SAMPLE DELIVERABLE 7.1.

OVERSAMPLING DESCRIPTION-PART I

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.:	D 7.1.
Deliverable name:	Oversampling Description-Part I
WP n.:	1.2.
Lead beneficiary:	1
Nature:	Report
Dissemination level:	PU
Status:	Final
Due delivery date from Annex I:	July 31th , 2009
Actual delivery date:	November 20th, 2009
Project co-ordinator name:	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

OVERSAMPLING DESCRIPTION-PART I

Project Acronym:	SAMPLE
Project Full title:	Small Area Methods for Poverty and Living Conditions Estimates
Grant agreement No:	SSH - CT - 2007 – 217565
WP no.:	1.2.
Deliverable n.	D7.1.
Document title:	Oversampling Description-Part I
Status:	Final
Editors:	Monica Pratesi, Alessandra Coli
Authors:	Monica Pratesi (m.pratesi@ec.unipi.it), UNIPI - DSMAE Alessandra Coli (a.coli@ec.unipi.it), UNIPI-DSMAE Michela Casarosa (m.casarosa@provincia.pisa.it), PP-UROPS
Document editing:	Marta Garro
Language revision:	Marta Garro

Summary

1. Introduction

The EU SILC wave 2008 oversampling for the Province of Pisa was scheduled for last fall. The data collection operations lasted till the end of October 2008. However, the data are not yet available, since the Italian Institue of Statistics (Istat) needs one year for checking the quality of the collected data (see Istat - Metodi e Norme, n. 38, 2009). The check concerns editing and imputation, and consistency evaluation of the collected data in comparison with the administrative data files maintained by the Central Government Agencies.
We foresee to have the final data set by the end of 2009. This is the reason why the Deliverable 7 (D7)- *Oversampling description* has been divided into two separated and successive reports.

The first report (D7.1. *Oversampling Description* - Part I) is structured into three parts:

- The EU SILC oversampling for the Province of Pisa: the state-of-the-art of EU SILC oversampling for the Province of Pisa. The process is still in progress, final microdata should be released by Istat in December 2009.
- Institutional contacts: a synthetic overview of the local contacts led by the Province of Pisa in order to gain access to administrative data.
- Methodology for the combined estimation at LAU2 and LAU1 levels.

The second and final report (D7.2. *Oversampling Description* - Part II) will integrate this first release describing the results of the multidimensional analysis of poverty, vulnerability and deprivation in the Province of Pisa with a first sub-provincial comparative analysis.

It is important to underline that in this report we do not refer to the methodology of combination and/or integration of EU SILC data with administrative data in order to improve the measurement of poverty. This will be the objective of Deliverable 9.
Here we refer to the methods of estimation of poverty indicators based on the EU SILC oversampling data and auxiliary information available at area level or at unit level. Recent developments in the field of sub-national poverty estimates make possible the use of Small Area Estimation (SAE) statistical methods, which are more sophisticated than the simple direct estimators. These methods are referred to as "combined" estimators. They are combined in the sense that they combine direct estimators with model based estimators. Small area combined estimators "borrow strength" from related areas using auxiliary information which is supposed to be correlated to the variable of interest (Rao, 2003).
We focus here on the direct estimation of the target quantities under the EU SILC sample design. This is the design-based part of the combined small area estimator (small area combined estimation). We refer to Deliverable 4 for a description of the models currently used to build the model-based part of the combined estimator.

2. The Eu-silc oversampling for the Province of Pisa[1]

The main data source used in SAMPLE for estimating poverty and social exclusion indicators is EU SILC.
For the 2008 wave, the Consortium commissioned to Istat an oversampling for the Province of Pisa. The purpose is threefold: i) getting direct estimates of poverty and social exclusion indicators for

[1] We thank Cristina Freguja, Andrea Cutillo and Sabina Giampaolo from Istat for their helpful advices and support.

the Province of Pisa; ii) improving the SAE methodology through the combination of LAU1 and LAU2 estimates and the use of local administrative information; iii) 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. Oversampling is fully integrated in the EU SILC standard procedure.
The EU SILC process is still in progress, and final microdata should be released by Istat in December 2009.
This work describes the steps already accomplished, i.e. the sample design and selection, the fieldwork for data collection and a first analysis of response rates. The correction procedure for non response (including the integration with registers) as well as the weighting procedure will be described in the updated version of the report, planned for February 2010.

2.1 Sample design and selection

EU SILC aims at providing comparable and timely cross-sectional and longitudinal data. Cross-sectional data focus on income, poverty, social exclusion and other living conditions whereas longitudinal data restrict to income, labour and a set of non-monetary indicators of social exclusion. For these *target primary areas*, data have to be collected yearly. *Target secondary areas* are investigated for the cross sectional component only, on a less-than-yearly frequency. For the year 2008, the secondary issue is over-indebtedness and financial exclusion.

Sample design: the "integrated approach"
Eurostat recommends a single sample design to fulfil both the cross-sectional and the longitudinal requirements; this model is called the *integrated approach*.
The design consists in selecting a fixed number of panels (subsamples or replications) at the first wave. The cross-sectional sample is composed by the sum of such subsamples. Each subsequent year a panel is dropped out and replaced by a new one. The greater the number of starting panels, the longer the duration of each panel.
Eurostat recommends four panels, which imply the minimum duration to fulfil the Commission Regulation requirement[2]. Italy, as most of the other countries, chose this sample design. This means that every year a fourth of the sample is renewed so that sampled individuals are traced on for a maximum of four-years.

Figure 1 – Illustration of the "integrated" design with four panels.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**…**
T-3	$A_{(4)}$	$B_{(3)}$	$C_{(2)}$	D_1					
T-2		$B_{(4)}$	$C_{(3)}$	D_2	E_1				
T-1			$C_{(4)}$	D_3	E_2	F_1			
T				D_4	E_3	F_2	G_1		
T+1					E_4	F_3	G_2	H_1	
T+2						F_4	G_3	H_2	…
T+3							G_4	H_3	…
…								…	…

Source: Istat (2008)

Figure 1 describes the rotational design in the case of four panels. For one year the cross sectional sample consists of four replications of the same dimension. Each panel remains in the survey for four years.
Let us describe the rotational process. At time T-3, the cross sectional sample is composed by the panels *A*, *B*, *C* and *D*. In order to start the panel rotation we behave as sample *A* were at its fourth and last wave, *B* at its third, *C* at its second and *D* at its first. The subscript stands for the wave

[2] The longitudinal component has to cover at least four years according to the EC Regulation, n. 1177/2003.

number; when in brackets, it indicates a fictitious wave. At time T-2, *A* is dropped and replaced by the new panel *D*, the subsequent year *B* is substituted with *F*, and so forth. Finally, at T+1, the sample is completely renewed.
As of time T, the system is fully established in that complete longitudinal samples are available. In this example, panel D is the first complete longitudinal sample.
The overlap between subsequent years is 75%. For example, between T and T+1 there are three out of four overlapping panels, i.e. *E*, *F* and *G*. The overlap falls short to 50% (panels *F* and *G*) and 25% (panel *G*) as the temporal lag increases to two (T+2) and three years (T+3).

The Italian sample design
Eurostat established the minimum actual sample size to be achieved for each Country on the basis of statistical and practical considerations. As a first step the Member States (MS) cross sectional sample sizes were determined in order to ensure a minimum precision for the most critical indicators, under the hypothesis of simple random sampling design: for Italy the required minimum size was of 5000 households for the cross-sectional sample and 4000 households for the longitudinal sample. As a second step, MS sample sizes were changed in order to reflect the relative weight of each country in terms of the number of residents. As a consequence, the Italian cross-sectional and longitudinal sample sizes increased respectively to 7500 and 5500 households.
Moving from these minimum sizes, individual countries were demanded to determine the effective sample dimensions taking into account the following aspects:

- the design and non response effects[3] on the precision of the estimates;
- the necessity to meet national requirements.

In Italy, the sample size was determined in order to fulfil both Eurostat requirements and a prefixed precision of the estimates at regional level (NUTS2). As a result, a starting cross-sectional sample of 32000 households was established, composed by four longitudinal subsamples of 8000 households each.
The four replications were drawn according to a stratified two-stage selection.
Municipalities (first stage units) were stratified by regional area (NUTS2) and demographic size for a total of 288 strata at national level.
For each region (NUTS2) three different kinds of strata were defined:

1. Self-Representative strata (SR). These strata include municipalities with a number of residents over an established threshold (larger municipalities). Each stratum includes one single Municipality so that in each region there are as many SR strata as the number of regional municipalities which exceed the threshold.
2. Non-Self-Representative strata of the first kind (NSR1). Each stratum includes few middle-size municipalities.
3. Non-Self-Representative strata of the second kind (NSR2). These strata include small-size municipalities.

At the first stage, a sample of municipalities is selected from the defined strata, according to the following pattern:

- all municipalities belonging to SR strata;
- two municipalities from the NSR1 strata, selected with probability proportional to the number of residents – such municipalities enter two panels each;
- four municipalities from the NSR2 strata, selected with probability proportional to the number of residents – each Municipality enters one panel only.

At the second stage, households are selected from the register office of each sampled Municipality following a systematic sampling.

[3] Including an evaluation of the attrition effect.

Every year one panel is dropped and replaced with a new one. It is worth stressing that the renewal concerns the second stage units only (households) since first stage units (municipalities) are established once and for all at the first wave. Particularly, municipalities update their samples partially or totally depending on the stratum they belong to: the SR municipalities substitute 25% of the sample every year, NSR1 municipalities substitute 50% of the sample every two years and NSR2 substitute 100% of the sample every four years.
Let us go back to Figure 1, in order to try to explain how the composition of panels change from one wave to the next. Let us consider time T, when the cross-sectional sample is composed by the longitudinal samples *D*, *E*, *F* and *G*, at their fourth, third, second and first wave respectively. According to the sampling design, each panel is composed of three subsets of households: a group of households is drawn from the SR municipalities, a group is selected from one of the two sampled NSR1 municipalities and a group of households is selected from one of the four NSR2 municipalities. At time T+1, with the dropping of group *D*, 25% of households drawn from SR municipalities exit from the cross sectional sample, as well as 50% of households drawn from the NSR1 Municipality[4] and 100% of households drawn from the NSR2 Municipality. The new replication *G* is obtained with the selection of new households from the same first stage units. As a consequence, the SR municipalities renew 25% of their samples, the NSR1 Municipality substitutes 50% of its sample and the NSR2 Municipality renews its sample completely.

The oversampling for the Province of Pisa
According to the standard sample design, the Province of Pisa sample was expected to be of 162 households in 2008, with the composition described in Table 1.
The sample of Pisa (SR Municipality) was expected to be composed by 53 households, 16 of which coming from previous waves. San Giuliano Terme (NRS1) was expected to interview 37 households, all selected in 2007. Lari (NSR1 Municipality) would be in charge of a sample of 22 households coming from previous waves, whereas Calcinaia and Pomarance (both NSR1 Municipality) would interview respectively 23 and 27 new households.

Table 1 – Number of sampled households by Municipality in the Province of Pisa, year 2008, before oversampling

First stage units (municipalities)	Stratum typology	Number of new households selected in 2008	Number of households already in the sample	Number of households to be interviewed in 2008
Pisa	SR	16	37	53
San Giuliano Terme	NRS1		37	37
Pomarance	NRS2	27		27
Lari	NRS2		22	22
Calcinaia	NRS2	23		23
Province of Pisa		66	96	162

Data source: Istat

The oversampling has increased the Pisa Province sample size to 818 households with an impact on the selection of both the first stage and second stage units. Following the sample design, a stratified two-stage selection has been performed.
The first stage units increased from five (see Table 2) to 25 municipalities.

[4] The remaining 50% belongs to another panel which continues to be part of the cross sectional sample even at time T=T+1

For the selection of the new 20 municipalities, an *ad hoc* stratification was applied. In fact, according to the Consortium requirements[5], municipalities were grouped by relevant Health Society (HS), which are public non profit institutions in charge of the planning of social and health assistance at local level.
Table 2 shows the number of sampled households by Municipality and Health Society.
The HS "Bassa Val di Cecina" and "Valdarno Inferiore" were considered as Self Representative strata; as a consequence all the municipalities belonging to these HS were considered as first stage units. "Area pisana", "Val d'Era" and "Alta Val di Cecina" were considered as Non-Self-Representative strata. For these strata municipalities were selected with probability proportional to demographical size with the exception of the five municipalities already belonging to the standard sample of the Province of Pisa (see Table 1).
Summing up, in the Province of Pisa 25 municipalities have been involved in EU SILC survey for the year 2008, with 20 municipalities taking part in the survey for the first time. For what concerns the five municipalities selected in previous waves, three have increased the size of their samples whereas two have kept their standard samples.
Finally, it is worth stressing that the Province of Pisa oversampling affects the 2008 cross-sectional sample only, because the new sampled individuals (even when eligible) are not going to be traced in next waves.
The selection of the samples took place from July to September 2008.

Table 2 – Number of sampled households by Municipality and HS – Province of Pisa, year 2008, after the oversampling (to be continued)

Health Society	Municipality	Sample size
Bassa Val di Cecina	Castellina M.ma	34
Bassa Val di Cecina	Guardistallo	33
Bassa Val di Cecina	Montescudaio	33
Bassa Val di Cecina	Riparbella	33
Valdarno inferiore	Castelfranco di S.	38
Valdarno inferiore	Montopoli V. A.	38
Valdarno inferiore	San Miniato	38
Valdarno inferiore	S. Croce sull'Arno	38
Alta Val di Cecina	Castelnuovo VdC	33
Alta Val di Cecina	Montecatini VdC	33
Alta Val di Cecina	Pomarance	33
Alta Val di Cecina	Volterra	34
Val d'Era	Bientina	24
Val d'Era	Calcinaia	29
Val d'Era	Casciana Terme	29
Val d'Era	Lari	30
Val d'Era	Ponsacco	29

[5] Health Societies are considered the reference units for monitoring health and social assistance provided by public and non profit institutions. For this reason the oversampling has been planned for obtaining direct estimates of poverty indicators at HS level.

Health Society	Municipality	Sample size
Val d'Era	Pontedera	29
Val d'Era	S. Maria a Monte	29
Area Pisana	Calci	28
Area Pisana	Cascina	28
Area Pisana	Pisa	53
Area Pisana	San Giuliano T.me	36
Area Pisana	Vecchiano	28
Area Pisana	Vicopisano	28
Bassa Val di Cecina	*Total*	*133*
Valdarno inferiore	*Total*	*152*
Alta Val di Cecina	*Total*	*133*
Val d'Era	*Total*	*199*
Area Pisana	*Total*	*201*
Province of Pisa	**Total**	**818**

Data source: Istat

2.2 The fieldwork

In September 2008 Istat organized a training meeting in Pisa for the interviewers and the Municipality staff in charge of the survey (supervisors). Previously, municipalities were contacted in order to be informed on the reasons and purposes of the oversampling.
Table 3 provides details on the training meeting attendance and on the level of know-how of the municipalities. We observe that 11 municipalities out of 25 declared to be unexperienced in managing sample surveys. Fortunately, the attendance to the training meeting was high, with a good level of participation, according to the Istat report.

Table 3 – Qualitative indicators on the know-how of the sampled municipalities (to be continued)

Municipality	Municipalities selected before 2008 (yes/no)	Municipalities with experience in managing sample surveys (yes/no)	Who attended the training meeting ?	Interviewer with experience (yes/no)
Bientina	No	No	Supervisor + interviewer	No
Calci	No	No	Supervisor + interviewer	No
Calcinaia	Yes	Yes	Supervisor + interviewer	Yes
Casciana terme	No	Yes	Supervisor + interviewer	Yes
Cascina	No	No	Supervisor + interviewer	Yes
Castelfranco di Sotto	No	Yes	Supervisor + interviewer	No
Castellina Marittima	No	Yes	Supervisor +interviewer	No
Castelnuovo di Val di Cecina	No	No	Interviewer	No
Guardistallo	No	No	Supervisor + interviewer	No
Lari	Yes	Yes	Supervisor + interviewer	Yes

Municipality	Municipalities selected before 2008 (yes/no)	Municipalities with experience in managing sample surveys (yes/no)	Who attended the training meeting ?	Interviewer with experience (yes/no)
Montecatini Val di Cecina	No	No	Supervisor + interviewer	No
Montescudaio	No	No	Supervisor + interviewer	No
Montopoli in Val d'Arno	No	No	Supervisor + interviewer	Yes
Pisa	Yes	Yes	Supervisor + interviewer	Yes
Pomarance	Yes	Yes	Interviewer	Yes
Ponsacco	No	No	Interviewer	Yes
Pontedera	No	Yes	Supervisor + interviewer	No
Riparbella	No	No	Supervisor + interviewer	No
San Giuliano Terme	Yes	Yes	Supervisor + interviewer	Yes
San Miniato	No	Yes	Interviewer	Yes
Santa Croce sull'Arno	No	Yes	Interviewer	No
Santa Maria a Monte	No	Yes	Interviewer	Yes
Vecchiano	No	No	Supervisor + interviewer	Yes
Vicopisano	No	No	Supervisor + interviewer	No
Volterra	No	Yes	-	Yes

Data source: Istat (2008)

Data collection took place from the end of September to half of November. During this period, supervisors were asked to communicate weekly, by web, both the number of interviewed and not interviewed households. On the basis of this preliminary information it emerged that around 7 % of households refused to participate to the survey.
Once the fieldwork operations were concluded, the survey material was returned to Istat.

The material consisted of the REG form, the FAM form and the IND form.

The interviewer fills in the **REG form** when she/he gets in touch with the family for the first time. Municipalities must return one REG form for each sampled family, independently of the contact outcome. In case of positive contact, the interviewer registers personal information on each member of the family such as gender, date of birth, occupational status, main source of income, etc. In case of negative outcome, the interviewer must record the reason of the missed interview.
The number of REG forms corresponds approximately to the sample dimension.

The **FAM form** is the household questionnaire. It collects data on lodging, housing costs and on the general economic status of the family. Municipalities must return to Istat one FAM form for each interviewed household.

The **IND form** is the Individual questionnaire. It must be compiled by members aged 15 or more at the moment of the interview. The questionnaire collects information mainly on education, health, status in employment, income and savings. In case of missed interview, the interviewer must specify the reason.

2.3 First Analysis of non-response rates

Istat carried out a first check on the forms and questionnaires in order to assess how many households and individuals were actually interviewed.

Table 4 – EU SILC Response and Non-response rates by Municipality, year 2008

Municipality	Number of REG forms	Number of FAM forms	Response rate = FAM/REG (%)	Non-response rate (%)
Bientina	24	23	95,8	4,2
Calci	28	21	75,0	25,0
Casciana Terme	29	29	100,0	0,0
Cascina	28	28	100,0	0,0
Castelfranco di Sotto	38	34	89,5	10,5
Castellina Marittima	33	26	78,8	21,2
Castelnuovo di Val di Cecina	33	30	90,9	9,1
Guardistallo	33	23	69,7	30,3
Montecatini Val di Cecina	33	27	81,8	18,2
Montescudaio	33	16	48,5	51,5
Montopoli in Val d'Arno	38	32	84,2	15,8
Ponsacco	29	16	55,2	44,8
Pontedera	29	25	86,2	13,8
Riparbella	33	32	97,0	3,0
San Miniato	38	35	92,1	7,9
Santa Croce sull'Arno	40	37	92,5	7,5
Santa Maria a Monte	29	27	93,1	6,9
Vecchiano	28	26	92,9	7,1
Vicopisano	28	21	75,0	25,0
Volterra	34	12	35,3	64,7
Municipalities sampled in 2008	**638**	**520**	**81.5**	**18.5**
Calcinaia	29 = 23+6	23	79.3	20.69
Lari	29 = 22+7	28	96.6	3.33
Pomarance	33 = 27+6	30	90.9	9.09
Municipalities present in previous waves (*with* oversampling)	**91**	**81**	**89.0**	**11.0**
Pisa	53	47	88.7	11.3
San Giuliano Terme	36	27	75.0	25.0
Municipalities present in previous waves (*without* oversampling)	**89**	**74**	**83.1**	**16.9**
Pisa province	**818**	**675**	**82.5**	**17.5**
Tuscany	**2441**	**2023**	**82.9**	**17.1**
Italy	**26042**	**21470**	**82.4**	**17.6**

Data source: Istat

Table 4 shows the number of received REG and FAM forms for each sampled Municipality, for the province of Pisa, for Tuscany and for Italy.
The response rate is calculated as the ratio between FAM and REG forms. The province of Pisa rate (82.5%) is approximately equal to the Tuscany (82.9) and Italy (82.4) rates. The non-response rate is higher for municipalities involved in EU SILC for the first time (18.5% against an average of 17.5%).

Table 5 shows the number of compiled IND forms. At present it is not possible to calculate a response rate since we do not know the number of persons eligible to be interviewed in the sampled households. This information will be available when the content of REG forms will be read and recorded.
Column 3 compares the number of returned individual questionnaires with the number of people aged 15 or more on January 2008, as it results from registry office records.

Table 5 – Returned IND questionnaires and number of individuals aged 15 or more by Municipality

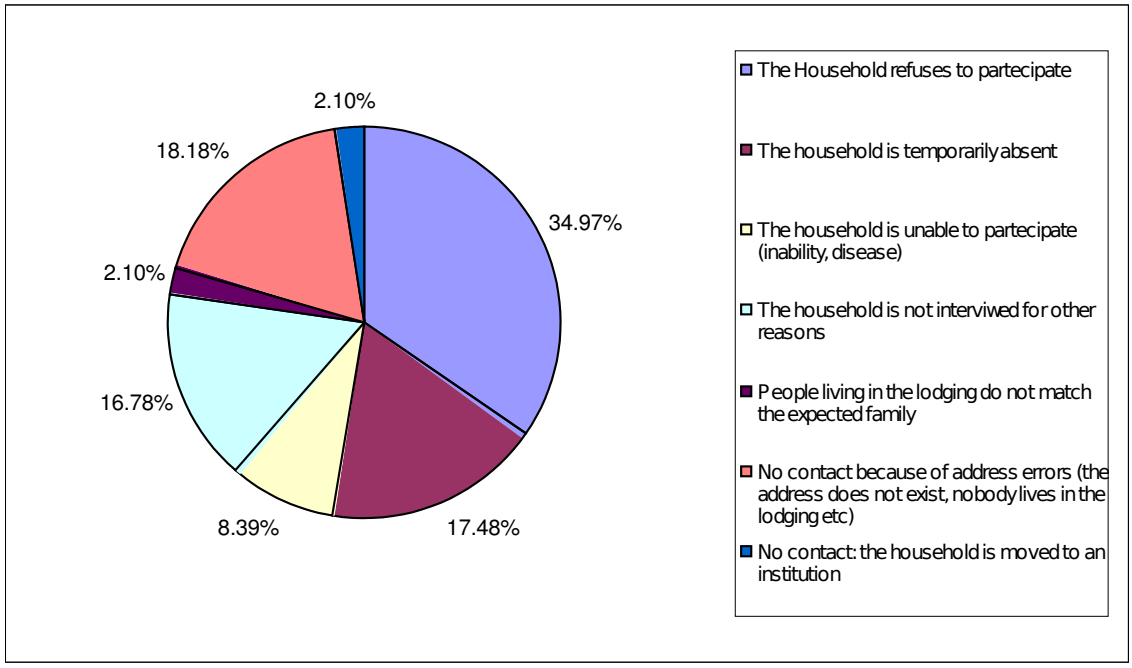
Municipality	**Number of IND forms**	**People aged 15 or more on January 2008***	**IND/(people aged 15 or more)**
Bientina	50	6,147	0.81%
Calci	46	5,494	0.84%
Casciana Terme	64	3,186	2.01%
Cascina	60	36,745	0.16%
Castelfranco di Sotto	90	10,704	0.84%
Castellina Marittima	48	1,744	2.75%
Castelnuovo di Val di Cecina	70	2,146	3.26%
Guardistallo	34	1,083	3.14%
Montecatini Val di Cecina	62	1,750	3.54%
Montescudaio	32	1,597	2.00%
Montopoli in Val d'Arno	73	9,328	0.78%
Ponsacco	39	12,631	0.31%
Pontedera	51	24,397	0.21%
Riparbella	63	1,383	4.56%
San Miniato	80	24,143	0.33%
Santa Croce sull'Arno	89	11,508	0.77%
Santa Maria a Monte	60	10,422	0.58%
Vecchiano	64	10,618	0.60%
Vicopisano	43	7,189	0.60%
Volterra	26	10,012	0.26%
Total municipalities sampled in 2008	**1,144**	**192,227**	**0.60%**
Calcinaia	47	9,026	0.52%
Lari	67	7,463	0.90%
Pomarance	63	5,525	1.14%
Municipalities present in previous waves (*with* oversampling)	**177**	**22,014**	**0.80%**
Pisa	94	77,972	0.12%
San Giuliano Terme	61	27,313	0.22%
Municipalities present in previous waves (*without* oversampling)	**155**	**105,285**	**0.15%**
Pisa province	**1,476**	**355,297**	**0.42%**
Tuscany	**4,334**	**3,222,720**	**0.13%**
Italy	**45,963**	**51,252,247**	**0.09%**

Data source: Istat

Figures 2, 3 and 4 show the non-response reasons for Pisa, Tuscany and Italy. Missed interviews arise both from failed contacts (errors in address, moving of the household to an institution or

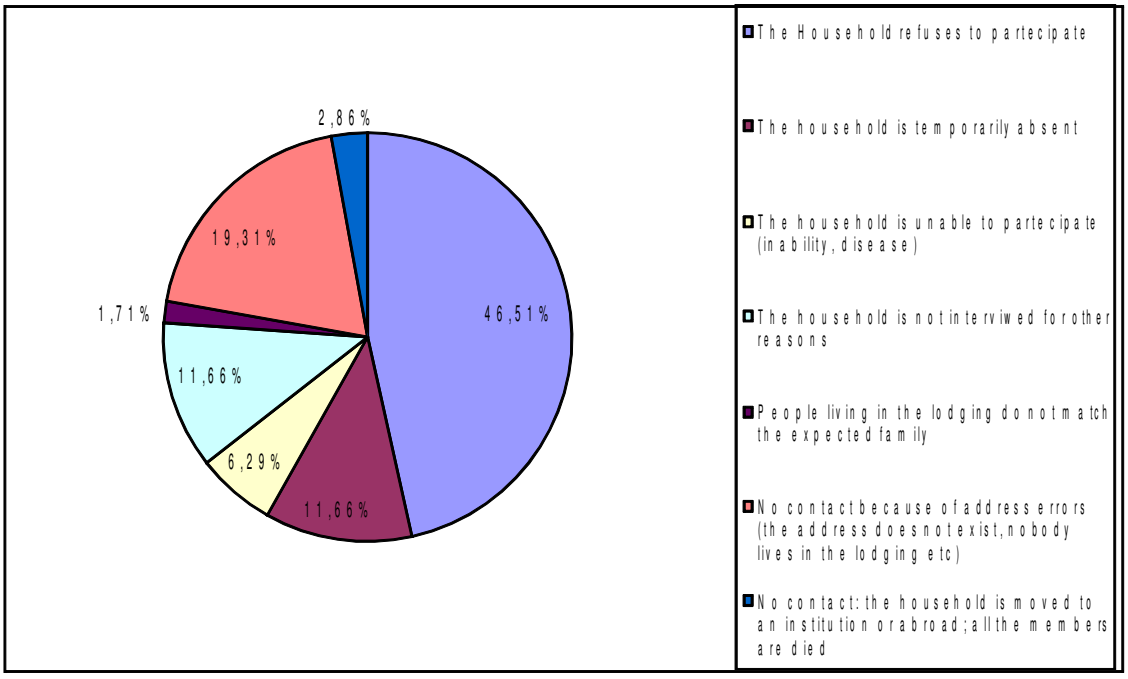
abroad etc..) and following a positive contact (refusal to participate, the households is temporary absent etc).

Figure 2 – Reasons for non-response - Province of Pisa – EU SILC 2008



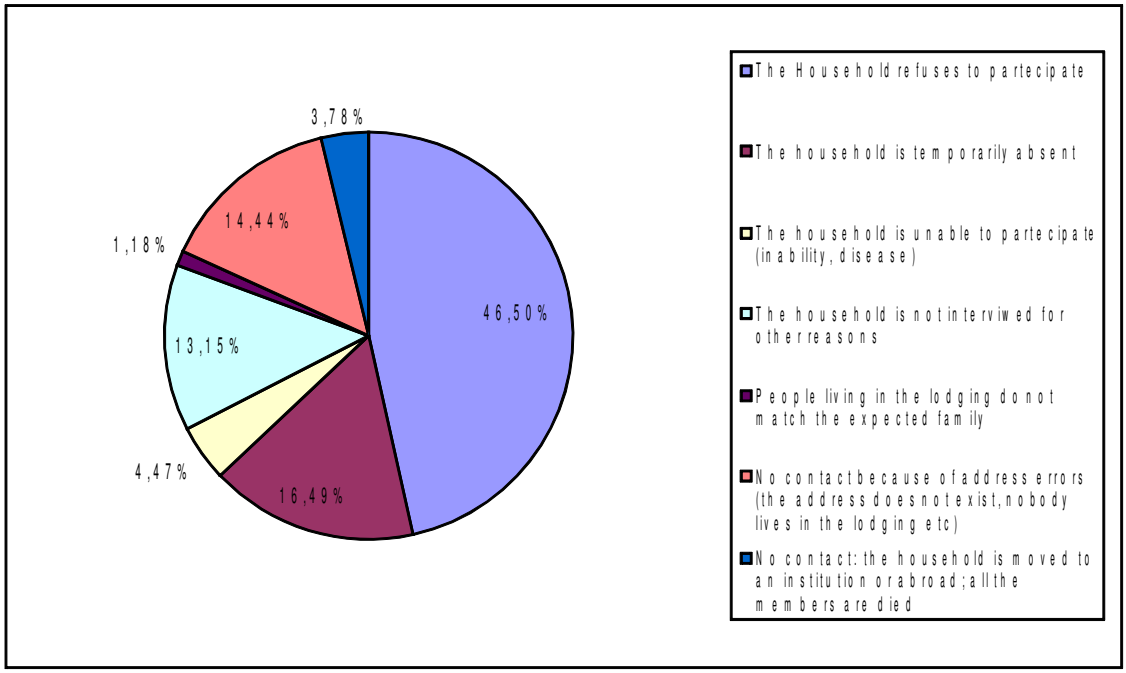
Data source: Istat

Figure 3 – Reasons for non-response - Tuscany – EU SILC 2008



Data source: Istat

Figure 4 – Reasons for non-response - Italy – EU SILC 2008



Data source: Istat

The non-response main reason is the *refusal to be interviewed* which accounts for 46.5% of non responses both in Tuscany and Italy. In the Province of Pisa the percentage of refusals is lower (about 35%), with comparatively higher rates for missed interviews due to the inability of the household to participate (8.39%) or other not specified reasons (16.79%).

In the following tables, response and non-response rates are further analysed by age, gender and citizenship of the head of the household (or reference person)[6]. Analysis are provided for the Province of Pisa, Tuscany and Italy.

Table 6 – Response and non-response rates by age of the head of the household, Province of Pisa

Interview?	<=35	36-45	46-55	56-65	>=65	Total
No	20	46	29	18	30	143
	14.0%	32.2%	20.3%	12.6%	21.0%	100.0%
	17.1%	27.4%	19.6%	12.5%	12.3%	17.4%
Yes	97	122	119	126	213	677

[6] In Istat surveys, the status of "head of the household" is derived directly from the Municipality register office records. The status is established on the basis of a pronouncement of the household itself at the moment of its first recording in the register.

	14.3%	18.0%	17.6%	18.6%	31.5%	100.0%
	82.9%	**72.6%**	80.4%	87.5%	**87.7%**	82.6%
Total	117	168	148	144	243	820
	14.3%	20.5%	18.0%	17.6%	29.6%	100.0%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Data sources: Istat

Table 7 – Response and non-response rates by age of the head of the household, Tuscany

Interview?	**<=35**	**36-45**	**46-55**	**56-65**	**>=65**	**Total**
No	71	123	62	59	107	422
	16.8%	29.1%	14.7%	14.0%	25.4%	100.0%
	23.8%	25.8%	14.7%	13.7%	13.1%	17.3%
Yes	227	353	361	371	711	2023
	11.2%	17.4%	17.8%	18.3%	35.1%	100.0%
	76.2%	**74.2%**	85.3%	86.3%	**86.9%**	82.7%
Total	298	476	423	430	818	2445
	12.2%	19.5%	17.3%	17.6%	33.5%	100.0%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Data sources: Istat

Table 8 – Response and non-response rates by age of the head of the household, Italy

Interview?	**<=35**	**36-45**	**46-55**	**56-65**	**>=65**	**Tot**
No	616	941	813	724	1478	4572
	13.5%	20.6%	17.8%	15.8%	32.3%	100.0%
	22.7%	19.6%	16.6%	15.9%	16.3%	17.6%
Yes	2097	3864	4079	3842	7588	21470
	9.8%	18.0%	19.0%	17.9%	35.3%	100.0%
	77.3%	80.4%	83.4%	**84.1%**	83.7%	82.4%
Tot	2713	4805	4892	4566	9066	26042
	10.4%	18.5%	18.8%	17.5%	34.8%	100.0%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Data sources: Istat

As expected, the response rate depends on the head of the household age (Chi-squared test *p-value* less than 0.01 for any geographical context). Data show that younger households are interviewed less frequently than older ones. Differences between response rates by age class are higher for Tuscany and the Province of Pisa (standard deviation equal to 5) with respect to Italy (standard deviation equal to 2.4).

Table 9 – Response and non-response rates by gender of the head of the household, Pisa Province

Interview?	**Gender**		**Total**
	Female	**Male**	
NO	42	101	143
	29.4%	70.6%	100.0%
	17.6%	17.4%	17.5%

YES	196 29.0% 82.4%	480 71.0% 82.6%	676 100.0% 82.5%
Total	238 29.1% 100.0%	581 70.9% 100.0%	819 100.0% 100.0%

Missing values: 1

Data sources: Istat

Table 10 – Response and non-response rates by gender of the head of the household, Tuscany

Interview ?	**Gender**		**Total**
	Female	**Male**	
NO	135 32,2% 17,9%	284 67,8% 16,9%	419 100,0% 17,2%
YES	621 30,7% 82,1%	1399 69,3% 83,1%	2020 100,0% 82,8%
Total	756 31,0% 100.0%	1683 69,0% 100.0%	2439 100.0% 100.0%

Missing values: 6

Data sources: Istat

Table 11 – Response and non-response rates by gender of the head of the household, Italy

Interview?	**Gender**		**Total**
	Female	**Male**	
NO	1564 34.4% 19.5%	2986 65.6% 16.6%	4550 100.0% 17.5%
YES	6441 30.1% 80.5%	14960 69.9% 83.4%	21401 100.0% 82.5%
Total	8005 30.8% 100.0%	17946 69.2% 100.0%	25951 100.0% 100.0%

Missing values: 91

Data sources: Istat

The head of the household gender does not affect response rates in Tuscany and in the Province of Pisa (Chi-squared test p-values are 0.59 and 0.99 respectively). On the contrary, at National level households were the reference person is a male are significantly associated with a higher response rate (Chi-squared test p-value equal to 0.00).

Table 12 – Response and non-response rates by citizenship of the head of the household, Pisa

Interview?	**Citizenship**		**Total**
	Italian	**Foreing**	
	136	7	143

NO	95,1%	4,9%	100,0%
	17,3%	19,4%	17,4%
YES	648	29	677
	95,7%	4,3%	100,0%
	82,7%	80,6%	82,6%
Total	784	36	820
	95,6%	4,4%	100.0%
	100.0%	100.0%	100.0%

Data sources: Istat

Table 13 – Response and non-response rates by citizenship of the head of the household, Tuscany

Interview?	**Citizenship**		**Total**
	Italian	**Foreing**	
NO	384	38	422
	91,0%	9,0%	100,0%
	16,5%	34,2%	17,3%
YES	1950	73	2023
	96,4%	3,6%	100,0%
	83,5%	65,8%	82,7%
Total	2334	111	2445
	95,5%	4,5%	100.0%
	100.0%	100.0%	100.0%

Data sources: Istat

Table 14 – Response and non-response rates by citizenship of the head of the household, Italy

Interview?	**Citizenship**		**Total**
	Italian	**Foreing**	
NO	4226	346	4572
	92,4%	7,6%	100,0%
	16,9%	31,2%	17,6%
YES	20707	763	21470
	96,4%	3,6%	100,0%
	83,1%	68,8%	82,4%
Total	24933	1109	26042
	95,7%	4,3%	100.0%
	100.0%	100.0%	100.0%

Data sources: Istat

According to our data, the citizenship of the head of the household affects response rates in Italy and Tuscany, with significantly higher values for households with an Italian reference member. In the Province of Pisa, on the contrary, the response rate does not significantly change in case of Italian or foreign head of the household (chi-squared test p-value equal to 0.92).

3. Institutional contacts

In the first 18 months of the SAMPLE project, The Province of Pisa-UROPS has activated important contacts with public and third sector organisations with the aim of involving them in the implementation of the project. In particular, the actors which have been contacted are:

- Social Observatories Regional Network;
- Health's Societies and local public agencies;
- Third sector organisations;
- Local Public Agencies.

A) Social observatories regional network

UROPS participated at the Social Observatories Regional Network coordinated by the Regional Social Observatory. In the Regional Network there are the ten Social Observatories of theTuscany Region.
The network has created a regional common set of indicators (more than 200) in order to monitor social policies and a common methodology for the acquisition of data.
In this list there are many indicators that are important for the estimation of the multidimensional measure of poverty at supra-municipal level (demographic profile, health state, essential level of territorial health care, elderly persons, families and youngsters, immigration, disability, mental health, dependences).
In November 2009, the Tuscany Region will approve specific guidelines for social policies planning which will include the monitoring of the set of indicators. These guidelines will be addressed to the Health Society, i.e. the public organisation that has the competence in social policies in Tuscany.
Simurg (SR) and UROPS have actively participated to the methodological definition of poverty and social exclusion indicators and also disseminated the SAMPLE experience during this meetings. Both partners are working to integrate this system within the Sample Project.

B) Health Societies

The **Health Societies**, (SdS according to the Italian abbreviation) are public organisations which have the function of social and sanitary planning at supra-municipal level.
The Health Societies are formed by health care companies and Municipalities. They represent one of the most important innovations in the integrated health program and social assistance and they have a specific planning task for the integrated social and health policies.
In this context, UROPS has organized informal and formal meetings with the menagers of Health Societies in order to propose them the subscription of a formal agreement aiming at:

- the sharing of the list of common indicators which is defined by the Social Observatories Regional Network;
- the collaboration for the involvement of third sector and local stakeholders in the construction of the set of indicators (Delphi method);
- the participation at the provincial Observation System to monitor poverty, vulnerability and social exclusion.

All Health Societies have given their verbal assent to subscribe the agreement and they have nominated a person in charge to collaborate with the Sample project Consortium.

C) Local public agencies

UROPS has also activated contacts with three important local public agencies in order to have access to their administrative databases:

- The Provincial Jobcentre, the office that works to ease the matching of labour demand and supply;
- The Department of Finance - Revenue Agency Organisation, the public organisation that manages tax returns;
- The Italian Social Security Service – INPS, the main Social Security Administration in Italy. Most workers are registered to INPS (private and public workers, self employed, etc.).
  The acquisition of the database is an important task for the developing of the integrating model, which is a crucial task in WP3 and also in the whole project.

These are the results that UROPS has achieved by now:

Revenue agency database

- The Revenue Agency uses the SIATEL System, which contains data on tax returns from 2004 to 2007.
- In a first moment UROPS has been authorised from the Department of Finance to access individual data using a password.
- Afterwards UROPS requested the authorisation to export the whole database and on 12 October the Revenue Agency accepted our request (by email) and sent us also the structure of the records. They asked us to select the fields in which UROPS were interested.
- UROPS sent them a proposal and is now waiting to receive the cd with data.

Jobcentre database

- The provincial Jobcentre uses the IDOL System to store data; the IDOL database contains data related to people registered as unemployed and to the start and cessation of jobs provided by companies. The IDOL database also contains information upon income and familiar conditions of unemployed people and upon workers' professional status.
- People who are stored in the Provincial Jobcentre's database result to be 30.574 at 31/12/2008, a number to which must be added the quota of workers communicated by companies.
- UROPS has gained the authorisation to access the IDOL database with individual code.
- After many discussions and meetings, UROPS selected the database records fields and is now waiting to receive, the Cd with the IDOL data.

INPS Database

During the first 18 months of the project, UROPS has established many formal and informal contacts with local, regional and national INPS delegations (at provincial, regional and national level) in order to explore possible technical and administrative solutions to gain access to their databases.

In the end, UROPS made two formal requests to INPS (in January and in September 2009) for accessing and acquiring on electronic format data which are contained in record related to the following three internal databases:

- Database of **Active Positions-** it contains workers' data;

- Database of **Pensions-** it contains data of pensions according to amount, pensions supplemented by guaranteed minimum income, etc.
- **The ISEE** Database (Indicator of the Equivalized Household Economic Position)- it contains ISEE declaration's data. The ISEE declaration must be submitted by people who want to obtain reduced tariffs in order to gain access to services provided by Public Administrations (school, social housing, etc…).

UROPS has just received a denial from INPS. They explained that in order to grant access to their database, they should have been formal partner in the Sample Project from the beginning. However, they will give to the project aggregated data at municipal level.

D) Caritas

One of the most important local organisation who realises actions to contrast poverty is Caritas. Caritas has more than 200 counselling centres in Tuscany and they are part of the Mirod Network. This Network, created in 2003, has designed an unique database, that contains the materials collected in all Caritas' counselling centres.
UROPS has gained the access to the MIROD database and has started the collaboration with Caritas within the Sample project (at the end of this month this collaboration will be consolidate with a formal agreement).

E) Local stakeholders

The involvement of local stakeholders (institutional and not institutional) has three main objectives:
- the sharing of poverty indicators (Delphi Metod);
- the development of the "Observation System to monitor poverty, vulnerability and social exclusion";
- the analysis of their information systems.

With the collaboration of Health Societies and Caritas UROPS has selected 573 stakeholders using this method:
- Extraction from the Provincial Register of third sector organisations. The regional law of social policies previews that all third sector organisations have to be registered in a Provincial Register. In our Register there are 602 organisations.
- Selection (with the aid of the head of Provincial third sector) of the organisations that realise actions against poverty in a multidimensional concept or that have a particular point of view about living conditions of local population. Exclusion of the other organisations.
- Sharing of the list with Health Societies and Caritas.
- Selection of local associations or key persons (not included in Provincial Register) like self help associations, parish priests, medicals, etc.

4. Small Area combined estimation at LAU1 and LAU2 levels

4.1 The problem

The Italian EU SILC is based on a stratified two-stage sampling design. First stage units are given by municipalities, stratified according to Administrative Province and demographic size (288

strata). Among municipalities those with at least 30,000 inhabitants are considered self representative and form a take-all stratum. Secondary sampling units are given by households.
In Italy, the survey is designed to obtain reliable estimates at the level of 20 Administrative Regions (NUTS2 according to the EU "Nomenclature of Units for Territorial Statistics"; see http://europa.eu.int/comm/eurostat/ramon/; see also Istat, 2008).
Since our domains of interest are obtained subdividing the population of Administrative Regions (NUTS2 level) by Province and then by "Health Society" and Municipalities (NUTS3, NUTS4 level – or LAU1, LAU2 level), the sample in many of these domains can be too small to obtain enough reliable estimators for meaningful analyses. In terms of households, the domain specific sample sizes range from a minimum of 24 to a maximum of 201; 25th, 50th and 75th percentiles are respectively 29, 33 and 38. Eventually, we note that Provinces and Municipalities are in some way planned domains (i.e. they are strata), while the other domains of interest Health Society can be obtained as a union of strata (i.e. as union of Municipalities).
Thus, the domains of interest are Municipalities (LAU1 level) or aggregation of Municipalities (LAU2 level). These administrative regions correspond to levels of governance for which policy makers need to have relevant and accurate figures on poverty and living conditions.
The target parameters are poverty indicators. These are mainly linear or non linear combinations of totals and means (see Deliverable 4 for a review of traditional poverty indicators). Here we refer to the estimation of area totals or area means. Moreover, we focus here on direct estimation of the target quantities under the EU SILC sample design. In Deliverable 8, on the other hand, the reader can find the description of the model-based approach and more details on the estimation of linear and non linear combinations of totals and means, and on the estimation of quantiles of the cumulative distribution function of income.

4.2 Direct estimators

Even if the domains we consider are combinations of planned domains, we plan to modify the official final weights published in the EU SILC data set, in order to have weights calibrated on the distribution of the Province of Pisa population by Municipalities and Health Society. We recall that final published weights are obtained by a double calibration correction of basic weights that are defined as the inverse of inclusion probabilities. The first step adjusts basic weights for non-response, while the second step modifies these intermediate weights to calibrate them to known totals as suggested in the Eurostat guidelines for the EU SILC survey (Istat, 2008). In particular, the distribution of the population by gender, age class and geographical region will be considered.
In order to obtain weights calibrated on the distribution of the population in the domains of interest (i.e. Municipalities and Health Societies), we plan to start from the survey intermediate weights and re-make the second step, considering the following calibration variables: Administrative Region of residence, household type[7], gender and age (5 classes). More precisely, the weights will be calibrated to the population of Administrative Regions classified by household typology and to the same population classified by age classes.
In the calculation of the calibration weights, we will refer to methods which have the advantage of producing always positive weights (see Deville and Särndal, 1992 for more details).

[7] Nine household typologies are considered in the EU-SILC survey. These typologies are defined by simultaneously considering the household size, the presence of children and the age of components. They are defined as follows: 1. One person households; 2. Two adults, no dependent children, both adults under 65 years; 3. Two adults, no dependent children, at least one adult 65 years or more; 4. Other households without dependent children; 5. Single parent household, one or more dependent children; 6. Two adults, one dependent child; 7. Two adults, two dependent children; 8. Two adults, three or more dependent children; 9. Other households with dependent children.

4.3 Variance of the direct estimators

In the production of small area estimates using a procedure such as combined estimation, a major requirement is that of evaluating the reliability of the direct estimates. To achieve this goal we basically need to estimate their variances and, eventually, to apply small area multivariate models (Rao 2003). We need also to estimate the covariances between estimators of different rates obtained for the same domain.

Evaluating the variances and covariances of the direct estimators is not an easy task for the EU SILC survey, since: i) the considered poverty rates are non-linear functions of data; ii) the underlying design is complex; iii) the weights used in their computation incorporate, as it has been previously described, two stages of calibration corrections. Other works in this field (Verma and Betti, 2005; Betti and Verma, 2006) suggest a solution based on re-sampling algorithms and in particular on the bootstrap estimation strategy.

For stratified multistage designs one solution is that proposed by McCarthy and Snowden (1985). This bootstrap procedure is an asymptotically valid method in assessing the variability of direct estimators. Specifically, every bootstrap sample is obtained drawing with-replacement a random sample of *n-1* PSU's out the *n* at disposal in the stratum. After every re-sampling, the original weights are properly rescaled and the bootstrap variance estimate of the corresponding indicator is obtained by the usual Monte Carlo approximation based on the independent bootstrap replicates.

For every LAU1 or LAU2 area the variance of the direct estimator is calculated in the following way

$$\psi_i = V\left(\hat{\theta}_i\right) = \frac{1}{B-1}\sum_{b=1}^{N}\left(\hat{\theta}_i^{*b} - \hat{\theta}_i^{*}\right)$$

where $\hat{\theta}_i$ is the direct estimator in the *i*-th small area, $\hat{\theta}_i^{*b}$ is the bootstrap estimator for the same area obtained by the *b*-th bootstrap and $\hat{\theta}_i^{*}$ is computed as

$$\hat{\theta}_i^{*} = \frac{1}{B}\sum_{b=1}^{N}\hat{\theta}_i^{*b}.$$

Finally the standard error of direct estimator in the *i*-th area is defined as:

$$SE\left(\hat{\theta}_i\right) = \sqrt{V\left(\hat{\theta}_i\right)}.$$

For an application to Polish EU SILC data see Zięba and Kubacki (2009). They replicated the sampling procedure $B = 500$ times.

Other bootstrap variance estimators have been proposed and analyzed for sampling designs as general as multi-stage designs with stratification of primary units. See Rao (1999) for more details.

Anyway, all these estimators rely on the assumptions that the number of strata is large and that few primary units (but at least two) are sampled from each stratum, so that the sampling fraction at the first stage is negligible. This latter assumption is not met in the EU SILC Italian design, since in some strata all the primary units enter in the sample (Municipalities).

For this reason, we plan to refer to the bootstrap algorithm proposed by Fabrizi et al. (2008), in which any bootstrap sample is the union of two sub-samples, one taken re-sampling the population in the non self-representative strata, and the other drawn from the stratum of self-representative

municipalities, where the sampling design is actually single stage. After it is drawn, each bootstrap sample undergoes the same calibration adjustment of weights to known totals applied to the original sample. The algorithm has been tested by the authors by means of simulation exercises and it provided estimates close to those obtained using the linearization method for simpler parameters (i.e. averages) for which this latter method may be applied.
The algorithm has a limitation: variances cannot be estimated for domains in which there are no 'poor' households in the sample. In fact, we would have an estimate of 0 in all bootstrap samples and 0 estimates of the variances (and covariances). A solution could be that of referring to the model based predictions of poverty rates for these domains, obtained using the methodology illustrated in Deliverables 4 and 8.

4.4 Auxiliary information for calibration

The direct calibrated estimators described in section 4.1. make use of auxiliary information, that is of information on the domains of interest available from sources independent from the EU SILC survey, such as Censuses or Administrative archives, which may be used to improve the estimation of area-specific poverty rates.
Totals are obtained from the same data sources used in the derivation of the final official weights for all the variables, except for the distribution of the population by household type within administrative regions, which can be obtained as an average of the quarterly Labour Force Survey results in 2007.
Other studies (Fabrizi et al., 2008) have analyzed similar data sets, in which domains are given by Administrative Regions. They found evidence that poverty rates are strongly correlated with the unemployment rate. These estimates are based on the Italian Labour Force Survey (ILFS; Istat, 2003). In fact, the correlation is rather high (around 0.7).
At our level of governance (Municipality and Health Society) Istat does not routinely calculate and publish estimates of unemployment rates. Istat is unlikely to provide us with the estimates of the annual average unemployment rates for our domains or at Province level in year 2007 (the income reference period). By the way, although calculated on a much bigger sample (the ILFS has an overall annual sample of around 300,000 households), estimates at the level of disaggregation we are interested in are characterized by a considerable level of uncertainty, in particular for typologies for which the rate of participation to the labour market is low. This uncertainty has to be accounted for in the analysis.
We can also consider auxiliary information at our level of governance using the regional section of the National System of Accounts, the ILFS and other administrative archives as data sources. The following variables resulted to be relevant in previous poverty studies: per-capita consumption of the household sector, per-capita GDP, per-capita employee income, per-capita expenditure for leisure and culture, per-capita taxable income, share of workers/value added in the manufacturing industry, school abandonment rate, annual average unemployment rate.
Note that all these variables, being estimated at our level of governance, are characterized by a level of uncertainty that may be considered in the implementation of the models to obtain the combined estimates.

References

Deville, J. C. and Sarndal, C. E. (1992), *Calibration estimators in survey sampling*, Journal of the American Statistical Association, 87, 376-382.

Eurostat (2007), *Comparative EU statistics on Income and Living Conditions: Issues and Challenges* - Proceedings of the EU-SILC conference (Helsinki, 6-8 November 2006)

Freguja, C. (2009), *Ampliamento del campione Eu-silc per la Provincia di Pisa: primi risultati relativi alla partecipazione delle famiglie all'indagine*, Istat, rapporto interno.

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

McCarthy, P. J. and Snowden, C. B. (1985). *The bootstrap and finite population sampling*. In Vital and Health Statistics 2–95. Public Health Service Publication 85-1369, U.S. Government Printing Office, Washington, DC.

European Community (2003) *European Parliament and Council Regulation (EC) N° 1177/*2003 of 16/6/2003, Official Journal of 3/7/20

Fabrizi, E., Ferrante, M.R. and Pacei, S. (2008), *Measuring Sub-National Financial Poverty by using a Small Area Multivariate Approach*, The Review of Income and Wealth Volume 54 Issue 4 , Pages 597 - 615 (December 2008)

Rao, J. N. K. (1999), *Some current trends in sample survey theory and methods*, Sankhya, Series B, 61, 1-55

Rao, J.N.K. (2003) *Small area Estimation*, Wiley Series in Survey Methodology, John Wiley and Sons.

SAMPLE Consortium - Deliverable 2

SAMPLE Consortium - Deliverable 4

SAMPLE Consortium - Deliverable 8

Verma, V. and Betti, G. (2005), *Sampling errors for Measures of Inequality and Poverty*, Classification and Data Analysis 2005, Book of Short Papers, CLADAG, 175-178.

Betti, G. and Verma V., (2006), Approximate variance estimates of poverty measures for the application for EBLUP for small area estimation, Working Paper n. 57, Dipartimento Di Metodi Quantitativi, Universita di Siena.

Zieba, A. and Kubacki, J. (2009), *EU-SILC and Income Poverty Measures at sub-national level in Poland*, " SAE 2009 conference, http://cio.umh.es/sae2009/AbstractList.asp