provincia.pisa.it) PP-UROPS
	Michela Casarosa (m.casarosa@provincia.pisa.it) PP-UROPS
Document Editing:	Marta Garro (garro@sample-project.eu), UNIPI-DSMAE
Language revision:	Marta Garro (garro@sample-project.eu), UNIPI-DSMAE

1. Introduction (PP-UROPS)

The main objective of WP3 is to exploit administrative and third sector locally available data in order to calculate indicators able to monitor social exclusion and poverty and useful to define effective local social policies.

This report will collect the developments achieved by the partners in the data acquisition process; a description of the theoretical setup; and the developed model of data integration (EU-SILC, Local Data Sources).

In Deliverable 7 (D7) we showed in details the data acquisition process. In particular, the selection process of three local public agencies and one third sector organisation, with the aim of having access to their databases: *the Revenue Agency Organisation* of the Department of Finance; *The provincial Jobcentre*; *The Italian Social Security Service – INPS*; *Caritas*[1].

Since the delivery of D7, UROPS has organised specific meetings with the local responsible of the Caritas/ MIROD network. During those meetings Caritas:

- Communicated its interest in collaborating with the SAMPLE project;
- Explained how the provincial area of Pisa is parted in three different Caritas' detachments (San Miniato, Volterra and Pisa) which include also municipalities of other provinces;
- Illustrated the contents of the questionnaire used by counselling centres and explained how to have access to its data.

Afterwards, UROPS has obtained the authorisation for accessing the MIROD database (see appendix A.3.) and has started a collaboration with the three territorial Caritas, in order to organize the collaboration within the Sample project through a formal agreement.

The objectives of the formal agreement are:

- the access to the MIROD database;
- the involvement of all provincial counselling centres (there are 102 counselling centres in the whole Tuscany) in the stakeholders survey;
- the involvement of Caritas in the Observation System to monitor poverty, vulnerability and social exclusion.

In the next few days UROPS will have access to the cd with all data.

1.1. Exploration of other administrative and third sector databases

Within SAMPLE, task 3.4 aims at creating an Observation System in order to monitor poverty, vulnerability and social exclusion. This task is partially integrated with WP1 (Task 1.4 "Indicators for Local Government") concerning the involvement of local stakeholders in the selection of poverty indicators (Delphi method).

In this concern, partners of WP3 have achieved two important goals: the creation of an early list of local stakeholders (about 600 institutional and non institutional stakeholders) and the composition of a questionnaire to be sent to the stakeholders. The first section of the questionnaire include three questions about their modality of data storing, in order to explore stakeholders' information system. The first results of this survey are expexted at the end of December 2009.

[1] See D7, pp. 18-19

Meanwhile, other databases potentially useful for this task of the Sample project will be searched and selected.

2. Data description (SR)

2.1. Administrative data file: an overview

2.1.1. General perspective

Administrative data are produced as a result of or in connection with the administrative procedures of organizations. Administrative data are becoming an increasingly important data source for the production of statistics by National Statistical Institutes (NSI), since the use of administrative data drastically reduces the costs and response burden on enterprises and persons. Furthermore, they often represent the only source of data for Local Government Areas (LGAs), used for local policy planning.

Although register-based statistics are the most common form of statistics there isn’t many literature in this field. A first step towards a systematic theory and methodology on this topic is the book of Wallgren A. and Wallgren B. (2006), widely based on the Nordic countries experience. The Nordic countries have a long tradition in using administrative registers in the production of official statistics. In these countries administrative registers are becoming the main data source for the production of official statistics. This trend can also be observed in other European countries (mainly The Netherlands) and is dictated by:

- Cost reduction: direct data collection (surveys) is much more expensive than register-based statistics;
- Reducing response burden;
- Detailed information requirements: because administrative data often completely cover whole populations, it is particularly well suited for the creation of detailed information on subpopulations and at small area level;
- Longitudinal information requirements: administrative data often cover whole populations over longer periods of time.

The use of administrative data is further enhanced also by the increasing use of information and communication technology in public administrations (e-Government). As a result of this development, more and more administrative data is becoming available in an electronic form.

When public law allows the NSIs or the LGAs to use these electronic administrative repositories, they have the potential of becoming increasingly important data sources for the production of statistics and for a better planning and monitoring of national and local policies.

2.1.2. Concepts, principles and definitions

Administrative data are produced on the basis of some administrative processes, and units and variables are defined out of administrative rules and demands. The definitions may differ from the needs of the official statistics, but the data are usually of good quality for their administrative purposes. As a definition, administrative data have the following features that differ from classical statistical survey:

- in contrast to most statistical surveys, the agent that supplies the data to the statistical agency and the unit to which the data relate are usually different;
- the data were originally collected for a non-statistical purpose that might affect the treatment of the source unit;
- complete coverage of the target population;

- control of the methods by which the administrative data are collected and processed are under the administrative agency control.

Wallgren A. and Wallgren B. (2006) formulate four important principles to describe how administrative registers should be used:

1. A statistical office should have access to administrative registers kept by public authorities. This right should be supported by law as the protection of privacy.
2. These administrative registers should be transformed into statistical registers. Many sources should be used and compared during this transformation.
3. All statistical registers should be included in a coordinated register system. This system will ensure that all data can be integrated and used effectively.
4. Consistency regarding populations and variables are necessary for the coherence of estimates from different register-based surveys.

There are some main concepts and definitions we should keep in mind dealing with administrative data. According with UNECE (2007), we can define the following main concepts:

Administrative data source: Comprise in principle all kind of sources used for administrative purposes. In this report all administrative data sources mentioned are registers.

Administrative register: Register primarily used in an administrative information system. This means that the registers are used in the production of goods and services in public or private institutions or companies, or that the information is a result of such production. Administrative registers used for statistical purposes are normally operated by the state or jointly by local authorities, but registers operated by private organizations are also used.

Base register: *Administrative* base registers are kept as a basic resource for public or private administration. The function is to keep stock of the population and to maintain identification information. *Statistical* base registers are based on the corresponding administrative registers. Their principle tasks are to define important populations and contain links to other base registers.

Derived variable: New variable formed by using existing variables.

Link: One or several connecting variables that identify individual units. Links (or *keys*) are used when several registers are matched.

Primary register: Most often equivalent to administrative registers, but also used for statistical registers in areas where no central administrative register exists.

Register: Systematic collection of unit-level data organized in such a way that updating is possible. Updating is the processing of identifiable information with the purpose of establishing, updating, correcting or extending the register.

Register-based census: When all data is collected from statistical registers, we call it a (totally) register based population and housing census. A census based on combined data from registers and questionnaire is called a partially register-based census.

Register-based statistical system: Statistical registers are included in a common and coordinated system.

Register-based statistics: Statistics produced by using register data only. Data from other sources (sample surveys) may be used indirectly, for instance for imputation, calibration of models or quality assessments.

Register owner: Authority responsible for an administrative register. Also called *register keeper*.

Specialized register: Register, which unlike base registers, serves one specific purpose or a clearly defined group of purposes. Specialized registers often receive information on the population and some basic data from a base register, but supply other data themselves.

Statistical register: Register processed for statistical purposes. A statistical register could be based on one or several administrative registers. Statistical registers are also referred to as *secondary registers*.

2.1.3. The administrative data in the SAMPLE project

As stated before, the aims of WP 3 within the SAMPLE project is to explore all the possible administrative data sources collecting information related to income, poverty and social exclusion phenomena. The minimun requirements of the databases to be collected are:

- Local level: databases should be micro-data at individual level with structured geo-location or aggregated at census area level;
- Updated on a regular base;
- Structured: should be collected by the mean of structured modules and with normalized methodologies;
- Total coverage: should cover the whole provincial area and could be generalized at national and even European level.

The datasets initially selected, on the base of previous PP-UROPS experiences, were the following:

- INPS: the **Active Positions** Database **which** contains workers' data;
- INPS: the **Pensions** Database**, which** contains data of pensions according to amount, pensions supplemented by guaranteed minimum income, etc.;
- INPS: the **ISEE** Database (Indicator of the Equivalized Household Economic Position), that contains ISEE declaration's data;
- Revenue Agency: the SIATEL database, that contains data on tax returns from 2004 to 2008;
- Provincial Jobcentre: the IDOL database, that contains data related to people registered as unemployed and to the start and cessation of jobs provided by companies;
- Caritas: the MIROD database contains data about people accessing to Caritas' counselling centres.

As explained in the introduction, so far we have gained access to the following databases:

- Revenue Agency: the SIATEL database;
- Provincial Jobcentre: the IDOL database;
- Caritas: the MIROD database.

2.2. The Revenue Agency: the SIATEL database

2.2.1. Source description

The Italian Revenue Agency (IRA) is a non ministerial public body and performs the functions and obligations imposed by law in the field of taxes and fiscal duties. The IRA works on the basis of a full managerial and operational responsibility under the supervision of the Minister of Economy and Finance, who holds responsibility for policy making. The IRA has full autonomy in regard to regulations, management, assets, organization, accounting and finance. At territorial level the IRA is divided into 19 Regional Directorates and local Offices.

The IRA collects data about all taxpayers in Italy. The data are collected yearly based on taxpayers revenue declarations and stored in a wide datawarehouse system. Public Administration authorized access to these data is possible through the SIATEL system (Sistema Interscambio Anagrafe Tributaria Enti Locali).

The SIATEL database contains declarations about personal revenue of each resident.

2.2.2. Reference population and coverage

The reference population of SIATEL is every person having perceived an income in the fiscal year and consequently it is a sub-group of resident population; it includes also data about legal entities presenting fiscal declarations. Persons without an income and not having a family member perceiving an income are not included.

SIATEL contains revenue declarations presented with the following fiscal forms:

- *Modello Unico Persone fisiche*
- *Modello 730*
- Modello Unico Società di Persone
- Modello Unico Società di Capitali
- Modello Unico Enti non Commerciali
- *Modello 770 semplificato (form for natural persons not presenting declarations)*

The form in *italic* refers to natural persons, the others to legal entities.

Each person could present more fiscal declarations, so we can find duplicates in SIATEL (to be verified – gathered from the table structure provided by our SIATEL contact).

At this stage we cannot determine precisely the coverage quota of the SIATEL database with respect to the population of the province of Pisa. Referring to 2006 (the last year in which there are published fiscal data[2]) the taxpayer persons in the province of Pisa were 235.439, while the resident families in the same year were 165.429 and the total resident population was 399.881. Other local research[3] has determined that SIATEL data cover about 95% of resident eligible population.

2.2.3. Data update

Data are collected on a yearly base and are referred to the revenue of the previous year (e.g. 2008 declarations refers to 2007 revenues). SIATEL collects data from many sources, that have to be normalized and integrated one with each other. Consequently, data are published with about one year delay in respect to declarations, two years delay in respect to income.

2.2.4. Description of the variables

The SIATEL database contains basic individual socio-demographic variables (sex, age, family status and composition, job position) and many variables about the revenue composition and source. The variables that we have requested to the Revenue Agency are listed in A1.

2.2.5. Indicators from SIATEL

At this stage, we have to explore the data in order to decide the relevant and significant variables for the SAMPLE project purposes. They are all particularly relevant to calculate monetary indicators about income distribution to be compared with the EU-SILC indicators coming from the province of Pisa 2008 oversampling.

[2] See: http://www.finanze.gov.it/dipartimentopolitichefiscali/fiscalitalocale/distribuz_addirpef/lista.htm?r=1&pagina=toscana.htm&anno=2009&pr=PI

[3] See: Giovanni Bigi e Giuliano Orlandi (Ufficio statistico del Comune di Modena), Michele Lalla e Daniela Mantovani (CAPP), "L'integrazione fra banche dati locali", in Meeting on "Politiche locali e disuguaglianze. Strumenti e metodologie di conoscenza", Modena, 22 June 2006 (http://www.capp.unimo.it/WS_FEG/WS_FEGCAPP.htm)

2.3. The Provincial Job centers: the IDOL database

Provincial Jobcenters are public offices, depending from the Province of Pisa, and existing in every Italian province, having the legal responsibility of managing regular job positions concerning employees (employers and self-employed are not included)[4].

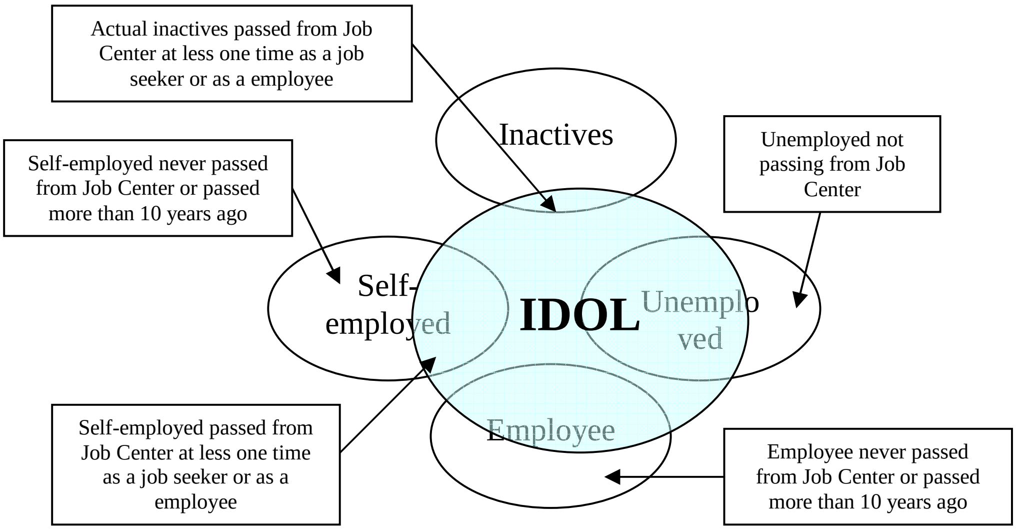
2.3.1. Source description

IDOL is a complex datawarehouse system containing stock and flow data. The flow data (derived from communications about job positions changes) update in real time the stock data (anagraphic). We are mainly interested in stock data about individuals registered in the IDOL archives.

The IDOL register (stock) contains all the data about professional conditions of people working in economic activities (public and private) localized in the province of Pisa. According to the Italian law, each change in job position has to be communicated to the Job Center. Unemployed people have to be registered to the Job Center to gain access to social provisions and to be selected for some kind of jobs. For these reasons Job Centers collect and manage data for an increasing share of active populations.

The eligible (theoretical) population is represented by people working (or seeking for a job) as employee on the labour market of the Pisa province. So it should include also non resident persons working in public and private organisations localized in the provincial territory. Self-employed individual are not included, such as retired, inactive persons, students (partially) and children. In practice, the real coverage of the eligible population is a bit different from the theoretical coverage. The IDOL individuals register contains tracks of a wider population than the eligible one, because it keeps track of each person that in a certain moment enters in (or exit from) the labour market as an employee or as a job seeker. This implies, for example, that, if an individual now working as a self-employee was registered at the Job Center one time in his life, he/she remains registered in the IDOL database, even if with a particular status flag (cancelled). The very complex situation is synthesized by the following schema:

Figure 1 The IDOL database


Actual inactives passed from Job Center at less one time as a job seeker or as a employee
Inactives
Unemployed not passing from Job Center
Self-employed never passed from Job Center or passed more than 10 years ago
Self-employed
IDOL
Unemployed
Employee
Self-employed passed from Job Center at less one time as a job seeker or as a employee
Employee never passed from Job Center or passed more than 10 years ago

[4] Co.co.pro (freelancer) are included.

2.3.2. Data update

The IDOL database is continuously updated every time the front offices of the Provincial Job Centers register a change in the job status of every employee or job seeker. Employers are obliged to communicate to the Provincial Job Center every hiring, dismissal or changing in job position of every employee within 30 days. From 2007 these communications are compulsorily transmitted online to the Job Center and they are directly registered in the IDOL database.

Therefore we can consider IDOL data updated in real time.

2.3.3. Description of the variables

The IDOL database contains basic individual socio-demographic variables (sex, age, marital status, family status and composition, ethnic origin, education level, etc.) and many variables about the job positions over the time. The variables that we have requested to the Job Centres of the Pisa Province are listed in appendix A.2.

2.3.4. Indicators from IDOL

IDOL allows to calculate indicators about local labour market comparing similar to standard labour market indicators like activity rate, unemployment rate, etc. But, for the reasons underlined in the previous paragraphs, they are biased indicators in respect to the eligible population. IDOL allows also to calculate some socio-demographic indicators (mean age, sex rate, marital status, family composition, etc.) and some monetary ones (wage amount and distribution). All these possible indicators must be tested with real data to evaluate their reliability and the kind of bias affecting them.

2.4. Caritas: the MIROD database

2.4.1. Source description

Caritas is a confederation of 162 Roman Catholic relief, development and social service organisations operating in over 200 countries and territories worldwide. Their mission is to work to build a better world, especially for the poor and oppressed. Caritas Italiana is the Pastoral Body created by the Italian Episcopal Conference in order to promote the charity commitment of the Italian ecclesiastical community, with particular attention to the poor. Caritas Italiana coordinates and performs concrete operations so as to support poor people (counselling centres, dormitories, lunchrooms, vouchers, clothing, benefits, etc….) and contrast the problem of poverty in Italy.

In order to facilitate contacts with other institutional and non institutional dealers, Caritas Toscana created the OPR “Observatories of Poverty and Resources”. The OPR are part of the MIROD Network. This Network, created in 2003, has at first designed an unique database, which contains the materials collected in all Caritas’ counselling centres.

2.4.2. Reference population

The eligible population is not “a priori” determined. Each Caritas’ counselling centre is opened to receive every person asking for help. The Local Diocesis guidelines suggest to counselling centres to follow a kind of territorial competence; this means that if an individual apply for help, but he/she doesn’t live in the area of competence of the contacted centre, the operators should send him/her to the right centre. This is not always possible or appropriate.

In principle, Caritas counselling centres receive people of three possible categories:

- Residents in the Pisa province;
- Residents outside the Pisa province;
- Homeless.

The Pisa province area is not entirely covered by the MIROD network: firstly, because the Caritas counselling centres have a fragmented diffusion on the territory; secondly, because not all counselling centres use the MIROD software.

These are the areas actually covered in the three dioceses of the Pisa province:

- Pisa’s Diocesi: 5 Caritas counselling centres and 3 soup kitchens and clothing distribution centres;
- Volterra’s Diocesi: 1 counselling centre;
- Valdarno’s Diocesi: 14 counselling centres;

In 2008 they registered 2322 accesses.

2.4.3. Data update

By now, the MIROD database is updated through periodical data transfers from each Caritas counselling center to the regional database. Data extraction comes from the regional database, so it isn't in real time. Caritas is planning to update the datawarehouse system migrating from the current offline one to a web based system, updated in real time by each counselling center.

2.4.4. Description of the variables

The MIROD database exportation gives two tables:

- The personal data table;
- The expressed needs table, linked to the personal data table;

The variables are listed in appendix A.3..

2.5. EU-SILC

2.5.1. Source description

As described in D7, “EU-SILC (Community Statistics on Income and Living Conditions) is an instrument aiming at collecting timely and comparable cross sectional and longitudinal multidimensional micro data on income poverty and social exclusion. EU-SILC is the main data source used in the SAMPLE project for estimating poverty and social exclusion indicators. For the year 2008, SAMPLE has commissioned to Istat an over-sampling for the province of Pisa. The purpose is threefold:

- getting direct estimates of poverty and social exclusion indicators for the Province of Pisa;
- improving SAE methodology through the combination of NUTS3 and NUTS4 estimates and the use of local administrative information;
- getting a larger set of units to be linked or matched with local registers.

Istat is in charge of the whole data production procedure, from the sample design to the release of microdata. Over-sampling is fully integrated in the EU-SILC standard procedure.”[5]

At present, the EU-SILC process is still in progress, microdata being expected at December 2009. For more information about the sampling process and first analysis of the response rate see SAMPLE Deliverable 7.

Istat will release to SAMPLE the microdata with the full results of 2008 interviews realized in the Pisa province. The microdata files will have the standard format described in “Cross Sectional UDB (User Data Base)” IT-SILC 2006 (Eurostat 2009). The following paragraphs are extracted from this technical guide.

[5] See D7, pp. 3-4

2.5.2. Reference population

The reference population of EU-SILC is all private households and their current members residing in the territory of the province of Pisa at the time of data collection. Persons living in collective households and in institutions are excluded from the target population.

In terms of the units involved, four types of data are gathered in EU-SILC:

- variables measured at the household level;
- information on household size and composition and basic characteristics of household members;
- income and other more complex variables termed “basic variables” (education, basic labour information and second job) measured at the personal level, but normally aggregated to construct household-level variables;
- variables collected and analysed at the person-level “the detailed variables” (health, access to health care, detailed labour information, activity history and calendar of activities).”

2.5.3. Description of the variables

The domains and areas covered by the survey are listed below and are collected at two different levels:

a) Household level:

BASIC DATA (B)	Basic household data including degree of urbanisation
INCOME (Y)	Total household income (gross and disposable)
	Gross income components at household level
SOCIAL EXCLUSION (S)	Housing and non-housing related arrears
	Non-monetary household deprivation indicators, including problems in making ends meet, extent of debt and enforced lack of basic necessities
	Physical and social environment
HOUSING (H)	Dwelling type, tenure status and housing conditions
	Amenities in dwelling
	Housing costs

b) Personal level:

BASIC DATA (B)	Basic personal data
	Demographic data
EDUCATION (E)	Education, including highest ISCED level attained
LABOUR INFORMATION (L)	Basic labour information on current activity status and on current main job, including information on last main job for unemployed
	Basic information on activity status during income reference period
	Total number of hours worked on current second/third … jobs
	Detailed labour information
	Activity history
	Calendar of activities
HEALTH (H)	Health, including health status and chronic illness or condition
	Access to health care
INCOME (Y)	Gross personal income, total and components at personal level

Following the structure of the main database, the different variables are distributed in four different files:

- Household Register (D)
- Personal Register (R)
- Household Data (H)
- Personal Data (P)

The complete list of the variables included in EUSILC dataset is in appendix A.4.

3. Objective of the data processing (DSMAE/SR)

3.1. Objectives of the analysis of the administrative data files: to provide new (even old) indicators of poverty and vulnerability at small area level (SR)

Given the main objective of WP3 within SAMPLE, the basic idea of the project is to explore local administrative datasets rich of information about these phenomena and to find out methods to calculate not-biased statistical indicators related to the Laeken indicators. The mean to correct the typical bias of administrative dataset is the linkage with the EUSILC 2008 oversampling results in Pisa province.

Before the linking of the selected administrative dataset with EUSILC oversampling results we should:

- define methods and procedures to quickly asses the quality of administrative data for statistical use in a standardized way;
- analyse the data contained in these dataset, normalize them and determine which kind of bias affects the calculated indicators. It’s a sort of preliminary explorative analysis that will produce a “biased set of indicators” (or “not corrected set of indicators”) that also could compose the information system of the local Observatory.

The analysis of the administrative data file has also the objective of defining a procedure to be replied on a yearly base to produce the indicators set.

3.1.1. Quality assessment

To assess the quality of the administrative data used in SAMPLE project we refer to Eurostat guidelines (Eurostat 2003). Eurostat has proposed to evaluate 12 aspects/dimensions for the determination of the quality of administrative data (Eurostat 2003, pp. 12-14). These aspects are:

- *Clarity*: the result of the evaluation of the metadata documentation of the administrative dataset;
- *Administrative concepts*: ability to understand the administrative concepts of the data source. The population units, variables and administrative procedures used should be described by the register holder;
- *Coverage*: the extent of the coverage of the administrative dataset. A precise definition of thepopulation units included in the dataset should be given;
- *Reference time*: the reference time of the records in the dataset. Is the time recorded the occurrence or the registration of the event or are both recorded? ;
- *Data freshness*: the time that has lapsed since the last update of the administrative dataset and the likely extent to which the data are outdated;
- *Errors in the data*: all errors that exist in the data (e.g. measurement, processing and nonresponse errors). This usually cannot be assessed directly and might imply the assistance of the register holder;
- *Completeness*: if the administrative data in the register covers all the data needs about the product.
- *Record matching ability*: ability to match the records with those in the (statistical) registers provided by a statistical source (NSI or LSI). Any existing common identifiers of population units in the data file should be listed. When this is not the case, the result of the use of other fields for record matching and an evaluation of the effectiveness should be reported;
- *Confidentiality and privacy protection*: any issues related to confidentiality or privacy protection that may impose constraints on the availability of administrative data to the NSI at a desired level of detail must be reported;

- *Compatibility between file formats*: comparison between the format in which the administrative data can be made available and the format that can be imported by the NSI. The effect of any conversion efforts should be included;
- *Comparability of administrative datasets in time*: all necessary information to assess the comparability of the data through time;
- *Envisaged use of the data*: this item must state what the potential expected use of the data is.

According with these guidelines, we assess the quality of the administrative databases to be used in SAMPLE project following three stages:

1. First step: before the dataset acquisition, we are evaluating the metadata quality using a quality checklist (see appendix) proposed by Daas P.J.H. and Fonville T.C. (2007).
2. Second step: after the dataset acquisition, we evaluate:
   - the *coverage:* to determine which units of the dataset are included or not included in the reference population and calculate the under-coverage or the over-coverage as a percentage of total population;
   - the *overall reliability*: The data in the register is explored in this stage by the means of very simple explorative data-analysis (e.g. determination of frequencies, average, medians and totals) to reveal any inconsistency in the data, missing values, mis-classifications, etc.
3. Third step: Data that is found to be correct in the first step of data evaluation needs to be studied in more detail. The following quality aspects need to be further investigated.
   - *Timeliness* (data freshness): The data in the register should describe recent events;
   - *Continuity*: The register holder should assure that the register will be maintained for a certain period in the future.
   - *Linking*: Register data has to be linked with existing data on the micro level.
   - *Validity*: the data of a selected set of variables should be compared with those of similar data already available.
   - *Expected use*: relevant indicators that should be obtained from the dataset.

3.1.2. Biased indicators calculation

When the metadata and the data of a register have been completely evaluated, it should be possible to conclude whether a register is useful or a useful addition for the creation of statistics on poverty and social exclusion.

The quality evaluation process allow us to determine which reliable (even if biased) statistical indicators can be obtained from the dataset.

This set of biased statistical indicator is a valuable output of SAMPLE project that will compose the informations system of the Local Observatory on Poverty and Social Exclusion (see WP 3.4) and will be included in the application developped in WP 4.

3.2. Objectives of the data integration: to refer the information from ad files and from EUSILC data to the same entity. (DSMAE).

In the economy of the SAMPLE project we must condition our analysis to the existing administrative information on Poverty and Deprivation. To be relevant for the SAMPLE project the information must be accessible to local policy makers at their level of governance (Local Administrative Units 2 or their aggregations as the Societies of Health (Società della Salute) or the Provinces.

In this context data integration is the process of alignment of the information from administrative files (Anagrafe Tributaria, Idol file, Caritas file) with that from Eusilc Survey. The way to adjust the information from administrative sources with that from survey data is to refer it to the same entity. Entity in this context is a virtual unit of observation common to both administrative and survey data or easily definable on both files. This entity can be the individual subject or the local level of governance to which the individual belongs. This coincides with the administrative geographical area where the individual reside and it is identified by the official European Nomenclature of Territorial Units for Statistics (NUTS).

The integration opens the possibility to apply different methods of adjustment of the Poverty and Deprivation Indicators defined on the administrative data files (Anagrafe Tributaria, Idol and Caritas files) in order to correct their self-selection bias and to compute a measure of their statistical accuracy (see Deliverable 11 for the detailed procedure)

The integration is the result of a linking process of the Eusilc data (Archive 1, from now on) with each administrative file (Archive 2) separately. The linkage procedure can be exact or statistic (probabilistic). Details on the linkage methodologies are in section 4. Whatever the linkage procedure be, the result is a matrix where the rows are the entities and the columns are the variables coming from Archive 1 or from Archive 2. This matrix is called aligned matrix.

Individual level alignment

The individual is identified by the fiscal code that is a unique identification key composed by the given name(s), surname, sex, place and province of birth (or country of birth if foreign) and the date of birth of each individual. The Italian fiscal code, officially known as Italy's Codice Fiscale, is the tax code in Italy; similar to a Social Security Number (SSN) in the United States. The tax code in Italy is an alphanumeric code of 16 characters. The code serves to identify, unambiguously for tax purposes, individuals residing in Italy irrespective of residency status.

The result of the linkage procedure is called “individual level alignment”, see Table 1 for a model of the resulting table. In the table the variables from Eusilc are in italics. As it is shown in the table, there are four possibilities:

a. the individual is present in both archives
b. the individual is present only in archive 1
c. the individual is present only in archive 2.
d. the individual is not present in archive 1 nor in archive 2.

Case d. happens when the individuals are not sampled in Eusilc nor present in Archive 2.

The most interesting cases are cases b. and c. In these cases the direct exact record linkage failed. There can be several reasons why this happens: see section 4 for a detailed description. Here the focus is on the fact that in those situations, where there is no unique identifier to perform the matching, a probabilistic linkage is used. A record on the first file is linked to a record in the second file with a certain probability, and then the decision is made on whether this link is a true link or not. As a result of the linkage process the so-called “linkage weights” are produced reflecting the degree to which the pair (j,k) [unit j from Archive 1, unit k from Archive 2] is a true link. On the basis of the Archive 1 sampling weights and using the linkage weights (Lavallee and Caron 2001; Deville, Lavallee 2006; Pratesi and Salvati, 2005) it is possible to apply the Generalized Weight Share Method and measure the accuracy of the estimates defined on Archive 2 variables.

So the alignment of the data is useful to SAMPLE goals because

- when the individual is present in both archives (case a) this allows to pass on directly the data from Archive 2 to the sampled individual. This is crucial because it allows for the weighting of the administrative data by the Eusilc sampling weights. In other words the administrative variables can be treated as they were survey variable. The administrative Poverty and Deprivation indicators obtained can be referred to the same target than Eusilc Survey.

- when the direct reference to Eusilc sampled unit is not possible the probabilistic linkage can result in a linkage weights matrix that is the basis to make inference from estimates defined on Archive 2 through the GWSM methodology.

Table 1. Individual level alignment

case	Archive 1 code	Archive 2 code	*X*	*W*	Z
a	1	18	*X1*	*W1*	Z18
a	10	29	*X10*	*W10*	Z29
b		3			Z3
b		4			Z4
b		16			Z16
c	2		*X2*	*W2*	
c	23		*X23*	*W23*	

Level of governance alignment

Each Member State is divided into several level of governance. They mainly correspond to the regions identified by the official European Nomenclature of territorial units for statistics – NUTS - Statistical Regions of Europe. Local Administrative Units (LAU) are the basic components of NUTS regions. The Nomenclature of Territorial Units for Statistics (NUTS) is defined only for the Member States of the European Union. For the candidate countries awaiting accession to the EU, for the other European Economic Area (EEA) countries and for Switzerland, a coding of Statistical Regions has been defined by Eurostat in agreement with the countries concerned (see http://ec.europa.eu/eurostat/ramon/nuts/home_regions_en.html).

The level of governance of interest here coincide with the Municipality (Local Administrative Unit 2) or aggregations of Municipalities. Among the possible aggregations here the focus is on the Societies of Health (http://www.sds.zonapisana.it/sdspisa/visualizza?chi_siamo), or the Province (http://www.provincia.pisa.it/), (NUTS3, 2006).

The Societies of Health are the level of Governance at which many social services as health care are planned, offered and assigned. The Province is the level of governance immediately after the Regions. Each Region (NUTS2 level) is partitioned into several provinces.

Municipalities are partitioned by Enumeration Districts. They are not a level of governance: they are small areas defined in occasion of the Population Census and used to canvass the territory in order to interview households and individuals. Enumeration Districts are very useful to define a virtual common unit of observation. They are defined on the basis of the addresses of the households and they can be directly known from the archive (as in EuSilc case) or they can be reconstructed using a geo-codification process of the household/individual address (as in the Anagrafe Tributaria, Idol file and Caritas file).

In the economy of the SAMPLE project the alignment of the two archives is done at Enumeration District level and at Municipality level.

These two levels are relevant because:

- merging at ED level allows for a crude reference for the results of the administrative indicators as computed with the individual level alignment.

- Merging at Municipality level is important because the figures obtained with the individual alignment can be compared also with the data collected by National Security Service. In addition the direct estimates from the Archive 1 can be a statistically significant reference at Municipality level because of the increase in the sample size due to the Oversampling procedure.

Table 2. Enumeration district level alignment

Archive 1 code	Archive 2 code	ED code	Municipality code	*W*	Z
1	18	1	45	*W1*	Z1
	20	1	45		Z2
10	29	1	45	*W2*	Z3
2	3		45		
28		2	45	*W3*	
1		1	46		
1	19	1	46	*W1*	Z1
	24	10	46		Z2
10	23	10	46	*W2*	Z3
2	3	2	48		
23		2	48	*W3*	
24		2	48		

4. Methods of data integration (DSMAE)

This section investigates the possibility of merging the administrative files described in section 1 and 2 and the Eusilc oversampling dataset by linkage of individual records. The focus is on the linkage of Eusilc records versus each single administrative data source.

As a first attempt, *record linkage* is suggested in order to identify pairs of records which correspond to the same population unit. The method is described in section 4.1.

Only a small part of records will probably match through record linkage. *Statistical matching* is then suggested in order to integrate the non-matched records. Section 4.2 is devoted to the description of this technique.

Each record of the integrated dataset will contain data from both Eusilc and the administrative files. Having access to both the administrative data and Eusilc sampling weights will make it possible to estimate administrative-data-based poverty and deprivation indicators corrected for the self-selection bias (cfr. 3.2). This is the aim which motivates the present study on data sources integration methods.

4.1. Record linkage

Record linkage is a technique which compares records contained in two files *A* and *B,* in order to determine pairs of records pertaining the same population unit. Through record linkage it is possible to obtain a new file where information form *A* and *B* is available for population units represented in both files. The *A* and *B* files are supposed to contain identical units that have to be found according to an identifier (like the social security number) or a set of identifying variables (k variables) present in both files. Record linkage is also known as exact matching and computerized matching.

Figure 2 Illustration of record linkage

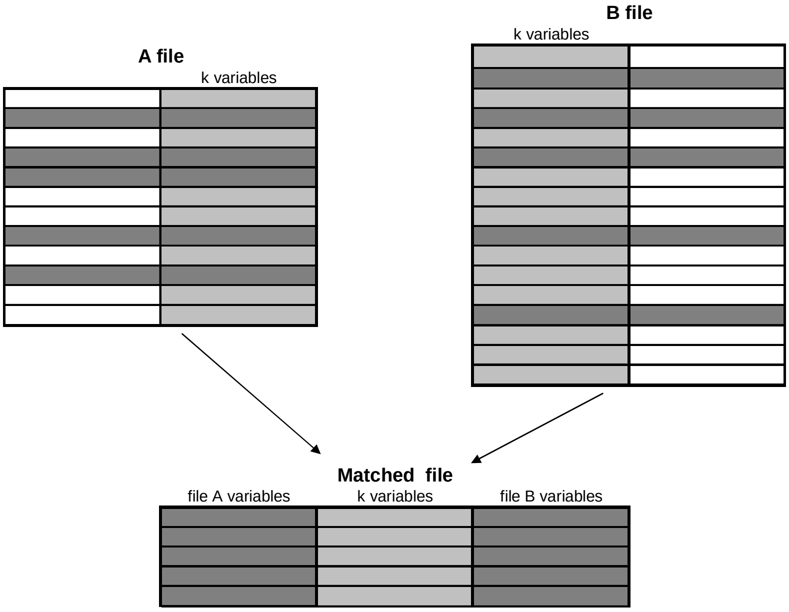


Figure 1 illustrates the principle of record linkage. Darker rows in *A* and *B* identify the same population units present in the files. Through record linkage, record pairs relating to the same population units are singled out and recorded in a new file called matched file.

In order to apply record linkage three requirements must necessarily be met (Scanu 2003, p. 17):

1. the files must have a non-empty set of units in common;
2. the files must have an identifier (for example the social security number for individuals) or a set of variables (key variables) in common which jointly allow to identify the units present in both files;
3. the multiple variable $K = (X_1, \cdots, X_k)$ derived from the k variables identifies the units univocally, in that there must be one-to-one correspondence between k-values sequences and population units.

Record linkage between two files is very simple provided that each record in both files contains the same identifier and this identifier is recorded without errors. In this case the problem is solved by simply picking out the records (if any) with the same identifier value.

Two main complications may occur (Copas and Hilton 1990):

i) Errors may occur because incorrect information is obtained from the individual, or because information is incorrectly recorded. Due to such errors two records for the same person may not agree, and two records which agree may refer to different people.
ii) Some values of the k variables may be missing so that the K-variable may not be known exactly for some of the records in *A* or *B*.

The k variables have to be chosen among statistically accurate permanent variables as the date of birth, the name or the gender. Non-permanent variables like education or the marital status may change over time so that different values could actually refer to the same population units, differences being due to different reference periods only.

Unfortunately the number of matched records will be further reduced by errors and missing values in the k variables. Conversely, some of the matched records could refer to different population units.

Just to give an idea of the relevance of the errors and missing fields problem we reproduce here the results of the analysis carried out by Copas and Hilton (1990) on a study file. The file consisted of 8601 pairs of records, each pair corresponding to the same population unit. The studied fields were: Nationality, Sex, Date of birth (two digits each for year, month, day) sound codes of names (Family name, up to three forenames).

Table 3 Agreements and disagreements in the study file, by field

Field	Both observed		One missing	Both missing	Proportion of pairs (disagree)
	agree	disagree			
Nationality	8374	227	-	-	0.026
Sex	8397	93	110	1	0.024
Birth year	8268	311	22	-	0.039
Birth month	7830	198	264	309	0.054
Birth day	7762	259	262	318	0.061
Family name	7276	1325	-	-	0.154
1st forename	6843	1732	24	2	0.204
2nd forename	2864	1123	1252	3362	0.276
3rd forename	245	201	413	7742	0.071

Copas, Hilton(1990), p. 291

Table 3 shows the number of record pairs with both values present, one missing or both missing, the pair with both values present being divided into those agreeing and those disagreeing. A double blank (both missing) is considered an agreement.

Disagreements occur more often in the Names fields. Sex records the lowest proportion of disagreements, which is mostly due to missing values. About 4% of records pair disagrees for the Birth year values, especially for recording errors. Birth month and Birth day record higher proportions of disagreements as well as a significant number of both missing fields.

The presence of this kind of errors considerably affects the record linkage quality as well as the quality of the statistical analysis based on the resulting matched file.

Record linkage methods have been studied and applied for years. Generally speaking we can group these methods into the following categories (Scanu, 2003):

i) ad hoc methods, which consider the record linkage problem mainly as a computational issue;
ii) statistical methods which formalize the linking procedure into a statistical model.

With ad hoc methods a sort of blind matching is run without any detail on the probability of making errors. On the contrary, with statistical methods record linking results can be evaluated by measuring the probability of generating false-matched-pairs and false-unmatched pairs.

The following section is devoted to the description of the statistical model developed by Fellegi and Sunter (1969).

4.1.1 The theory

Fellegi and Sunter (1969) provide a theoretical framework for a computer oriented solution of record linkage which is still nowadays considered a milestone. In the following we recall the main aspects of this theory.

Let us consider the number of pairs composed by the *A* and *B* units:

$$A\times B=\{(a,b):a\in A,b\in B\}.$$

Record linkage aims at partitioning the $A\times B$ set into the disjunctive subsets *M* and *U*, where:

$$M=\{(a,b)\in A\times B:a=b\}$$

$$U=\{(a,b)\in A\times B:a\neq b\}$$

The *M* and *U* subsets are named *matched* and *unmatched* datasets respectively. Each unit in the population is identified by the k variables recorded values.

Two distinct record generating processes, one for each of the two population, give rise to one record for each population unit. These records, denoted as α(a) and β(b), contain the k variables values observed on the *a* and *b* units respectively.

The assignment of a unit pair to the *M* or *U* subsets depends on the k variables values observed on the *a* and *b* units. A comparison is to be made in order to decide whether or not the compared units represent the same person.

A comparison vector is thus defined as a function of the records α(a) and β(b):

$$\gamma[\alpha(a),\beta(b)]=\left(\gamma^1[\alpha(a),\beta(b)],\cdots,\gamma^k[\alpha(a),\beta(b)]\right)$$

The simplest way of defining the γ for the *h*-th variable is:

$$\gamma^h=\begin{cases}1 & if\ X_a^h=X_b^h\\ 0 & otherwise\end{cases}$$

The comparison set of possible realizations of γ is denoted by Γ .

Three decisions can be made:

- The first decision, denoted by *A₁*, is called *positive link*: $(a,b)\in M$
- The second decision, denoted *A₃*, is called *positive-non-link* : $(a,b)\in U$
- The third decision, denoted *A₂*, is called *possible-link*: cases in which we find ourselves unable to make either of the previous decisions

The *A₁* and *A₃* decisions may imply errors, in that linked records could correspond, in fact, to different persons or non-linked records could in fact correspond to the same person.

The probabilities of these errors are defined as:

$$\mu=P(A_1\mid U)=\sum_{\gamma\in\Gamma}u(\gamma)P(A_1|\gamma)$$

$$\lambda=P(A_3\mid M)=\sum_{\gamma\in\Gamma}m(\gamma)P(A_3|\gamma)$$

where:

$u(\gamma)=P\{\gamma[\alpha(a),\beta(b)]\ |(a,b)\in U\}$ and $m(\gamma)=P\{\gamma[\alpha(a),\beta(b)]\ |(a,b)\in M\}$

$u(\gamma)$ and $m(\gamma)$ representing the conditional probability of γ , given that $(a,b)\in U$ or $(a,b)\in M$.

Fellegi and Sunter (1969) procedure defines the decisional rule for labeling each pair of records as *positive link* (decision *A₁*), *non-positive link* (decision *A₂*) or *possible link* (decision *A₃*).

As a first step, the comparison vector $\gamma[\alpha(a),\beta(b)]$ is transformed into a real number (called weight) as follows:

$$t(\gamma)=\frac{m(\gamma)}{u(\gamma)}$$

Higher $t(\gamma)$ values are more probably generated by matched pairs.

As a second step, two threshold values T_μ and T_λ are calculated in order to identify the $t(\gamma)$ intervals corresponding to each decision. The method allows to calculate such thresholds values given the required μ and λ probabilities of errors.

For each $t(\gamma)$ value:

- If $T_\mu \leq t(\gamma)$ A_1 decision is taken: positive link
- If $t(\gamma) \leq T_\lambda$ A_3 decision is taken: positive non-link
- If $T_\lambda \leq t(\gamma) \leq T_\mu$ A_2 decision is taken: possible link

The authors suggest an *optimal* linkage rule $L(\mu, \gamma, \Gamma)$ which assigns probabilities $P(A_1 | \gamma), P(A_2 | \gamma), P(A_3 | \gamma)$ to each possible realization of Γ in order to minimizes the probability of failing to make a positive link ($P(A_2 | \gamma)$), given fixed levels of the μ and λ probabilities.

From a practical perspective, the implementation of the described theory requires the following steps:

- Identification of the k variables
- Computation of the comparison vector values for each pair of records (a,b): $\gamma_{a,b} = (\gamma_{a,b}{}^1, \cdots, \gamma_{a,b}{}^k)$;
- Estimation of the $m(\gamma)$ and $u(\gamma)$ probabilities for each distinct realization of vector γ ;
- Calculation of the weight value $t(\gamma)$ for each pair of records;
- Calculation of the threshold values T_μ and T_λ

Fellegi and Sunter (1969) outline a method for calculating the threshold values corresponding to the required levels of errors μ and λ . Moreover they propose two different methods for calculating the quantities $m(\gamma)$ and $u(\gamma)$. The procedure allows to select the set of matched record pairs (*M* set) providing a measure of the error probabilities μ and λ . This is crucial since it is possible to evaluate the quality of the matched file.

4.1.2. Record linkage of Eusilc and administrative files

Let us indicate F_E as the Eusilc file, and F_S, F_J and F_C as the SIATEL, Idol and Caritas files respectively.

Given the population covered by each data source (see section 2) we can assume that a number of identical individuals be present in each of the compared files. This authorizes a record linkage procedure in order to build a file where the F_E. records are extended with data from the linked administrative source.

At its simplest the problem is stated as follows. F_E contains N records, one for each of N individuals. F_S (F_J, F_C) contains M records with data on individuals who may or may not be among those represented in the F_E file. Given a common set of variables, we have to evaluate the evidence that the i-th record from F_E and the j-th record from F_S, (F_J or F_C) relate to the same person.

F_E contains the following personal items which could be used as k variables:

- X_1: Birth day
- X_2: Birth month
- X_3: Birth year
- X_4: Gender

- X_5: Place of birth
- X_6: Place of residence (Municipality or Enumeration District)
- X_7: Nationality

F_S, F_J and F_C record these variables directly or they can be derived from other information present in the files.

As a first step the k variables have to be checked in order to account for errors and missing values. For what concerns Eusilc a checking procedure has already been run by Istat as a part of the survey validation process (see Istat (2008), section 4). Particularly, personal items as individual gender and birth date have been corrected taking into account data from Municipalities record registers from which Eusilc samples are selected.

Let us define $\gamma_{e,s}$ as:

$$\gamma^h = \begin{cases} 1 & if\ X_e^h = X_s^h \\ 0 & otherwise \end{cases}$$

For k=7, the cardinality of the comparison set Γ will be $2^7 = 128$. For each realization of vectors γ, the probabilities $m(\gamma)$ and $u(\gamma)$ should be estimated, by using information directly available in the compared datasets or by using prior information on the distribution of the k variables as well on the probabilities of the different kinds of errors.

4.2. Statistical matching

Once record linkage has been performed, both Eusilc and the administrative datasets will extend the available information for each matched population unit. However, both files will continue to present the pre-linkage information for the unmatched records (Fig. 4.2).

Unmatched pairs, i.e. pairs belonging to the *U* subset may correspond to:

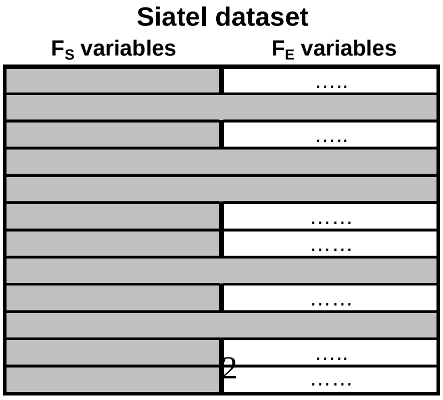
i) Units present in one of the compared files only because of differences in the eligible populations of each source; for example, people who do not earn money are not supposed to enter the Siatel file; conversely homeless people, included in the Caritas file, are not surveyed by Eusilc;

ii) Units appearing in the administrative file but not in the Eusilc file because not sampled for the survey;

iii)Units present in both data sources which did not match because of errors or missing values in the linking variables.

For the unmatched pairs belonging to the last two categories it is possible to recover a proper sampling weight from Eusilc file. On the contrary this solution cannot be applied for those units which do not belong to the Eusilc eligible population (such as homeless).

Figure 2 Siatel dataset after record linkage



We can look at the unmatched records as if they were affected by missing values for the fields corresponding to the Eusilc variables. These last can be imputed through statistical matching by singling out proper donors among the Eusilc units.

Statistical matching is a data integration procedure which is used to integrate two or more datasets provided that: i) the datasets contain both a set of common variables (*matching variables*) and a set of specific variables; ii) the units observed in the datasets have been drawn independently from the same population.

Point ii) is particularly relevant in that it allows to clearly separate statistical matching from record linkage. With record linkage the two datasets contain identical individuals; with statistical matching the number of identical individuals in both datasets is typically small if not zero (Rässler S. 2002).

Let us consider files *A* and *B*. Some variables *Y* appear only in *A* whereas some variables *X* appear only in *B*. In both samples a set of matching variables *Z* can be observed. Through statistical matching, an artificial data set is generated where each unit records *Z*, *Y* and *X* values.

Various methods can be used to match two files *A* and *B*.

The nearest neighbour match

The nearest neighbour match (or hotdeck method) is a non parametric method frequently used to integrate datasets at micro level. According to this procedure, statistical matching can be regarded as an imputation problem.

Let us consider sample *A* as an incomplete data set where *X* variables are missing. *A* is then defined as the recipient sample. For every unit a_i , with i = (1,2,….,n_A), one *x* value from the observations of the donor sample *B* is selected. The donor unit b_j with j = (1,2,….n_B) is searched among the units belonging to *B* for which *Z* values are identical to those of the recipient unit a_i. These are called exact matches (Rässler S., 2002). Whenever a perfect match in terms of the common variables is not possible (especially if some common variables are continuous), the donor unit is selected on the basis of a distance measure d(*Z*). The donor unit is the *nearest neighbour* i.e. the unit with the smallest distance. When more donors are identified a random selection is performed.

The *A* and *B* files are merged in a single new and complete data set, $\tilde{A} = \{(x_1, z_1, \tilde{y}_1), \cdots, (x_{nA}, z_{nA}, \tilde{y}_{nA})\}$. This artificial sample is considered representative of the true population of interest. Notice that $\tilde{A}$ has the same number of elements n_A as the recipient sample and that $\tilde{y}_j$ is the value of the donor unit belonging to *B*.

This method is relatively simple but it has a relevant undesirable implication. In fact, the application of traditional statistical matching implies the so called Conditional independence assumption (CIA) between Y and X given Z. Conditional independence is produced for the variables not jointly observed even when such variables are conditionally dependent in reality. Since the CIA cannot be tested from the dataset $A \cup B$, this assumption could be wrong and, hence, misleading.

CIA can be roughly satisfied when there exists a strong predictive relationship between common variables Z and recipient-donor measures. For this reason the choice of suitable common variables is a crucial aspect of statistical matching, even more important than the matching technique itself (Rässler S. 2002).

Propensity score matching

Propensities scores are used to generate suitable control groups in observational studies in order to properly measure the effect of treatment and no treatment on one single unit (Rosenbaum, Rubin (1983)).

In the contest of statistical matching, propensity scores may be used instead of the nearest neighbour match method in order to identify the donor unit (Rässler S. 2002).

Let us consider *A* as the donor file and *B* as the recipient file. For both files a new variable S is defined. $S_i = 1$ for all units of the recipient file whereas $S_j = 0$ for the donor sample units.

Considering sample $A \cup B$, a logit model is estimated with S as the dependent variable and the common variables Z as independent variables (the so-called covariates). $S_i = 1$ if unit *i*, *i*=1,2,…n_{A+B} belongs to the (treated) recipient sample and $S_i = 0$ if unit *i,* belongs to the (control) recipient sample.

Once the model has been estimated it is possible to calculate the propensity score $\hat{e}(z_i)$ for each unit belonging to the $A \cup B$ set.
Propensity score is defined as the conditional probability of a unit to belong to a certain treatment group given the covariates Z.

$$e(z_i) = P(S = 1 \mid Z = z_i)$$

The matching is performed on the basis of the estimated propensity scores $\hat{e}(z_i)$. For every recipient unit (S=1) a unit is searched in the donor file (S=0) with the same or nearest propensity score estimate (Rässler S. 2002). The y-values are thus imputed to the recipient unit.
Figure 3 illustrates the principle of propensity score matching.

Figure 3 Principle of propensity score matching

Recipient sample

Unit number	Common variables Z	Specific variables X	S	$\hat{e}(z)$
1			1	0.6758
2			1	0.2856
….			…	
n_A			1	0.7881

Donor sample

Unit number	Common variables Z	Specific variables Y	S	$\hat{e}(z)$
1			0	0.2112
2			0	0.6711
….			…	….
n_B			0	0.5502

Rässler S. (2002), p. 25

As in the case of the nearest neighbour match it is necessary to select the covariates among the variables common to the donor (Eusilc) and the recipient (F_S (F_I or F_C) files.

4.2.1 Statistical matching of Eusilc and administrative files

Let us consider the Eusilc dataset as the donor file and the F_S (F_I or F_C) dataset as the recipient file. Furthermore Y and X are the specific variables of Eusilc and the F_S (F_I or F_C) datasets respectively. Z is a set of common variables. Statistical matching is aimed at finding for each F_S (F_I or F_C) unmatched record (Fig. 2) a donor unit in the Eusilc sample so that Y values may be imputed. The donor is defined as the Eusilc unit most similar to the recipient one with respect to a set of common variables, named matching variables.

In order to match Eusilc and the administrative files the following phases have to be accomplished:

Harmonization of datasets

Samples have to be harmonized in order to make the data comparable. Harmonization concerns the definition both of population units and variables. Inconsistencies must be solved through recoding of variables, imposing assumptions etc.

Choosing of the matching variables (covariates)

Theoretically, all common harmonized variables can be used for matching the samples. However, computational efficiency trades off with the number of matching variables (covariates). For this reason it is advisable to consider only common variables statistically connected with Y and X.

Performing the matching

A matching technique between Eusilc and F_S (F_I , F_C) files is applied. In case of the nearest neighbour match a distance function is defined in order to compare every pair of units from the donor and recipient files with respect to the matching variables Z. If applying the propensity score matching, a logit (or probit) model is estimated in order to obtain propensity scores for each unit of Eusilc and F_S (F_I , F_C) files. For every unit $i, i = 1,2,\cdots,n_{F_S}$ of the administrative dataset, the donor unit $j, j = 1,2,\cdots,n_{F_E}$ is identified as the one with the smallest distance to *i* with respect to the observed variables Z or propensity score estimates. Finally, all the observed information of variables Y of the donor unit *j* is imputed to the recipient unit *i.*

Given the sizes of the datasets (see section 2 for details), some Eusilc records would be probably imputed more than once in the recipient files (F_S ,F_I , F_C). This could artificially modify the variability of the distribution of the imputed variables in the synthetic file (as in D'Orazio *et al.* (2006), p. 35)

Assessing the accuracy of the statistical matching procedure

Usually a statistical matching is considered successful if the marginal and joint empirical distributions of the variables in the donor file are preserved in the statistically matched file (Rässler, 2002). This means that the empirical distributions of the Z and Y variables as they are observed in Eusilc dataset must be nearly the same in the statistically matched file.

D'Orazio *et al.* (2006) show the peculiarities of each phase of the matching process through the description of an application aimed at integrating the Banca d'Italia survey on Households Income and Wealth (SHIW) and the Istat Household Budget Survey (HBS). Full results of this application are in Coli *et al*. (2005).

Finally it is worth stressing that the output of the record linkage procedure between F_E and F_S (F_I , F_C) files is a complete dataset where either (Y,Z,X) are observed. Such datasets would provide auxiliary information for improving the quality of statistical matching (see D'Orazio *et al*., 2006).

References (SR/DSMAE)

Coli 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 delle famiglie italiane e l'Indagine Banca d'Italia sui bilanci delle famiglie italiane*. Istat, Documenti 12/2006, Roma.

Consolini P. (2003). *Administrative data based statistics: the case of non-pension cash benefit.* Proceeding of the 17th roundtable on business survey frames, volume II, pp. 423-430, Roma, 26-31 October, 2003.

Consolini P. (2004). *L'indagine sperimentale sull'archivio fiscale modd. 770 anno 1999: analisi della qualità del dato e stime campionarie*. Roma: Istat. (Contributi Istat, n. 29)

Consolini P., M. Di Marco, R. Ricci e S. Vitaletti. (2006). *Administrative and survey microdata on self-employment: the Italian experience with the Eu-Silc project*. Iariw 29th general conference, Joensuu (Finland), 20-26 August, 2006.

Copas J. B., F. J. Hilton (1990) *Record Linkage: Statistical Models for Matching Computer Records*, Journal of the Royal Statistical Society. Series A (Statistics in Society), Vol. 153, No. 3 (1990), pp. 287-320

Daas, P, Jeurissen, E., Boonstra, H.J., Nieuwenbroek, N. (2005) *Register theory: Registers and Statistics* Netherlands (in Dutch) Repport-nr. TMO-R&D-2005-01-31-PDAS Statistics Netherlands,

Daas P.J.H. and Fonville T.C. (2007), *Quality control of Dutch Administrative Registers: An inventory of quality aspects,* Seminar on Registers in Statistics - methodology and quality 21 - 23 May, 2007 Helsinki.

Di Marco M. (2006), *International comparability of microdata on incomes: lessons from the Eu-Silc project*. VIII International meeting on quantitative methods for applied sciences, Certosa di Pontignano (Siena), 11-13 settembre.

D'Orazio M., M. Di Zio, M. Scanu (2006), *Statistical matching: Theory and Practice*, Wiley & Sons, New York.

Epland J. (2006). *Challenges in income comparability: experiences from the use of the register data in Norwegian Eu-Silc*. VIII International meeting on quantitative methods for applied sciences, Certosa di Pontignano (Siena), 11-13 settembre.

Eurostat (2000). Assessment *of the quality in statistics, Item4: Definition of quality in statistics*. 4-5 April, Luxembourg.

Eurostat (2003). Quality assessment of administrative data for statistical purposes, Working group "Assessment of quality in statistics", Sixth meeting, Luxembourg, 2-3 October 2003, Item 6.

Eurostat (2009). Cross Sectional UDB (User Data Base). Version 2007-2 from 01-08-09

Fellegi I. P., A. B. Sunter. (1969). *A theory for record linkage*, Journal of the American statistical association. vol. 64, pp. 1183-1210.

Deville J. C. , Lavallee P. (2006). *Indirect Sampling: the foundations of the generalized weight share method*. Survey Methodology, 32:165-176

Herzog T. N., Scheuren F. J. e W. E. Winkler. (2007). *Data quality and record linkage techniques*. New York: Springer ed.

Istat (2008) *L'indagine europea sui redditi e le condizioni di vita delle famiglie (Eu-silc)*, Istat, Metodi e Norme N. 37.

Lavallee, Caron (2001). *Estimation using the generalized weight share method. The case of record linkage*. Survey Methodology, 27:155-1

Maletic, J. I. e A. Marcus (2000). *Data cleansing: beyond integrity analysis*. Proceedings of the conference on information quality (IQ2000), Boston, pp. 200-209.

Nordic Statistical Institutes (2007) *Register-based statistics in the Nordic countries* . Review of the best practices with focus on population and social statistics. UNECE.

Pratesi M., Salvati N. (2005) *Sampling Strategies and Multifunctionality in Agricultural Surveys*, proceedings of the Convegno Intermedio SIS "Statistica e Ambiente", 21-23 settembre 2005, Messina.

Rässler S. (2002), *Statistical matching: a frequentist theory, practical applications and alternative Bayesian approaches*, Lecture notes in statistics, 168 – New York Springer.

Rendtel U. et al. (2004). *Report on quality of income data*. WP5 Chitex project, final version, January 2004.

Rosenbaum P. R., Rubin D. B. (1983) *The central role of the propensity score in observational studies for causal effects*, Biometrika, Vol. 70, No. 1, pp. 41-55

Scanu M. (2003). *Metodi statistici per il record linkage*, Roma: Istat. (Metodi e Norme, n. 16)

Spinelli V. (2007). *Processo di acquisizione e trattamento informatico degli archivi relativi al modello di dichiarazione 770*. Roma: Istat. (Documenti Istat, n. 4)

Statistics Finland (2004), *Use of Register and Administrative Data Sources for Statistical Purposes, Best Practices of Statistics Finland*. Handbook 45. Statistics Finland, Helsinki, Finland.

Thomas, M. (2005) *Assessing Quality of Administrative Data*. Survey Methodol. Bull. 56, 74-84.

UNECE (United Nations Economic Commission for Europe) (2007), *Register-based statistics in the Nordic countries: review of best practices with focus on population and social statistics*, United Nation.

Van der Laan P. (2000). *Integrating administrative registers and household surveys*. Netherlands Official Statistics vol. 15, Summer 2000.

Wallgren A. and Wallgren B. (2006): *Register-Based Statistics - Administrative Data for Statistical Purposes*. John Wiley & Sons, Ltd

Working group "Assessment of quality in statistics" (2003) Item6: *Quality assessments of administrative data for statistical purposes*. Sixth meeting, 2-3 October, Luxembourg.

Working group "Assessment of quality in statistics" (2003) Item 4.2d: Methodological documents, Handbook *How to make a Quality Report*. Sixth meeting, 2-3 October, Luxembourg.

Appendix A Dataset files structure (SR)

Table A.1 SIATEL dataset

Variable name (italian)	**Variable name (english)**
Codice fiscale del dichiarante	Fiscal code
Flag dichiarante/coniuge	If the partner of persons who declare the income
Cognome	Surname
Nome	Name
Codice catastale comune nascita	Municipality code
Data di nascita	Date of birth
Sesso	Sex
Stato civile	Marriage status
Coniuge a carico	If dependent spouse
Figli a carico	N° of dependent children
Codice catastale del comune del domicilio fiscale attuale	Code of the municipality of residence for tax current
Indirizzo attuale	Current Address
CAP attuale	Current CAP
Totale redditi dichiarati	Total income declared
Totale reddito dominicale imponibile	Total land income taxable
Totale reddito agrario imponibile	Total agricultural income taxable
Totale imponibile fabbricati	Subtotal buildings
Totale redditi lavoro dipendente e assimilati	Total income of employees and similar
Totale dei redditi assimilati al lavoro dipendente per i quali non spettano le deduzioni	Total income assimilated to employees for which no deductions payable
Tipologia di reddito prevalente	Type of income category
Tipo di contratto (determinato/indeterminato)	Type of contract (permanent / temporary)
Reddito di impresa di allevamento di spettanza dell'imprenditore	Farming business income attributable to the contractor
Reddito (o perdita)	Income (or loss)
Totale reddito di lavoro autonomo	Total income of self-employment
Reddito d'impresa (o perdita) di spettanza dell'imprenditore	Business income (or loss) attributable to the contractor
Totale reddito di partecipazione	Total income participation
Totale redditi di partecipazione in società esercenti attività d'impresa	Total income from shares in companies engaged in business activities
Redditi (o perdite) di partecipazione in associazioni fra artisti e professionisti	Income (or loss) of participation in associations between artists and professionals
Redditi di partecipazione in società semplici	Income from participation in simple societies
Redditi di capitale - totale	Total capital income
Redditi diversi - reddito netto	Other net income
Attività sportive dilettantistiche - reddito imponibile	Taxable income from sport activities
Altri redditi di lavoro autonomo - totale netto compensi, proventi e redditi	Other income from self-employment
Redditi a tassazione separata - tassazione ordinaria	Income with separate taxation – normal tax
Reddito imponibile (quadro RN)	Total taxable income
Imposta netta	Net tax
Plusvalenze	Capital gains
Riserve costituite prima della trasformazione (art.170, comma 4)	
Redditi soggetti a tassazione separata	
Redditi derivanti dalla cessione di partecipazione	
Rimborso di oneri dedotti in precedenti esercizi	Remboursement of tax payed previous years
Reddito imponibile	Total income
Addizionale comunale all'Irpef dovuta -casella esenzione	Lacal tax exemption
Addizionale comunale all'Irpef dovuta -importo	Local tax amount
Reddito imponibile	Total taxable income for local tax
Reddito complessivo netto	Total net income

Table A.2 IDOL dataset

Variable name (italian)	**Variable name (english)**
Codice fiscale	Fiscal Code
Sesso	Sex
Data di nascita	Birthdate
Comune di nascita	City of birth
Cittadinanza	Citizenship
Stato civile	Marital status
Comune di residenza	Municipality of residence
Indirizzo di residenza	Home address
Frazione di residenza	Locality of residence
Cap di residenza	Postcode of residence
Comune di domicilio	Common residence
Indirizzo di domicilio	Home address
Frazione di domicilio	Locality of residence
Cap di domicilio	Cap domicile
Titolo di studio (Il titolo più alto conseguto)	Degree (the highest title achieved)
Reddito dichiarato Per i carichi familiari	Declared income to meet family
Data dichiarazione reddito	Income Statement Data
Tipo di reddito Annuale, mensile, etc.	Type of annual income, monthly, etc.
Numero persone a carico Tratto da PERSONE_CARICO	Number of dependents handled by personnel charged
Invalido	Invalid
Carico familiare	Family burden
Tipo di carico	Load type
Situazione (vecchio flag) Campo che descriva se sono iscritti oppure se sono presenti in anagrafica a causa di una comunicazione	Location (old flag) field that describes whether they are members or are present in the registry because of a communication
Tipo iscrizione	Registration Type
Tipo comunicazione Se non iscritto, tipo di comunicazione che determinato presenza in anagrafica	Type statement If not registered, type of communication that determined presence in registry
Motivo cancellazione Se presenti, ma cancellati	Reason for cancellation if they exist, but erased
Data cancellazione	Cancellation Date
Data ultimo aggiornamento	Last update
Data inizio disoccupazione Per gli iscritti, la data di iscrizione, per i non iscritti non avviati la data dell'ultima cessazione	Start date unemployment For members, the date of registration for non-members not start the date of termination
Data ultimo avviamento Ultimo avviamento non ancora cessato	Last seed Last Goodwill not yet ceased
Tipo di contratto	Contract type
Tipo orario	Type time
Qualifica avviamento Qualifica professionale Istat	Qualification ZIP Professional qualification Istat
Contratto collettivo applicato	Applicable collective agreement
Livello di inquadramento	Level classification
Retribuzione/compenso Se possibile, retribuzione mensile o complessiva annuale (considerando tutti i rapporti di lavoro registrati)	Salary / compensation If possible, pay monthly or annual total (considering all employment relationships registered)
Tipo retribuzione (Orario, giornaliero, mensile, totale)	Type earnings Hourly, daily, monthly, total
Ore settimanali medie	Average weekly hours

Table A.3 MIROD dataset

Personal data table:

Variable name (italian)	Variable name (english)
abitazione	home
accoglienza presso	reception at
anno di arrivo in italia	year of arrival in Italy
assistente sociale	Social Worker
cedolino di richiesta/rinnovo permesso	cedolino request / permit renewal
centro operativo primo contatto	operations center first contact
cittadinanza	Citizenship
coabitazione	cohabitation
codice scheda	code card
cognome	surname
cognome e nome del coniuge/convivente	Name of spouse / partner
comune di residenza	municipality of residence
comune dimora abituale	common habitual residence
con chi vive	with those living
condizione professionale	professional status
condizione professionale nel paese di origine	professional status in the country of origin
convivente	cohabitant
data chiusura pratica	Closing date practice
data di nascita	birthday
data scheda	given tab
dimora abituale	usual residence
diocesi	Dioceses
età	age
figli altrove	children elsewhere
figli in italia conviventi	children living in Italy
figli in italia non conviventi	children not living in Italy
figli rimasti in patria	children remaining at home
ha figli	childless
ha un assistente sociale	has a social worker
luogo di nascita	birthplace
motivo di chiusura della pratica	reason for closure of the practice
motivo rilascio permesso di soggiorno	why issue a residence permit
nessun documento posseduto	possessed no documents
nessun documento presentato	No documents submitted
nomade	nomad
nome	name
nome gruppo nomade	nomadic group name
numeri telefonici	numbers
posizione nella professione	employment status
possesso del permesso di soggiorno	possession of a residence permit
possesso della carta di soggiorno	possession of a residence permit
professione in italia	profession in Italy
professione nel paese di origine	profession in the country of origin
provenienza coniuge/convivente	from spouse / partner
religione	religion
residenza	residence
residenza coniuge/convivente	resident spouse / partner
richiesta carta di soggiorno	request a residence permit
scadenza permesso di soggiorno	expired permit
seconda cittadinanza	second citizenship
servizio/associazione/parrocchia attualmente in contatto	service / association / parish currently in talks
servizio/associazione/parrocchia da cui proviene	service / association / parish from which
sesso	sex
stato civile	marital status
tipo dimora abituale	type usual residence

Variable name (italian)	Variable name (english)
titolo di studio	qualification
zona	area
conteggio di centro operativo nota	count operations center known
conteggio di data nota	count given note
conteggio di oggetto della nota	count subject of note
conteggio di tipo della nota	count type of Note
problemi familiari	family problems
povertà/problemi economici	Poverty / Economic
conteggio di centro operativo bisogno	count operations center need
conteggio di data fine rilevazione	count survey end date
conteggio di data inizio rilevazione	count data collection beginning
conteggio di durata rilevazione bisogno	count duration detection needs
conteggio di stato rilevazione	count detected
problemi di reddito	problems of income
handicap o disabilita'	handicap or disabled
problemi di salute	health problems
problematiche abitative	housing issues
bisogni in migrazione/immigrazione	needs in migration / immigration
problemi del lavoro	employment issues
problemi di occupazione/lavoro	problems of employment / work
detenzione e giustizia	detention and justice
dipendenza	dependence
problemi di istruzione	education problems
altri problemi	Other problems
handicap/disabilità	handicap / disability
dipendenze	dependencies
non individuato	unidentified

Expressed needs table:

Variable name (italian)	Variable name (english)
centro operativo primo contatto	operations center first contact
cittadinanza	Citizenship
codice scheda	code card
diocesi	Dioceses
sesso	sex
beni e servizi materiali - vestiario	material goods and services - clothing
sussidi economici - per pagamento bollette/tasse	Economic aid - to pay bills / taxes
beni e servizi materiali - viveri	material goods and services - food
sanità - farmaci	HEALTH - medicines
sussidi economici - per spese sanitarie	Economic aid - for health costs
beni e servizi materiali - biglietti per viaggi	material goods and services - tickets for travel
beni e servizi materiali - altro	material goods and services - other
non specificato	Unspecified
coinvolgimenti - coinvolgimento di persone o famiglie	involvement - involvement of individuals or families
ascolto - primo ascolto	listening - first listen
beni e servizi materiali - mobilio, attrezzatura per la casa	material goods and services - furniture, household equipment
lavoro - lavoro generico	work - Generic
lavoro - part time	work - part time
beni e servizi materiali - alimenti e prodotti per neonati	material goods and services - food and products for babies
istruzione - doposcuola e sostegno scolastico (lezioni)	education - after-school and school support (tuition)
beni materiali - vestiario	material goods - clothing
coinvolgimenti - coinvolgimento di parrocchie e/o gruppi parrocchiali	involvement - involvement of the parishes and / or church groups
ascolto - ascolto con discernimento e progetto	listening - listening with discernment and project
vitto - distribuzione viveri	food - food distribution
sanità - analisi, esami clinici	HEALTH - analysis, clinical
vitto - mensa	Food - canteen
alloggio - accoglienza a lungo termine (casa, appartamento in affitto)	accommodation - accommodation in the long term (house, apartment for rent)

Variable name (italian)	**Variable name (english)**
sussidi economici - per alloggio	allowances - for housing
coinvolgimenti - coinvolgimento enti pubblici	involvement - public involvement
consulenza professionale - legale	professional advice - legal
animazione promozionale - in comunita' di reiserimento o centri di riabilitazione	Promotional animation - in the community 'of reiser or rehabilitation centers
sussidi economici - per acquisto di alimentari	Economic aid - for the purchase of food
ascolto - ascolto (semplice ascolto/primo ascolto)	listening - listening (easy listening / first listening)
lavoro - altro	Work - other
lavoro - tempo pieno	work - full time
sussidi economici - restituzione prestito	Economic aid - loan repayment
animazione promozionale - coinvolgimento parrocchie e gruppi parrocchiali	Promotional animation - involvement parishes and parish groups
sussidi economici - microcredito/prestito	Economic aid - microcredit / loan
beni e servizi materiali - apparecchiature e/o materiale sanitario	goods and services and equipment - equipment and / or medical equipment
lavoro - tempo pieno convivente	work - full time partner
lavoro - saltuario, occasionale	job - casual, occasional
coinvolgimenti - coinvolgimento di gruppi laici di volontariato	involvement - involvement of lay groups of voluntary
orientamento - per esigenze abitative	guidance - for housing needs
coinvolgimenti - coinvolgimento enti privati o del terzo settore	involvement - involvement of private entities or third sector
lavoro - lavoro generico - tempo pieno	workers - Generic - full time
alloggio - pensionato	accommodation - pensioner
alloggio - accoglienza in casa famiglia/comunità alloggio	accommodation - welcome home family / community housing
lavoro - lavoro specifico - part-time	work - specific work - part-time
sanità - operazioni chirurgiche	healthcare - surgery
sussidi economici - sussidi a fondo perduto - per pagamento bollette	Economic aid - subsidies grants - to pay bills
sussidi economici - sussidi a fondo perduto - per documenti	Economic aid - repayable grants - for documents
alloggio - alloggio generico	Accommodation - Accommodation generic
orientamento - a servizi socio sanitari	orientation - a social-health services
alloggio - pronta e prima accoglienza (ostello, dormitorio, tende, ecc.)	alloggio - pronta e prima accoglienza (ostello, dormitorio, tende, ecc.)
beni e servizi materiali - mensa	material goods and services - cafeteria
alloggio - casa famiglia	accommodation - Family Home
animazione promozionale - coinvolgimento di gruppi di volontariato (non parrocchiali)	Promotional animation - involvement of volunteer groups (non-parochial)
lavoro - lavoro specifico	work - specific work
lavoro - lavoro generico - part time	workers - Generic - part time
istruzione - corsi di lingua italiana	Education - courses in Italian language
orientamento - per pratiche burocratiche, legali	Guidance - for paperwork, legal
animazione promozionale - coinvolgimento enti pubblici	Promotional animation - public involvement
alloggio - altro	Accommodation – another
beni e servizi materiali - buoni carburante	Material goods and services- fuel
animazione promozionale - coinvolgimento enti privati	Promotional animation- private involvement
coinvolgimenti - altro tipo di coinvolgimento	Involvement- other kinds of involvement
sussidi economici - altre richieste/interventi economici - per acquisto cibo, generi alimentari	Economic aid- other requests/economic interventions- food
consulenza professionale - psico sociale	Professional help- psycho-social
alloggio - comunita' alloggio	Housing- community housing
sussidi economici - per altri motivi	Economic aid- for other reasons
beni materiali - altro	Material goods- other
orientamento - per problemi occupazionali/pensionistici	Career guidance- for professional/pension-related problems
altre richieste/interventi - altre richieste/interventi	Other requests/interventions- other requests/intervention
consulenza professionale - altro	Professional consultancy- other
scuola/istruzione - corsi di lingua italiana	School/Education- Italian language courses
istruzione - corsi professionali	Education- training courses
ascolto - progetto di intervento	Listening- intervention project
sanita' - visite mediche (prestazioni specialistiche, consulenza sanitaria)	Health- medical assistance
lavoro - altro - part time	Work-other-part time

Variable name (italian)	**Variable name (english)**
segretariato sociale - per orientamento/invio a servizi	Social secretariat- for guidance/transferring to services
sanita' - medicinali	Health- medicines
altre richieste/risposte - igiene personale, bagni, docce	Other requests/answers- personal care, baths, showers
sussidi economici - altre richieste/interventi economici - per riscatto bagagli	Economic aid- other requests/economic interventions- to buy back luggage
beni materiali - mezzo di trasporto	Material goods- transportation mean
alloggio - dormitorio-ostello	Housing- dormitory-hostel
lavoro - lavoro specifico - tempo pieno	Work- specific work-full time
animazione promozionale - prevenzione (secondaria o terziaria)	Promotional animation- prevention (secondary or tertiary)
sussidi economici - altre richieste/interventi economici - per pagamento bollette	Economic aid- other requests/economic intervention- to pay bills

Table A.4 EUSILC dataset

Household Register (D-file)

Variable code	Variable name
DB010	Year of the survey
DB020	Country
DB030	Household ID
DB040	Region
DB060	PSU-1 (first stage)
DB062	PSU-2 (second stage)
DB070	Order of selection of PSU
DB075	Rotational group
DB090	Household cross-sectional weight
DB100	Degree of urbanisation
DB110	Household status

Personal Register (R-file)

Variable code	Variable name
RB010	Year of the survey
RB020	Country
RB030	Personal ID
RB040	Current household id
RB050	Personal cross-sectional weight
RB060	Personal base weight
RB070	Quarter of birth
RB080	Year of birth
RB090	Sex
RB100	Sample person or co-resident
RB110	Membership status
RB120	Moved to
RB140	Quarter moved out or died
RB150	Year moved out or died
RB160	Number of months in household during the income reference period
RB170	Main activity status during the income reference period
RB180	Quarter moved in
RB190	Year moved in
RB200	Residential status
RB210	Basic activity status
RB220	Father ID
RB230	Mother ID
RB240	Spouse/partner ID
RB245	Respondent status

RB250	Data Status
RB260	Type of interview
Variable code	Variable name
RB270	Personal ID of proxy
RL010	Education at pre-school
RL020	Education at compulsory school
RL030	Child care at centre-based services
RL040	Child care at day-care centre
RL050	Child care by a professional child-minder at child's home or at child-minder's home
RL060	Child care by grand-parents, others household members (outside parents), other relatives, friends or neighbours
RL070	Children cross-sectional weight for child care
RX010	Age at the date of interview
RX020	Age at the end of the income reference period
RX030	Household ID

Household Data (H-file)

Variable code	Variable name
HB010	Year of the survey
HB020	Country
HB030	Household ID
HB050	Quarter of household interview
HB060	Year of household interview
HB070	Person responding the household questionnaire
HB080	Person 1 responsible for the accommodation
HB090	Person 2 responsible for the accommodation
HB100	Number of minutes to complete the household questionnaire
HH010	Dwelling type
HH020	Tenure status
HH030	Number of rooms available to the household
HH031	Year of contract or purchasing or installation
HH040	Leaking roof, damp walls/floors/foundation, or rot in window frames or floor
HH050	Ability to keep home adequately warm
HH060	Current rent related to occupied dwelling
HH061	Subjective rent
HH070	Total housing cost
HH080	Bath or shower in dwelling
HH090	Indoor flushing toilet for sole use of household
HS010	Arrears on mortgage or rent payments
HS020	Arrears on utility bills
HS030	Arrears on hire purchase instalments or other loan payments
HS040	Capacity to afford paying for one week annual holiday away from home
HS050	Capacity to afford a meal with meat, chicken, fish (or vegetarian equivalent) every second day
HS060	Capacity to face unexpected financial expenses
HS070	Do you have a telephone (including mobile phone)?
HS080	Do you have a colour TV?
HS090	Do you have a computer?
HS100	Do you have a washing machine?
HS110	Do you have a car?
HS120	Ability to make ends meet
HS130	Lowest monthly income to make ends meet
HS140	Financial burden of the total housing cost
HS150	Financial burden of the repayment of debts from hire purchases or loans
HS160	Problems with the dwelling
HS170	Noise from neighbors or from the street
HS180	Pollution, grime or other environmental problems
HS190	Crime violence or vandalism in the area
HY010	Total household gross income

HY020	Total disposable household income
Variable code	Variable name
HY022	Total disposable household income before social transfers other than old-age and survivor's benefits
HY023	Total disposable household income before social transfers including old-age and survivor's benefits
HY025	Within-household non-response inflation factor
HY030G/HY030N	Imputed rent
HY040G/HY040N	Income from rental of a property or land
HY090G/HY090N	Interest, dividends, profit from capital investments in unincorporated business
HY050G/HY050N	Family/Children related allowances
HY060G/HY060N	Social exclusion not elsewhere classified
HY070G/HY070N	Housing allowances
HY080G/HY080N	Regular inter-household cash transfer received
HY100G/HY100N	Interest repayments on mortgage
HY110G/HY110N	Income received by people aged under 16
HY120G/HY120N	Regular taxes on wealth
HY130G/HY130N	Regular inter-household cash transfer paid
HY140G/HY140N	Tax on income and social contributions
HY145N	Repayments/receipts for tax adjustment
HX010	Change rate
HX020	Work intensity status
HX040	Household size
HX050	equivalised household size
HX060	Household type
HX070	Tenure status
HX080	Poverty indicator
HX090	equivalised disposable income
HX100	equivalised disposable income quintiles

Personal Data (P-file)

Variable code	Variable name
PB010	Year of the survey
PB020	Country
PB030	Personal ID
PB040	Personal cross-sectional weight
PB050	Personal base weight
PB060	Personal cross-sectional weight for selected respondent
PB080	Personal base weight for selected respondent
PB100	Quarter of the personal interview
PB110	Year of the personal interview
PB120	Minutes to complete the personal questionnaire
PB130	Quarter of birth
PB140	Year of birth
PB150	Sex
PB160	Father ID
PB170	Mother ID
PB180	Spouse/partner ID
PB190	Marital status
PB200	Consensual Union
PB210	Country of birth
PB220A	Citizenship 1

PE010	Current education activity
Variable code	Variable name
PE020	ISCED level currently attended
PE030	Year when highest level of education was attained
PE040	Highest ISCED level attained
PH010	General health
PH020	Suffer from any a chronic (long-standing) illness or condition
PH030	Limitation in activities because of health problems
PH040	Unmet need for medical examination or treatment
PH050	Main reason for unmet need for medical examination or treatment
PH060	Unmet need for dental examination or treatment
PH070	Main reason for unmet need for dental examination or treatment
PL015	Person has ever worked
PL020	Actively looking for a job
PL025	Available for work
PL030	Self-defined current economic status
PL035	Worked at least 1 hour during the previous week
PL040	Status in employment
PL050	Occupation (ISCO-88 (COM))
PL060	Number of hours usually worked per week in main job
PL070	Number of months spent at full-time work
PL072	Number of months spent at part-time work
PL080	Number of months spent in unemployment
PL085	Number of months spent in retirement
PL087	Number of months spent studying
PL090	Number of months spent in inactivity
PL100	Total number of hours usually worked in second, third… jobs
PL110	NACE (REV 1.1)
PL120	Reason for working less than 30 hours
PL130	Number of persons working at the local unit
PL140	Type of contract
PL150	Managerial position
PL160	Change of job since last year
PL170	Reason for change
PL180	Most recent change in the individual's activity status
PL190	When began first regular job
PL200	Number of years spent in paid work
PL210A	Main activity on January
PL210B	Main activity on February
PL210C	Main activity on March
PL210D	Main activity on April
PL210E	Main activity on May
PL210F	Main activity on June
PL210G	Main activity on July
PL210H	Main activity on August
PL210I	Main activity on September
PL210J	Main activity on October
PL210K	Main activity on November
PL210L	Main activity on December
PY010G/PY010N	Employee cash or near cash income
PY020G/PY020N	Non-Cash employee income
PY030G	Employer's social insurance contribution
PY035G/PY035N	Contributions to individual private pension plans
PY050G/PY050N	Cash benefits or losses from self-employment
PY070G/PY070N	Value of goods produced by own-consumption
PY080G/PY080N	Pension from individual private plans
PY090G/PY090N	Unemployment benefits
PY100G/PY100N	Old-age benefits
PY110G/PY110N	Survivor' benefits
PY120G/PY120N	Sickness benefits

PY130G/PY130N	Disability benefits
Variable code	Variable name
PY140G/PY140N	Education-related allowances
PY200G	Gross monthly earnings for employees
PX010	Exchange rate
PX020	Age at the end of the income reference period
PX030	Household ID
PX040	Respondent status
PX050	Activity status

Appendix B Quality aspects of the metadata checklist

Metadata aspects	**Explanation**
Purpose	What is the original purpose of the registration?
Basis	Legal basis on which the register is kept.
Law / Legal provision / Regulation / Agreements	Reference to the legal provision or agreement on which the register is based.
Population (conceptual def.)	The population(s) recorded in the register; the object type(s) should be described (e.g. persons, enterprises etc).
Geographic limit	The geographic area of the population(s) in the register.
Time limit	The period(s) for which the data in the population(s) is registered.
Identification keys	Unique keys in the register that can be used to identify the recorded object type(s). This could be more than one.
Collection	The way in which the data is collected by the register holder.
Maintenance *	The way in which the data is maintained by the register holder.
Editing *	If and how the data is edited by the register holder.
Selection	Often SN does not receive a full copy of the register but only a selected set. Check if and what sort of selection is made.
Time dimension *	What time event is recorded?
Occurrence	Is the time of occurrence recorded for each event.
Registration	Is the time of registration recorded for each event.
Quality control	Any form of quality control that is (regularly) performed by the register holder.
File format/Data structure *	The file format in which the data is made available.
Classifications / Variable description (key variables *)	Explanation of the classifications and variables used by the register holder.
Supplier agreement *	Agreement between the register holder (data supplier) and SN.
Privacy considerations *	If the register contains unit level identification keys there should be an agreement that the legal rights of the individual citizen with regard to the protection and integrity of his/her data is not violated.

Source: Daas P.J.H. and Fonville T.C. (2007)