SAMPLE
Small Area Methods for Poverty
and Living Condition Estimates

SAMPLE DELIVERABLE 11

INTEGRATING DATA MODEL – SECOND 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.	11
Deliverable name:	Integrating data model – Second release
WP no.:	3.3
Lead beneficiary:	1
Nature:	Report
Dissemination level:	PU
Due delivery date from Annex I:	30/5/2010
Actual delivery date:	26/07/2011
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 – SECOND 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.	11
Document title:	Integrating data model – Second release
Status:	Final
Editors:	Alessandra Coli, UNIPI-DSMAE (a.coli@ec.unipi.it)
Authors:	Alessandra Coli, UNIPI-DSMAE (a.coli@ec.unipi.it) Moreno Toigo, SR (simurg@simurgricerche.it) Stefano Marchetti, UNIPI-DSMAE (stefano.marchetti@for.unipi.it)
Document Editing:	Marta Garro (garro@sample-project.eu), UNIPI-DSMAE
Language revision:	Marta Garro (garro@sample-project.eu), UNIPI-DSMAE

Summary

1 Introduction

The SAMPLE Consortium got access to the following databases concerning the Province of Pisa:

- the Job centre database (JC)
- the Revenue agency database (RA)
- the Caritas database (CARITAS).

The first part of Deliverable 11 is devoted to the description and analysis of the acquired administrative datasets in order to understand what kind of information could be obtained for studying poverty at the local level. We find that it is possible to analyze the living conditions of two different population subgroups: the "socially -integrated" group, composed mainly by people with a house and/or a job; the "socially-emarginated" group composed mainly by Italian homeless and migrants. It is worth stressing how official statistics do not generally permit to analyze this last segment of poor population.

The second part focuses on the integration of PI-Silc data with the administrative datasets through a record linkage procedure.

Among the final outcomes of the SAMPLE project there was the building of a model for estimating administrative indicators corrected for self selection bias. The method (see Deliverable 9) relied on the possibility of linking the PI-Silc and the administrative databases at individual level.

Unfortunately we could not access complete and detailed input data. This reshaped our objectives as follows: i) the building of an integrated database for the year 2008 having PI-Silc as the core dataset and the linked administrative data sets (JC and RA) as satellites for in depth analysis on specific aspects (labour, income, taxes); ii) The estimate of poverty indicators using the matched data.

The Deliverable is organized as follows. In § 2 we describe the quality assessment procedure applied to the administrative datasets. Section 3 is devoted to the analysis of the administrative datasets. Section 4 concerns the matching procedure between the PI-Silc and the administrative datasets (JC and RA). Finally section 5 contains some concluding remarks.

2 Quality assessment of the administrative datasets [SR]

2.1 The Job Centre database

2.1.1 Dataset optimisation

Provincial Job Centers are public offices, depending by 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)[1]. The functions and the administrative activity of these offices are managed by a complex datawarehouse system called JC (Incontro Domanda Offerta di Lavoro). See D9 for the description of the JC data source.

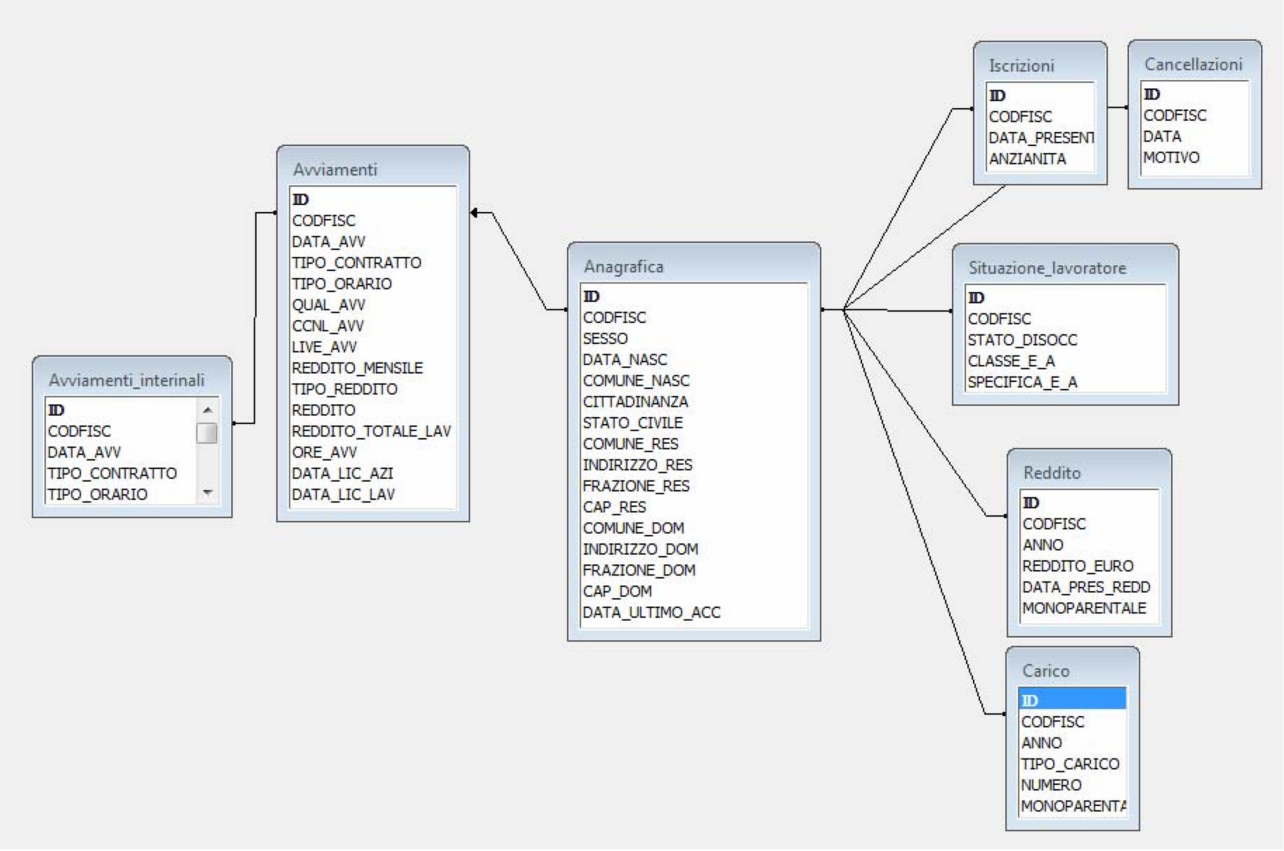
The SAMPLE Consortium obtained access to the following tables:

Table 1: JC: tables in the provided dataset

Table	Description	Type	N. records
anagrafica.csv	Register of personal data of job seeker and employee	Stock	231.443
carico.csv	Register of personal data of job seeker and employee - Family detail	Stock	33.692
reddito.csv	Register of personal data of job seeker and employee - Revenue detail	Stock	22.431
titoli_studio.csv	Register of personal data of job seeker and employee - Education detail	Stock	180.282
avviamenti.csv	Register of events concerning job position changes over time	Flow	189.208
avviamenti_interinali.csv	Register of events concerning job position changes over time (detail)	Flow	18.726
cancellazioni.csv	Register of persons cancelled from Job Center register	Flow	55.619
iscrizioni.csv	Register of enrollements at Job Center list	Flow	77.432

The tables are linked by job seeker/employee personal code (Codice Fiscale) as showed by the following figure.

Figure 1: JC: dataset schema and relations between tables



[1] Cocopro (freelancer) are included.

The provided dataset is the results of queries on a more complex datawarehouse used to manage Job Center data and procedures. The datawarehouse is structured and optimized mainly for administrative purposes, but also statistical needs have been taken into account in the software planning. Therefore, all the fields in the provided tables are well structured and can be treated and processed as statistical variables.

Such as all the administrative data sources, the JC dataset is affected by various kind of errors (non-response errors, measurement errors, processing errors, keying errors, coding errors). The non response error in this case depends on the fact that some eligible target units (the unemployed and the employees) is not registered. Measurement errors could affect mainly some variables like earnings (job wage) and family composition presenting many outlier. Moreover, these data come from statements by the users at the time of the first registration and they are not regularly updated. Processing errors could have affected the export phase from the JC datawarehouse to the tables provided for the SAMPLE project. Keying and coding errors are more common. The data input is realized by administrative employees of the Job Center and there aren't internal procedures to control and to assess the quality of the data. Most input fields are gathered through drop-down lists, but some fields are gathered as free text. Therefore they need to be optimized to correct input errors and to standardize the categories adopted to classify the information.

The JC dataset has been optimized and standardized mainly to allow the probabilistic linkage with EUSILC data. We have done the following data treatment:

- Calculate new derived individual variables;
- Fill the missing value with estimated or proxy data;
- Correct the keying and coding errors;
- Deal with outliers;
- Standardize free text variables;
- Georeference the dataset referring to the address of residence.

New variables created:

- *Education attainments*: this variable (categorical) is derived from the joining of the register with the table containing the qualifications documented by the users at the moment of the registration and the all subsequent updates; we selected the higher qualification inserted in the system and then we attached to the corresponding record in the register;

- *Wage*: we calculated the amount of the yearly wage for the users who have at least one job position recorded in JC archives; we selected the last wage recorded in the archive and assigned this value to a new variable linked with the personal register; it has been associated with each individual record also a time variable (the date when the employment contract starts).
- *Income*: The Job Centres records for a small number of users also few information regarding the family total income; this information is gathered when the user asks or wants to maintain facilitations or benefits, so it is recorded in very few cases (only 8%). We have attached this new variable (numeric) to each populated record in the personal register.
- *Family burden*: For similar reasons and situations (the request of certain benefits) the Job Centres records also information about the user's family composition; this information has been used to calculate two new variables: the variable "family dimension" (numeric) and the variable "single-parents family" (dicothomous).

Regarding missing value we didn't do any statistical treatment for the missing value replacement, except for the address variable. We tried to reduce the number of missing addresses because we needed to georeference the data for the linkage with EUSILC data (see paragraph 3). We created a new variable as a result of the mix of the address of residence with the address of domicile. In this way we lowered the number of missing cases from 16.241 to 9.832.

The resulting address toponyms have been normalized to improve the georeferencing process. In particular, we have corrected the street odonym (street names) correcting errors and standardizing the input variability as showed in the following synthetic table:

Original value	Corrected value
"N."	""
"V."	"VIA"
"LOC"	"LOCALITA' "
"C.SO"	"CORSO "
"P.ZZA"	"PIAZZA "
"POD."	"PODERE "
"P.ZA"	"PIAZZA "
"L.GO"	"LARGO "
"V.LE"	"VIALE "
"P.LE"	"PIAZZALE "
"NS.D."	""
"LOC."	"LOCALITA'"
"B.GO"	"BORGO "
"c/o"	""
","	" "
"LGO"	"LARGO"

"LUNG'ARNO"	"LUNGARNO"
"LOC.LA"	"LOCALITA'"
"P.ZZALE"	"PIAZZALE"
"FRAZ."	"FRAZIONE"
"L.NO"	"LUNGARNO"
" . "	" "
"-"	" "
"PZZA"	"PIAZZA"

Geocoding: For the probabilistic linkage between EUSILC, JC and RA we needed some common key variables (sex, age, place of residence) in every datasets. EUSILC is provided by census section, but not with the full residence address. RA and JC have the residence/domicile address but not the census section. To calculate the census section associated with each record in RA and in JC datasets we needed to *geocode* [2] the postal address and then to run a geographic query to link the proper census section to the related coordinates. With thousands of records could be a very hard work. Thankfully, we could take advantage of a free geocoding batch service provided by Tuscany Region[3] doing the kind of processing we need. The geocoding process inserted in the database structures the following new variables:

- Results of geocoding
- Input string
- Province Code
- Commune name
- Hamlet name
- Street name (odonym)
- Toponym
- Number
- Istat Region Code
- Istat Province Code
- Istat Commune Code
- Istat Census Section Code (1991)
- X coordinate of centroid Istat Census Section (1991)
- Y coordinate of centroid Istat Census Section (1991)
- Postal code
- X coordinate of street start
- Y coordinate of street start
- X coordinate of street end

[2] Geocoding is the process of finding associated geographic coordinates (expressed as latitude and longitude) from other geographic data, such as street addresses, or zip codes (postal codes).

[3] This service is provided through the site http://mappe.rete.toscana.it/. The user must prepare the dataset in a given fixed format, upload it to the web service and then download the geocoded datasets.

- Y coordinate of street end
- X coordinate of civic number
- Y coordinate of civic number
- Istat Census Section Code (2001)
- X coordinate of centroid Istat Census Section (2001)
- Y coordinate of centroid Istat Census Section (2001)
- Input key code
- Description of hamlet reject
- Description of street reject

Example of a geocoded address:

|0|VIA GIUSEPPE VERDI, 24 |CANTAGALLO |PO |aoo19_2010| |59025 |PO |CANTAGALLO | |VIA |GIUSEPPE VERDI |24 |09|100|001|0000000|0000000|0000000|59025 | | | | |1668871|4877158|0000001|0668811|4877101|AOOCCAN | | |

Geocoding results:

Table 2: JC geocoding results

Geocoding results	v.a.	%
Recognized address	195.342	90,4
Ambiguous address	1.884	0,9
Discarded address	18.811	8,7
Address withoud street name	147	0,1
Totale	**216.184**	**100,0**

Processing detail:

```
ESITO ELABORAZIONE : OPERAZIONE CONCLUSA CORRETTAMENTE
 NOME PROGETTO              : SAMPLE3_20100721144218
 FILE DI INPUT              : tbIndirizziGeo.txt
 RECORD TOTALI              : 000216185
 RICONOSCIUTI               : 000195206
 % RICONOSCIUTI             : 090.29
 GEOREFERENZIATI SUL CIVICO : 000187962
 % GEOREFERENZIATI          : 096.28
 GEOCODIFICATI ISTAT        : 000194429
 % GEOCODIFICATI ISTAT      : 099.60
 SCARTATI                   : 000020979
 % SCARTATI                 : 009.70
 DATA                       : 2010/07/21
 ORA INIZIO ELABORAZIONE    : 14:42:19
 ORA FINE ELABORAZIONE      : 14:50:35
```

2.1.2 Overall reliability

The overall quality of the dataset is very good for the main demographic variables (sex, date and place of birth,) and all the field concerning the job status and the job position, somewhat poor for some other variables (marriage status) very poor for other variables (education attainment, income, wage). Missing value for the main variables of the JC dataset are shown in the table below.

Table 3: Missing value for the main variables in the JC anagrafic dataset

Variable	Missing values	%
Sex	2	0
Date of birth	2	0
Place of birth	2	0
Municipality of residence	0	0
Municipality of domicile	0	0
Citizenship	100	0
Job status	357	0
Address of residence	16.241	7
Address of domicile	9.914	4
Marriage status	37.452	16
Education attainment	47.502	21
Wage	59.492	26
Income	212.898	92

For each relevant variable (sex, marriage status, educational attainment, citizenship, job status, wage, income, family burden) we also deployed an explorative statistical analysis to identify keying and coding errors and outliers, correcting the input errors where possible or replacing them with missing value.

2.2 The Revenue Agency database

2.2.1 Dataset optimisation

RA database contains declarations about personal revenue of each resident.

The reference population of RA 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 is not included.

RA contains declarations with the following fiscal forms:

- *Modello Unico Persone fisiche*
- *Modello 730*

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

The forms in *italic* refer to natural persons, the others to legal entities.

After a long negotiation with the IRA we obtained the following dataset:

- prpisa_a2007_730.txt – fiscal forms "Modello 730" (M730) presented by employee and pensioners;
- prpisa_a2007_upf.txt – fiscal forms "Modello Unico Persone Fisiche" (MUPF) presented by persons who receive compensation for land, buildings, participation, employment, self-employment occasional or continuous, business and pension and other income.
- prpisa_a2007_u50.txt, prpisa_a2007_u60.txt, prpisa_a2007_u61.txt – fiscal forms "Modello Unico Società di persone", presented by companies.

The datasets contain the declarations presented in 2008 referring to 2007 incomes.

Only the first two datasets are interesting for our purposes. Unfortunately, we could not obtain another important dataset from IRA, the "Modello 770" (M770), including the personal incomes of people who have not presented any declaration, but whose taxes have been payed by the employers ("sostituti d'imposta") [4].

The two datasets of our interest contains the following information:

- Personal anagrafic data;
- Income declaration data.

The total record number in the datasets is 181.675 (respectively 81.350 records for MUFP and 100.325 for M730). In the same year there were 236.791 taxpayers in the Pisa province[5]. Therefore, we could estimate the persons who presented only M770 as about 55.000 units. The obtained dataset covers about 77% of persons with a fiscal declaration in 2007. Referring to the eligible population (the whole resident population) our dataset does not include:

[4] As documented in various progress reports we encountered many difficulties in obtaining RA data from IRA, mainly depending on privacy concerns, but also depending on the lack of a clear institutional framework ruling the relations between IRA and Province of Pisa. The provided data, therefore, have less fields (variables) than expected. The main limitation is the lacking of an identification field permitting us to find duplicates and to link directly RA data with EUSILC dataset and JC dataset.

[5] Official data available from Minister of Finance
http://www.finanze.gov.it/dipartimentopolitichefiscali/fiscalitalocale/distribuz_addirpef/lista.htm

- The persons without a declared income (including also tax dodgers);
- The persons having income that presented only M770;

We cannot determine precisely the coverage quota of the RA database with reference to the population of the province of Pisa. Local researches[6] have determined that RA data cover about 95% of resident eligible population.

The fiscal data have been then processed to calculate the following variables:

- Total taxable income
- income from land (land and buildings)
- employed income
- self employment income
- capital income
- income from commercial activities and businesses
- other income

Geocoding: As described in the previous paragraph (see 2.1.1), we needed to geocode the RA dataset to assign to each record the proper census enumeration area (CEA). The following table shows the results of the geocoding process for MUPF and M730 datasets.

Table 4: RA geocoding results

Geocoding results	**MUPF**		**M730**	
	v.a.	**%**	**v.a.**	**%**
Recognized address	73.273	90,1	89.799	89,5
Ambiguous address	686	0,8	1.051	1,0
Discarded address	7.387	9,1	9.462	9,4
Address without street name	4	0,0	13	0,0
Totale	**81.350**	**100,0**	**100.325**	**100,0**

2.2.2 Overall reliability

The RA datasets contain good quality data. Missing cases and inconsistencies in the data are avoided by the input process of the tax returns[7]. In order to assess the overall reliability of RA dataset an explorative statistical analysis was developed to detect any left mistake.

[6] 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 Caritas database

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 CARITAS Network. This network, created in 2003, has at first designed an unique database, which contains the materials collected in all Caritas' counselling centres.

By now, the CARITAS 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.

The eligible population is not "a priori" determined. Each Caritas' counselling centres is opened to receive every person asking for their help. The Local Diocesis guidelines suggest to counselling centres to follow a kind of territorial competence; this means that if an individual applies 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:

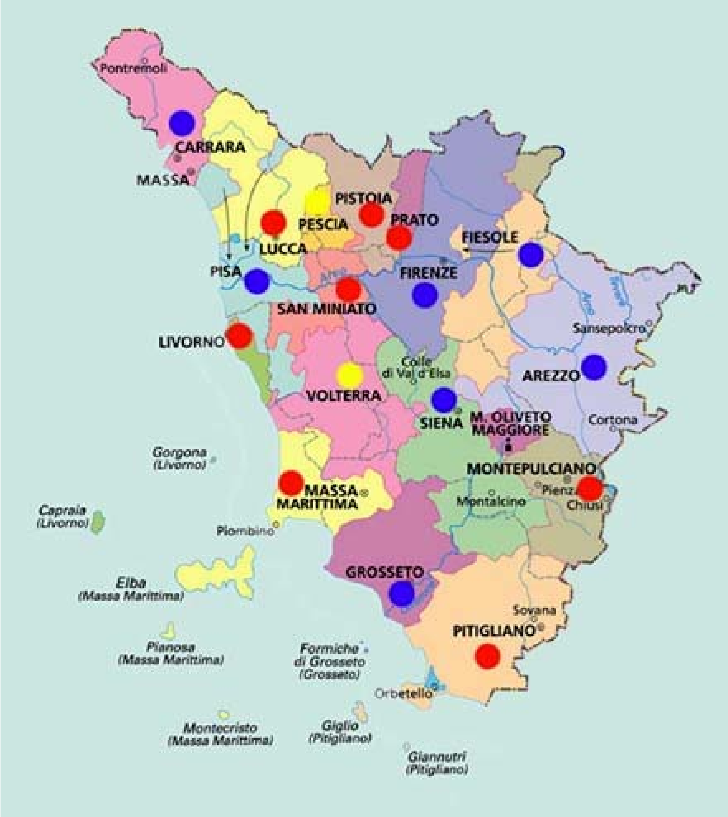
- Residents in the Pisa province;
- Residents outside the Pisa province;
- Homeless.

The Pisa province area is not entirely covered by the CARITAS network: firstly, because the Caritas counselling centres have a fragmented diffusion on the territory; secondly, because not all counselling centres use the CARITAS software. These are the areas actually covered in the three dioceses of the Pisa province:

[7] Since 2005 taxpayers must submit the tax return exclusively via the internet, directly or through a recognized intermediary. The software checks for any inconsistency in the data and it allows the tax returbntrasmission only if every needed field is completely filled.

- 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;

Figure 2: Maps of Diocesis in Tuscany region



We have received from the three Diocesis in the Pisa province the 6 files extracted in excel format from CARITAS database containing two type of data:

- anagraphic register of the persons who have addressed to one of the Caritas Counselling centres localized in the area of each Diocesis (see the maps above)
- services demanded/supplied by/to the persons who have addressed to one of the Caritas Counselling centres localized in each Diocesis.

The anagraphic register records all the persons who applied to the Counselling Centre since CARITAS system was activated; the services register contains only the services demanded/supplied in the reference year (in this case 2008). The two tables are linked by a personal ID (the personal card number).

Table 5: CARITAS: tables in the provided dataset

Table	Description	Type	N. records
pisa 2008 anag-bis.xls	anagraphic register - Diocesi of Pisa	Stock	1.718
pisa 2008 ric.xls	services demanded/supplied - Diocesi of Pisa	Flow	908
san miniato 2008 anag-bis.xls	anagraphic register - Diocesi of San Miniato	Stock	450
san miniato 2008 ric.xls.xls	services demanded/supplied - Diocesi of San Miniato	Flow	403
volterra 2008 anag-bis.xls	anagraphic register - Diocesi of Volterra	Stock	41
volterra 2008 ric.xls.xls	services demanded/supplied - Diocesi of Volterra	Flow	64

Since 2005 the Caritas counselling centres of Pisa province have registered 2.208 accesses, 85% is of people resident in the area (about 1.900 persons)[8]. In 2008, 1.458 persons have requested some services to a Caritas Counselling Centre.

2.3.1 Overall reliability

The CARITAS dataset quality is weak. Data input is made by volunteers whose main task is to give assistance and help to persons in very critical situations, having great needs. Their concern for personal forms filling and for data input is not so high. Caritas in recent years has tried to improve the reliability of the CARITAS observatory system, but the provided data lack of many of the features and the requirement needed to transform these administrative private datasets in a statistical source:

- The territorial coverage is not complete: only some (the most part) of the Counselling Centres inputs data in the CARITAS system.
- The case and time coverage is not regular and continuous: not every operators compiles on a regular basis the forms and input them in CARITAS system;
- The criteria and the rules to follow in the forms compilation, in data input and in information classifications are not strict and codified;
- There are a high quota of missing values and input errors.

This doesn't mean that CARITAS data couldn't provide valuable information and also statistical indicators giving precious and unique insights into the phenomena of poverty at local level. But there should be a greater commitment in improving the monitoring system in a statistical sense.

2.3.2 Dataset optimisation

The poor quality of the provided data leave very few possibilities of optimisation and standardisation. We try anyway to normalize the address of residence and to fill the missing values

[8] This statistics is calculated only on records with no missing address of residence.

in order to improve the results of the georeferencing process. The following table shows the results of geocoding process for CARITAS anagrafic dataset.

Table 6: CARITAS geocoding results

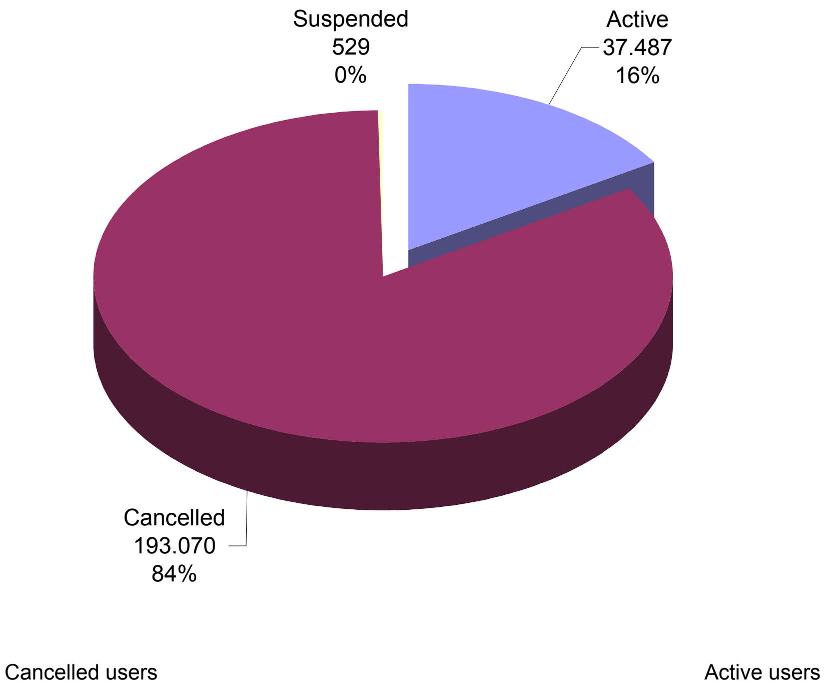
Geocoding results	CARITAS	
	v.a.	%
Records without address data	480	24,8
Recognized address	1.159	60,0
Ambiguous address	0	0,0
Discarded address	293	15,2
Address without street name	0	0,0
Totale	**1.932**	**100,0**

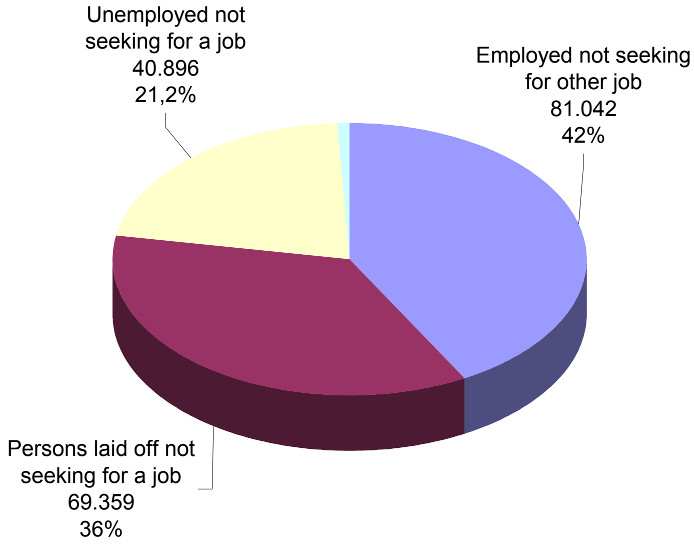
3 Studying poverty with the administrative data (SR/DSMAE)

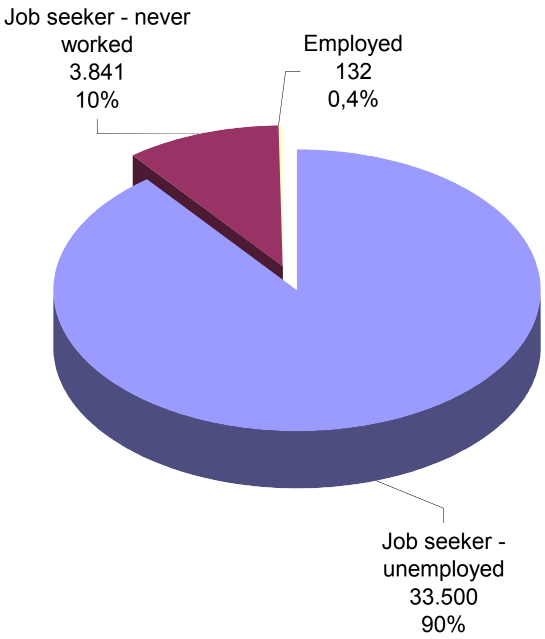
3.1 The JC data

The JC database contains 231.443 records. Only 37.487 (16%) are current job seekers (registered as "active users"). The most part are persons defined by the Job Center as "cancelled" (193.070).

Figure 1: JC: persons in the register by type of status







The “cancelled” belong to 3 main categories:

- (1) persons (21,2%) who have been enrolled in the past as a job seeker and who currently are not seeking for a job (because they just found one or because they become inactives);
- (2) persons who have been registered because their company communicated to the Job Center their hiring (42%) or (3) their layoff (36%);

Job seekers are mostly (90%) people who has been enrolled in the Job Center Register after the layoff from a previous job.

It is interesting to compare JC data and the Istat Labour Force Survey (LFS) results for the province of Pisa (2008 average data). First of all, we must consider that the ILSF is held on a sample that, for the province of Pisa, in the whole 2008 concerned about 1.500 individuals; JC datasets is an administrative source covering in principle the whole labour force. Secondly, we must take into account that the JC dataset can include also users resident outside Pisa province (3,8% of total registered).

Table 7: ILFS 2008 and JC: a comparison between data 2008 for province of Pisa

ILFS 2008			JC 2008	Total registered	Only residents in Pisa	
					Total	Only recent users
Population >15yo	**353.330**		**Total registered in Job Centers db**	**231.086**		
			Cancelled	*193.070*	*186.382*	*111.394*
Non labour force (inactive)	165.367		Unemployed not seeking for a job	40.896	39.089	9.328
Employed - self employed	48.063		Persons laid off not seeking for a job	69.359	67.060	27.195
			Other	1.773	1.706	663
Employed - employee	131.345	⇔	Employed - employee	81.042	78.527	42.088
Employed - total	179.407		*Active users ("conservato")*			
Unemployed	8.556	⇔	Job seekers	37.341	35.369	24.408
Labour force - total	187.963		Employed seeking for a job	132	129	119

Moreover, the definitions adopted in the ILFS are not coherent with the variables registered in the JC dataset. In particular, the criteria to identify the "unemployed" used in the ILFS are more restrictive than the ones used by the Job Center, even if the definitions of unemployed are very similar[9]. Therefore, as we can note in Table , the total number of job seeker registered in the Pisa Job Centers is more than four times higher than the number of unemployed estimated by ILSF (the yearly mean value). If we restrict the comparison only with job seekers residents in the Pisa province and who have updated their status in 2008 the number fall to 24.408, still three times higher than the ILFS estimate. This overestimation is well know[10]: in Tuscany in 2008 the unemployed registered by official statistics (ILFS) is on the average 30% of total registered at provincial Job Center. We must take into account also that about 25%-30% of unemployed registered by RCFL is not registered at Job Center.

Concerning the employed (as employee) the Job Center data underestimate the real number as estimated by RCFL (131.345 vs 78.527, i.e. 60,5%).

The JC data allows to calculate only proxy indicators about Labour Market.

[9] Job Centers, according to national law (DLGS 297/2002) and regional regulation, consider "unemployed" all the registered persons who: 1) are without a job; 2) declare to the job center of residence that they are available to accept a job; 3) participate to activities promoted by the Job Centre to improve their employability (training). This three conditions needs to be self-certified by the unemployed every year.

[10] Anastasia B., Gambuzza M. e Rasera M., La disoccupazione "amministrativa": un'approssimazione (o una finzione) irrinunciabile?, Veneto Lavoro, 2000 (report available on www.venetolavoro.it).

3.2 The RA data

This section focuses on the analysis of the RA data concerning people resident in the Province of Pisa.

According to the RA database, in 2008, 186.940 persons submitted a tax declaration, either through the M730 or the MUPF fiscal forms. As explained in previous sections the M730 form is submitted by employees and social benefits perceivers whereas the MUPF form concerns mainly self employed and capital income. It is worth stressing that M730 tax forms account for about 60% of total declarations.

The majority of taxpayers are male (54%) . Table xx presents their distribution by class of age and type of fiscal form.

Table 8: Taxpayers (% frequencies) by class of age and type of tax form

Class of Age	M730	MUPF	Total
< 21	0.27%	0.60%	0.01%
21-30	7.13%	9.18%	3.91%
31-40	20.96%	21.39%	26.80%
41-50	20.74%	22.15%	27.46%
51-60	18.02%	19.10%	20.58%
61-70	16.39%	14.37%	14.08%
71-80	11.13%	8.50%	5.65%
>80	5.36%	4.71%	1.51%
Total	100.00%	100.00%	100.00%

Unfortunately the SAMPLE consortium could access only part of the RA database. On the one side, it was not possible to access part of the declarations, moreover only a small set of economic variables were actually given. One of the main negative consequence is the lack of any insight in the pensioners economic condition. Moreover it is possible to analyse gross income only, given the lack of data on paid taxes.

Table 9 presents the main descriptive statistics for the main income categories.

Table 9: Descriptive statistics for the main income typologies

	Employees and pensioners income	self employed income	property income	Total taxable income
Mean	21591	24125	20422	22899
Standard deviation	19580	38947	35722	26114
25% quantile	12359	8851	5776	12022

50% quintile	18464	16037	13343	18508
75% quintile	25556	26803	23291	26519

.

Finally, the Gini index and the Income quantile share ratio (S80/S20) are measured for each income typology (see tab. 10)

Table 10: Gini index and quantile share ratio for income typologies

	employee income and social benefits	self employed income	property income	taxable income
Gini index	0.35	0.51	0.55	0.39
S80/S20 ratio	8.62	6.43	18.89	34.18

As expected variability and concentration are larger for self-employed and property income.

3.3 Caritas

The Caritas dataset allows to investigate a segment of the poor population, which can be hardly detected by official statistics. According to Istat [11], people in absolute poverty accounted for about 2,9% of people living in the Centre of Italy in 2008. Anyhow such estimate is obtained taking into account data from the Household budget survey i.e. interviewing people with a stable housing. On the contrary, Caritas data concern people asking for the Caritas services, be they resident or homeless people. As a consequence part of poor people captured by the Caritas dataset is not covered by the Istat absolute poverty index.

Unfortunately, the weak quality of the data (see § 2.3) affects the statistical analysis. Moreover the accessed dataset contains only part of the actually collected data.

People seeking assistance in the Caritas centres are typically foreign (about 70%) and female people. They are 42 years old on average, with foreigners younger than Italians (respectively 30 and 57 years old on average). Women live generally in a household (with relatives or friends) whereas a significant part of male people (23%) live on their own. Most of people, especially if stranger, asks for goods and services(food, cloth, furniture, taking a bath etc.). Italians more frequently seek an economic support.

[11] As defined by Istat (see: Istat Metodi e Norme, "La misura della povertà assoluta" del 22 Aprile 2009, http://www.istat.it/dati/catalogo/20090422_00/)

4 The integration of PI-SILC and the administrative datasets (DSMAE)

The JC and RA data sources seem to cover populations similar to the Silc population. As a consequence we expect to find some of the Pi-Silc sampled individuals in the administrative archives. The exact matching (or record linkage) is the technique used to identify and pick up such units.

Record linkage is a technique which compares records contained in two files *A* and *B*, in order to determine pairs of records referred to the same population unit. 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 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. This procedure is known as exact matching.

Unfortunately, some 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*.

Formalizing the linking procedure into a statistical model, it is possible to evaluate the matching by measuring the probability of generating false-matched-pairs and false-unmatched pairs. This procedure is known as probabilistic record linkage.

Coming to our application, we can state the problem as follows. The Pisa-Silc contains *N* records, one for each of the *N* interviewed individuals. On the other hand the RA (or JC) archive contains *M* records one for each registered subject. Given a set of common variables (*k*-variables) we have to evaluate the evidence that the *i*-th record from Silc and the *j*-th record from RA (or JC) relate to the same person.

Table 1 shows the personal items eligible to be used as *k*-variables in the three data sources. We dispose of a similar set of variables for PI-Silc and JC data. On the contrary, we cannot access the birth month in the RA data source.

Formalizing the linking procedure into a statistical model, it is possible to evaluate the matching by measuring the probability of generating false-matched-pairs and false-unmatched pairs (Fellegi and Sunter 1969).

Table 11: Personal items to be used as k-variables, in the PI-Silc, RA and JC data sources

Personal item	Data source		
	PI-Silc	RA	JC
Birth day			✓
Birth month	✓		✓
Birth year	✓	✓	✓
Gender	✓	✓	✓
Place of birth (municipality)	✓	✓	✓
Place of residence (municipality)	✓	✓	✓
Place of residence (census enumeration area)	✓	✓	✓
Nationality	✓	✓	✓

4.1 Integrating Silc with the local Revenue Agency archive

The PI-Silc and RA integration is particularly difficult because of the lack of identifiers on the RA side (see Tab. 1). Before describing the procedure, it is worth mentioning that revenue database consisted in data from the 730 and the Unico p.f. tax returns registers (see § 2). The first register (M730) contains data on employees and pensioners income, whereas the second (MUPF) collects mainly information on self-employed people.

- Step 1: The matching of 730 register and PI-Silc

Available data allow to build a match key made of the following personal items: “gender”, “birth year”, “birthplace”, “place of residence” and “nationality”. As a first step the linking procedure is run on the M730 taxpayers with a joint tax declaration[12]. Thus a 730 unit is linked to a PI-Silc unit if both the following conditions are met: i) the compared units share exactly the same match-key value ii) the compared units belong to the same family (i.e, the 730 couple records the same mach-key values as the corresponding PI-Silc couple). As a second step the linking procedure is run on the single declarations be they from the M730 register or the MUPF register. In the end, 411 Pi-Silc units find at leas one link with the fiscal records.

The matched data set contains the Silc variables as well as data coming from the RA database (730 and Unico p.f registers). Furthermore a linking probability is provided which helps evaluating the quality of the matching. This probability is an estimate of the probability that the coupled units represent a true link (the two record refer to the same unit) giving the observed values for the matching variables. Note that these probabilities are estimated by using the EM algorithm under the assumption of conditional independence. In practice, it is considered the vector $\gamma_{a,b}$ resulting by the comparison of the observed values for the matching variables on record *a* and record *b*; $\gamma_{a,b}^{(j)}=1$ if *a* and *b* show the same value for the variable *j* and 0 otherwise. The conditional independence assumption means that $\gamma_{a,b}^{(j)}$ is independent from $\gamma_{a,b}^{(k)}$ given that *a* and *b* are a true link. The values of the variables are being compared by using the method proposed by Jaro and Winkler (cf. Winkler, 1988). The probabilities are estimated by using the code made available in the package "RecordLinkage" (Borg and Sariyar, 2010) freely available for the R environment (R Development Core Team, 2010).

4.2 *Integrating PI-Silc with the local Job Centre archive*

The matching procedure is split in two separate procedures. As a first step a deterministic matching is applied for those units in database which present information concerning the census Enumeration Area (EA) in which they live; when this information is missing a probabilistic record linkage is applied. Note that the information concerning the census EA was not available for all the responding units at the Eusilc survey.

The exact matching procedure is straightforward. For each unit in the Eusilc survey is linked with the corresponding unit in the labour market database sharing the same information as far as the following variables are concerned: municipality, census EA, gender, birth month, birth year and a variable summarizing information about the birth place (country and NUTS3). Due to computational constraints the variable concerning the living municipality is used as a blocking variable (search is restricted to units living in the same Municipality). The units not linked in the

[12] Italian tax system allows couples to fill tax returns jointly, in this case the tax information of each member is reported on a distinct record.

exact matching phase are processed and linked in the record linkage phase. In practice units sharing the same values for the subset of the matching variables that not present missing values. In particular, the variables municipality, gender, birth month and year were available for all the units while some units had information missing for the census EA or for the birth place. In this step, again the Municipality is used as a blocking variable.

The first matching step based on exact linkage allowed to identify 529 couples of units corresponding. The second integration step allowed to find some other 404 couples of units. Note that for each linked couple of units it is estimated a probability of the linking quality. This probability is estimated with the same procedure used for the integration of PI-Silc with the Revenue agency archive.

4.3 Indicators from the matched data set

4.3.1 Indirect sampling

Indirect sampling starts from a sample s of n^s units obtained with a probability sampling design, it enlarges the initial sample using a link matrix and it associates estimation weights to the final sample units using the initial sampling weights and the link matrix. The sampling strategy is based on the Generalized Weight Share Method (GWSM – Lavallée, 1995; 2002; Deville and Lavallée, 2006) that can be viewed as a generalization of Network Sampling and also of Adaptive Cluster Sampling (Thompson, 1992; Thompson and Seber, 1996).

The objective is the estimation of a set of parameters of the RA and JC population using the PI-Silc sampling weights and the link matrix generated by the matching procedure (see § 4).

The set of individuals selected and surveyed by PI-Silc is considered a random sample s of size n^s. For each unit in the sample the set of units is defined on the result of the record linkage process between the Pi-Silc data and the JC (RA) datasets. Let indicate this additional set with c of size n^c. In such way the initial sample of PI-Silc units is enlarged adding the linked units in the JC (RA) data.

The correspondence between the sampled unit and the unit in the administrative data set can be represented by a link matrix $\mathbf{W}_{sc} = \left[w_{ji}^{sc}\right]$ of size $n^s \times n^c$ where each element the linkage weight $w_{ji}^{sc} \geq 0$. That is, unit j of s is related to unit i of c provided that $w_{ji}^{sc} > 0$, otherwise the two units are not related to each other. We assume that for any unit j the values of the link matrix can be obtained. Notice that in this application only one unit j has been finally selected for each sampled unit i. As a consequence the size of the target populations (JC/RA) is the same as the size of Pi-Silc sample.

The link matrix can be also defined in standardized form $\overline{\mathbf{W}}_{sc} = \mathbf{W}_{sc}\left[diag(\mathbf{1}_c^T\mathbf{W}_{sc})\right]^{-1}$ 13.

In this application we can not standardize the matrix because only one unit *j* has been finally selected for each sampled unit *i*.

Let π_j^s be the selection probability of unit j. We assume $\pi_j^s > 0$ for all j selected in s, we identify the units i of c that have a non-zero linkage weights correspondence. For each unit i of the set c, we measure a variable of interest y_i for each unit of the network. Let $\mathbf{Y}^c = \{y_1^c, y_2^c, ... y_{n^c}^c\}^T$ be the column vector of that variable of interest. The parameter is the total Y^c of the target population C where $Y^c = \sum_{i=1}^{N_c} y_i$.

Let $\check{\mathbf{s}}^s = \{\pi_1^s, \pi_2^s, ..., \pi_{n^s}^s\}$ and let $\mathbf{D}_s = diag(\check{\mathbf{s}}^s)$ be the diagonal matrix of size $n^s \times n^s$ containing the selection probabilities used for the selection sample s. Then we can directly form the unbiased Horvitz-Thompson estimator:

$$\hat{Y}^c = \mathbf{1}_s^T\mathbf{D}_s^{-1}\mathbf{W}_{sc}\mathbf{Y}^c \qquad (1.1)$$

Let us define the column vector $\mathbf{Z} = \mathbf{W}_{sc}\mathbf{Y}^c$ of size n^s. The variance of $\hat{Y}^c$ is directly obtained as the variance of the Horvitz-Thompson estimator (Särndal et al., 1992):

$$Var(\hat{Y}^c) = \mathbf{Z}^\mathbf{T}\Delta_s\mathbf{Z} = \mathbf{Y}^{c^T}\Delta_c\mathbf{Y}^c \qquad (1.2)$$

where $\Delta_s = \left[\left(\pi_{jj^T}^s - \pi_j^s\pi_{j^T}^s\right)/\pi_j^s\pi_{j^T}^s\right]$ is a non-negative definite matrix of size $n^s \times n^s$ and $\pi_{jj^T}^s$ is the joint selection probability of units j and jT from s, with $\Delta_c = \mathbf{W}_{sc}^T\Delta_s\mathbf{W}_{sc}$.

4.3.2 The application

The objective is the estimation of taxable average income from the RA population using the PI-Silc sampling weights and the link matrix generated by the matching procedure (see § 4).

[13] Note that in order for the matrix $\overline{\mathbf{W}}_{sc}$ to be well defined, we must have $\left[diag(\mathbf{1}_c^T\mathbf{W}_{sc})\right]^{-1}$ to exist, which is the case if and only if $wn_{+i}^{sc} > 0$ for all $i = 1...n^c$

Results are given in tab. Xx. The first column contains the unweighted mean and quantiles calculated on the RA-PISILC taxable income variable. The second column contains the same indicators estimated using the indirect sampling methodology.

It is worth stressing the general increase of average income due to the weighting procedure.

Table 12: Taxable income: comparison of weighted and unweighted indicators

	Taxable income (unweighted)	Taxable income (weighted*)
mean	19878	21561
25%	12138	13369
50%	17826	19774
75%	25233	27295

* weights are calculated following the indirect sampling methodology (see § 4.3.1)

5 Conclusions

The use of administrative data is necessary for local government in that official statistics only seldom provide accurate information at the local level. Moreover administrative data source provide a finer detail on issue only partially covered by official statistics surveys

This deliverable is mainly devoted to the integration of survey and administrative data sources. In particular we have tried to build a matched database having PI-Silc as the core dataset and the linked administrative data sets as satellites for in depth analysis on specific aspects (labour, income, taxes).

The Silc and administrative data sets have been integrated using a record linkage procedure. Finally, only a limited subset of records is successfully linked. This not entirely satisfying result is due to several reasons which can be summarized as follows:

- The Silc oversampling does not include all municipalities of the Province of Pisa (only 25 municipalities out of 39 have been involved);
- The administrative data sources cover only sub-populations of Silc (under coverage) i.e. part of the taxpayers (RA register) and people asking for the local job center services (JC register)
- There are errors and missing values in the variables used as identifiers; for example the JC register seems to contain outdated information on addresses, which lead to wrong census enumeration area.

- Data sources (RA in particular) contain only “weak” identifiers.

The linkage results have been used to correct the RA estimate of taxable income for the self selection bias applying the indirect sampling methodology.

Bibliography

Deville, J.C., Lavallée, P. (2006): Indirect Sampling: The Foundations of the Generalized Weight Share Method, *Survey Methodology, 32, 1*, pp.165-176.

Lavallée, P. (1995): Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method, *Survey Methodology, 21, 1*, pp.25-32.

Lavallée, P. (2002): Le Sondage Indirect, ou la Méthode généralisée du partage des poids, *Eds. Université de Bruxelles*, Bruxelles.

Särndal, C.E., Swensson, B., Wretman, J. (1992): *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Thompson, S.K. (1992): *Sampling*, John Wiley and Sons, New York.

Thompson, S.K., Seber, G.A. (1996): *Adaptive Sampling*, John Wiley and Sons, New York.