Global Strategy
IMPROVING AG-STATISTICS

# Research Plan

## Integrated Survey Framework for Agricultural Statistics

Piero Demetrio Falorsi

## 1. Background

*Strategic research objective*

This document illustrates the work plan of the research activity of the *Global Strategy* (GS) on *Integrated Survey Framework for Agricultural Statistics*. The proposal is focused on developing countries and takes into account that one of the *Global strategic objectives* (FAO, 2011) is the production of integrated and consistent (or coherent) agricultural statistics on phenomena related to three different target populations, the units of which are respectively:

- ✓ agricultural plots (for what concerns the environmental aspects);
- ✓ households (for the social aspect)
- ✓ farms (for the economic dimension).

Resorting to the integration of different data sources allows achieving a better coverage of specific statistics for which a suitable solution cannot be found only surveying specific units: e.g. livestock and fisheries are not related to agriculture plots, etc. The integrated survey framework also offers the opportunity to compare sample unit data across time, providing a major validation tool to improve data quality.

The research will mainly focus on:

- ✓ sampling design methods (how to collect the data of interest);
- ✓ estimation methods (how to produce estimators of the target parameters using existing data in an integrated framework).

The strategic goal is identifying economic and feasible survey strategies for assuring integrated and consistent agricultural statistics in developing countries.

However, an important prerequisite in this context is the clarifications of the concept of integration along with the identification of the mayor issues to be solved through this integration.

*Overview of the existing problems*

Indeed this objective is a complex task that may be solved using mixtures of different and reasonable methods (for sampling and estimation). The specific mixture for a given country can be properly defined on the basis of the specific informative context.

The GS indicates to build a Master Sampling Frame (MSF) where information on land parcels, households and farms are interlinked, thus allowing to simultaneously provide consistent and integrated statistics on the environmental, social and economic dimensions of agriculture. There are several ways of building the MSF, some of which are illustrated in FAO (2012). It is useful to underline that the present research activity does not deepen the construction of MSF, since this topic is the specific focus of another research activity within the GS. However the MSF is a relevant aspect to be considered when developing integrated survey strategies.

The MSF construction is a complex and expensive task and clear guidelines and proper implementation of the survey field operations are required in order to improve the quality of the linkage among different statistical data bases. Often, collected data have to be treated with sophisticated statistical procedures for refining the quality of the record linkage.

Although it is being an effective support, the MSF allows solving the problem of consistency of results only in an oversimplified context where:

- ✓ there is only one single survey selected in the MSF for collecting and disseminating data referred for the three target population of interest;
- ✓ linkage among the units of the different populations is performed without taking into account potential linking errors.

Even in this simplified context, traditional sampling methods do not allow to treat in a coherent and correct way the problem that, in dynamic environments, the MSF becomes quickly outdated, and the survey enumerators have to face a current situation, which can be quite different from that represented in the MSF: e.g. a farm might be divided, a new household may be formed by existing ones, and so on. These emerging problems could be tackled by adopting indirect sampling strategies (Lavallé 2007) as a unified approach both for sampling and direct estimation.

If the MSF is affected by a relevant linkage error, the production of the estimates related to the different target units, by calibrating on some totals available in the MSF might introduce a non trivial bias in the estimates. In these cases, for an efficient construction of the MSF, it is strategic to produce a reliable estimate of the linkage error.

In situations in which the MSF exists, but different surveys are conducted (either on units of the same type or on units of different types) the consistency of the estimates is not assured. Some methods dealing with this specific problem are developed considering the case of units of the same type (Kim and Rao, 2012; Knotterus, 2006). The above methods are based on a weighting approach - or a micro-integration approach - that generally shows good properties in terms of sampling design and models used for the inference.

*On the concept of integration*

The concept of data integration is not well defined in the literature, see, for example, Judson (2007), and it may mean different things to different researchers. For this specific research program the focus will be on the general issue of integrating data sets which share some common statistical units, for which– however-a perfect matching of these common units across the files is not possible. In this context a specific problem to be issued is that of making valid inferences with an integrated collection of data sets, providing statistical methods which automatically account for the uncertainty in the linkage step.

In finding some valid strategy for the integration we need to consider either the specific purposes of the integration and the possible informative contexts in which the integration is performed.

The integration of data from different sources can be desirable for many different reasons:

a) ***Observational reasons:*** in some informative contexts the sampling frame is only available for one of the relevant populations. This population can be observed using standard techniques; furthermore if its specific units can be linked with those units of other populations of interest, the available frame can be used to gather data related to the population without frame. An example may clarify this situation. Consider the case of employment status, a variable which is related to the households' population. As a general rule, this variable should be measured by a survey on households. However, if the frame is available only for the farms, it is necessary to collect the needed information through a survey on farms, asking, for instance, to each farm worker how many people in her/his family have worked in the reference week, obtaining in such a way the integration of data from households and farms.

b) ***Survey's goals:*** One of the shortcomings of current statistical systems in both industrialized and developing countries is that data are collected by sector, using different sampling frames and surveys. This division leaves no opportunity to measure the impact of an action in one sector on another. Surveys are often conducted independently one of each other and they are performed on an ad-hoc basis (e.g. with no links to a master sampling frame or the use of geo-referenced units). As a consequence, it is difficult to integrate data from various surveys for in-depth analysis with cross tabulation of variables. Data on crop and livestock production are often drawn from separate surveys, which are based on separate samples. The separate datasets provide no basis for analyzing the characteristics of farms that produce both crops and livestock, or for comparing them to farms that are specialized in one or the other. Household surveys are often conducted in isolation from production surveys with no coordination or with too small sample sizes to disaggregate the data into the rural and farm sectors. The results generated from these surveys are also not integrated into a common database for access by data users. A good solution for these issues seems to be the joint observation of units from different populations.

c) ***Efficiency reasons:*** In some circumstances (the) phenomena related to different units have correlation structure that could be taken into account through specific models. Considering these correlation structures in producing estimates allows to gain in efficiency and to calculate estimates which are automatically integrated.

Also the definition of consistency in an integrated framework is still an open research topic. It may be defined:

- ✓ with respect to the known auxiliary totals available either in the MSF or in the registers which represent the sampling frames for the units of the three target populations;
- ✓ with respect to the estimates of the target variables that are collected by the surveys;
- ✓ by considering a relation of strict equality, or a relation of vicinity which takes into account different factors of uncertainty deriving from: sampling, linkage errors, measurement error, models used for inference.

The statistical framework described above is even further complicated by the different informative contexts characterizing the developing countries, which may be broadly divided into two main categories.

- ✓ The first category, denoted as Countries with Integrated Statistical System (CISS) includes those countries which conduct the agricultural statistics by fostering the cooperation of the different organizations which have the responsibility of agricultural statistics. By considering the ongoing related FAO initiatives, this category may be individuated by considering the developing countries which have launched projects for the construction of a national MSF. At the end of these actions, in these countries, the surveys strategies may be planned in advance in an integrated way. The micro-data, either from surveys or registers, may be interlinked at different levels of aggregations.
- ✓ The second category, indicated as Countries with Non Integrated Statistical System (CNISS), in which the agricultural statistics are under the responsibility of different subjects who operate independently for all the phases of the surveys and do not share micro-data. Currently, this category includes most of the developing countries.

Actually, diverse intermediate levels of internal cooperation can be recognized in developing countries. Although being very rough, the above dichotomization allows identifying the main strategies which can be carried out for assuring the consistency of agricultural statistics. In CISS with it is possible either to build the MSF and to develop survey strategies which jointly consider the sample selection and the estimation. In CNISS, it is not possible to leverage the sampling design, the MSF cannot be built and the coherence of the survey estimates can be obtained only via estimation techniques.

The concept of integration obliges also to revise the well know frameworks on quality (Brakstone et al. 2001). The new framework should be developed by considering that the quality can represent a powerful leverage for enhancing the capacity building of developing countries. For instance, when indirect sampling is adopted, the estimation of the distribution of links among different statistical units (average, minimum and maximum number of links) is essential to model design effect and its main components (like weighting effect) for each sub-population of interest.

When a MSF or multiple registers are available, different sampling procedures can be carried out. Specific optimal criteria have to be specified by considering the several dimensions which characterize the sampling context: the populations of interest (parcels of the ground, households, farms); the different costs for the target units, the parameters to be estimated, the estimation procedure, the detail of the estimation domains, the field organization, etc.

In this context, the research should focus its attention on finding possible answers to the following **questions**.

1. What could be a reasonable formal definitions of the concept of consistency to be used in the context of integrated statistical surveys?
2. How to assure consistency among estimates from different surveys by considering both units of the same types and units of different types? In particular: How to interlink different statistical units in integrated survey strategies? How to deal with imperfect and not updated sample frames?
3. How to assess the quality of data for producing integrated agricultural statistics, i.e. to ensure a continuous improvement of the integrated statistics, considering the context of different statistical models (for record linkage and for predicting the variables of interest) and different approaches to inference (model assisted, model based and Bayesian)?
4. How to deal with the main key questions affecting statistical quality of integrated agricultural statistics, such as non-sampling errors (linkage error; nonresponse, etc., non-coverage), optimal sampling designs, and integrated estimation of small domains and of rare characteristics.

Finally, we note that a lot of different methods and practices are currently proposed in FAO for survey strategies in different domains (population, aquaculture, forestry, ecc.). Although most of the above methods can be well founded, it would be necessary to revise some of the existing practices on the basis of new research results in sampling methodology. This could help in standardizing and modernizing the statistical process in developing countries by defining a set of reasonable survey strategies which could represent recommended survey practices.

In section 3 below some research lines which address the above questions are presented clarifying the links among the questions and the expected research outputs.

## 2. Related work

Without having the ambition of the completeness, here in the following some related works are cited.

FAO has conducted a series of related work as those mentioned in section of References as FAO (2010 and 2012b).

The research activity *Improved methodology of Master Sample Frame* of the GS is surely connected to this line of activity. The GS research activities related to *estimation* (of crops, cost production , etc.) and *administrative data* are related; for instance, it could be useful to share the results of activity on *improving the quality of administrative data*. The Integrated estimation could be interlinked with the results of the activity on *Developing more efficient and more robust methods for administrative data*.

Statistics Netherlands has developed an Informative System to obtain consistency among tables of survey estimates (Knotterus, 2006).

Researchers of Statistics Canada are dealing with the theory of indirect sampling which allows dealing with imperfect frames and with the problem of reaching a target population starting from a list the units of which can represent a link to the target units (Lavalle, 2007, ).

The Italian National Institute of Statistics has conducted related research activities on: (*i*) Optimal sampling (Falorsi and Righi, 2008, 2012); (*ii*) Small Area Estimation[1] (D'Aló *et al.* 2012); (*iii*) Data integration (D'Orazio *et al.* 2006).

*Statistics Sweden* has carried out relevant research on model assisted survey sampling (Särndal et al., 2003) and on calibration and imputation in the treatment of nonresponse and frame imperfections (Lundstrom and Särndal, 2001).

The USDA has released important results on producing estimates with noncoverage and non response error (Kott *et al.*, 2006 and 2012).

The *Department Methods and models for Economics, Territory and Finance*[2] *of Sapienza University of Rome* has been in charge of research projects of strategic interest at European level (FP7) and at national level (co-financed by the Italian Ministry of University and Research) on the topics of integrating data from different sources (Alleva and Falorsi 2009; Liseo *et al.* 2006, Bramati 2013). Furthermore, it has conducted research on estimation in presence of linkage errors, published on top scientific magazines (Liseo and Tancredi, 2011 and Tancredi and Liseo, 2011) and on robust sampling design (Bramati, 2012).

The *University of Wollongong* has produced releventa research results on inference Based on Estimating Equations and Probability-Linked Data (Chambers et *al.*, 2009).

The Department of Statistics, University of Milan-Bicocca (Italy) has lead a deep study of a unified approach to estimation estimators in the context of Multiple Frame Surveys (Sing and Mecatti, 2011; Mecatti, 2007)

The *University of Ottawa* has produced relevant research results on the topics of combining data from independent surveys (Kim and Rao 2012).

The *Department of Mathematics and Statistics of the University of Laval (Quebec, Canada)* has produced relevant research results on the topics of capture-recapture sampling and estimation (Lavallée and Rivest, 2012), indirect sampling and integration of different sources of data (Rivest, Baillargeon, 2008), optimal model based stratification (Rivest, 2002).

The *Laboratory of Data Treatment Methodology (L.M.T.D.)of the Institute of Sociology, Université libre of Bruxelles* has produced relevant research results on both theoretical and applied survey sampling, focusing on the use of auxiliary information (Tillé, 1994), non-response (Droesbeke and Lavallée, 1996), sampling design and survey methodology (Tillè, 2006).

The *Institut de Statistique* of the Université de Neuchatel, has as main research interests in both theoretical and applied survey methodology, ranging from sample coordination (Nedyalkova, Tillé, and Pea, 2008, Nedyalkova and Tillé, 2008; Nedyalkova, 2009), model-based inference (Nedyalkova, and Tillé, 2012), nonresponse corrections using latent models (Matei and Ranalli, 2011), variance estimation via resampling methods (Antal and Tillé, 2011), measures of inequalities in sampling from finite population, (Langel and Tillé, 2013 and Massiani, 2012). Also, the Institute is actively involved in the *European Statistical Training Programme* and is responsible for the development and maintenance of the *sampling* package in R [3].

[1] Consider the result of the relevant ESSnet project available at: ; *http://www.cros-portal.eu/sites/default/files//WP6-Report.pdf* 2012.

[2] See http://www.memotef.uniroma1.it/newdip/inglese/dipartimento.php.

[3] See : http://cran.r-project.org/src/contrib/Descriptions/sampling.html.

In section 3. below some alternative strategies for achieving integrated statistics are outlined considering the different informative contexts deemed possible.

## 3. Strategies for achieving the integration

*The initial proposal of the Global Strategy*

In the Global Strategy to Improve Agricultural and Rural Statistics (FAO, 2010, pp. 24-26) the steps to implement an Integrated Survey Framework for Agricultural and Rural Statistics when a Master Sampling Frame is available are outlined. These steps are reported below:

A) Determine the set of core items for which at least annual data are required. For those core items not needed annually, group them by category, including economic variables such as farm structure expenditures and income; environmental measures such as the use of fertilizers and chemicals and land and water use; social variables such as household income and well-being; and other items of national interest (minor crop or livestock items, for example). Data for these items will come from rotating panel surveys based on a subsample of the core survey.

B) Select a replicated sample for the annual core items. In other words, instead of selecting one large sample, select several replicates. As shown in the following table (FAO, 2010, page 24), this allows a process to include some of the sample units in the survey across time for longitudinal analysis. The table below shows 12 replicates; 1 through 5 for year 1, 2 through 6 for year 2, etc. This provides longitudinal data, but also limits the number of times for respondent burden considerations.

C) Design a survey questionnaire to obtain the annual core data items. Each year the core questionnaire should contain supplementary questions regarding one of the subject matters described above. For example, in year 1 replicates 1 through 5 will be surveyed using the core questionnaire, which will also contain key questions about economic variables. The core questionnaire can either obtain all information required, or a subsample could be selected for the collection of the detailed data. In year 2, replicates 2 through 6 will be surveyed using the core questionnaire, which will also contain questions about environmental issues. By year 4, all of the subject matters will have been included.

The Table below provides an overview of a survey framework based on replicated samples that are surveyed each year for the annual core data items. In addition, each year the core questionnaire contains a set of supplemental questions for one of the subject matters that round out the minimum set of core data

***TABLE 1:*** *Example of a replicated* survey design with the use of an annual core questionnaire and rotating sets of supplemental questionnaires

| REPLICATE YEAR | REP 1 | REP 2 | REP 3 | REP 4 | REP 5 | REP 6 | REP 7 | REP 8 | REP 9 | REP 10 | REP 11 | REP 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | A | A | A | A | A | | | | | | | |
| 2 | | B | B | B | B | B | | | | | | |
| 3 | | | C | C | C | C | C | | | | | |
| 4 | | | | | D | D | D | D | | | | |
| 5 | | | | | A | A | A | A | A | | | |
| 6 | | | | | | B | B | B | B | B | | |
| 7 | | | | | | | C | C | C | C | C | |
| 8 | | | | | | | | D | D | D | D | D |
| 9 | | | | | | | | | A | A | A | A |
| 10 | | | | | | | | | | B | B | B |
| 11 | | | | | | | | | | | C | C |
| 12 | | | | | | | | | | | | D |

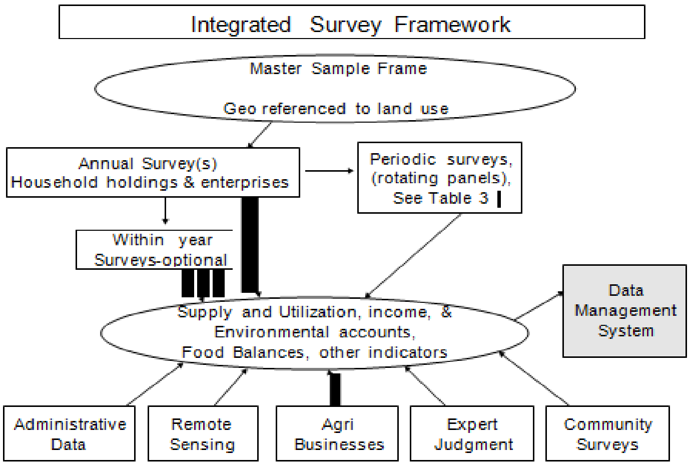
**Detailed Questionnaires for Rotating panel surveys**

Every replicate receives the same core questionnaire every year for annual core data items plus obtains data for one following rotating panels:

**A. Economic items including Farm structure, expenditures, income**

**B. Environmental items including inputs, chemicals, tillage, water use, land use**

The figure below provides an overview of how the annual and periodic surveys are connected in the data system. Note that within-year surveys can also be conducted using subsamples from the annual survey.


Integrated Survey Framework
Master Sample Frame
Geo referenced to land use
Annual Survey(s)
Household holdings & enterprises
Periodic surveys, (rotating panels), See Table 3
Within year Surveys-optional
Supply and Utilization, income, & Environmental accounts, Food Balances, other indicators
Data Management System
Administrative Data
Remote Sensing
Agri Businesses
Expert Judgment
Community Surveys

*Criticalities of the initial proposal*

The described strategy is certainly useful, since it helps in clarifying the expected results from the integration. However, it is only suitable in the case in which the Master Sampling Frame (MSF) is available and updated. This optimal situation is particularly rare and there might be many other informative contexts in which alternative solutions could be found to obtain the integration.

Note also that the availability of the MSF is a sufficient but not necessary condition for the integration. Moreover, in dynamic environments, the MSF becomes quickly outdated, and the survey enumerators may encounter an actual situation which is quite different from that represented in the MSF.

Although the MSF allows establishing an effective link between the different statistical units, important sampling and estimation problems still remain to be addressed, even in a simplified context where a single multipurpose survey is used to collect and disseminate statistics for the three target populations.

*Toward a holistic approach*

A key success factor to a good integration strategy in developing countries consists in taking into account, through a holistic approach, the various dimensions that identify the *country specific* informative context. In particular, the following dimensions play a fundamental role.

- ✓ **Country variables**

    - Geographical area;
    - Statistical level (quality of the output and of the entire process) in general and with regard to agricultural and rural statistics.
    - Subjects of interest for the country (target variables)

- ✓ **Features that define the types of databases and their integration**
    - *Role of the Census*
        - No Census (surveys on population and on agriculture);
        - Only the population census is available;
        - Only the agricultural census is available;
        - Both the population and the agricultural census are available.

    - *Periodicity of the different data sources for the integration*
        - Census data;
        - Administrative archives;
        - Judgments of experts (indirect)
        - Sample surveys.
    - *Specific types of integration*
        - Population Census – Agricultural census;
        - Census – Administrative Archives;
        - Census – Sample surveys;
        - Census – Administrative Archives – Sample surveys.

In the following sub-section the role of the census in defining the informative contexts is further clarified.

*Informative contexts*

As briefly discussed above, it is possible to define different informative contexts depending on the availability of the population and the agricultural census. A schematic representation of the situations that may be encountered in practice is given below.

***Informative context 1***: Countries in which there is neither the census of the population nor the census of the agriculture and only some surveys (with the related frame) are available.

***Informative context 2:*** Countries in which there is only the census of the population with some surveys on agriculture.

***Informative context 3:*** Countries in which only the census of the agriculture is available, with some surveys on the population.

***Informative context 4:*** Countries in which both the census of the population and the census of the agriculture are available.

In each defined informative context it is also possible to make a distinction between:

***a)*** Countries with integrated statistical systems (CISS), i.e. countries in which the census and/or the surveys are based on a single frame;

*b)* Countries with non-integrated statistical systems (CNISS), i.e. countries in which the census and/or the surveys are based on independent frames.

In synthesis, the informative context of a given country could be assigned to a specific cell of the table reported below.

| Informative context | Integration of the statistical system | |
|---|---|---|
| | **A : CISS** | **B : CNISS** |
| **1 : Only some surveys** | Ex. Guyana | Ex. Angola, Sierra Leone, Yemen |
| **2: Only population census** | Ex. Papua New Guinea, Nepal | Ex. Israel, Liberia, Mauritius |
| **3: Only agricultural census** | Ex. Tanzania | |
| **4: Agricultural and population census** | Ex. Canada, Mozambique, | Ex. Italy |

*Different strategies for different informative contexts*

The various informative contexts considered above underline the need of specific methodological strategies which allow the production of integrated statistics on agriculture. These strategies may operate at different phases of the statistical production chain.

- ✓ At the pre-processing phase, for producing linked data sets which are input for the successive phases.
- ✓ At the sample design phase in all the informative contexts where it is possible to plan the sample for the different target Populations (typically the informative context 4 and for the CISS.
- ✓ At the estimation phase for producing either integrated data sets or integrated statistics. In the less structured situations (typically the informative contexts 1b, 2b and 3b) it is possible to achieve the integration only via estimation.

## 4. Organization of the research and expected Outputs

*Organization of the research*

In the context of above sketched scenario, the research activity can be developed according to the following lines of action.

A. *Guidelines for consolidation of current survey practices*. This line deals with sampling design and estimation and proposes well founded, economic and feasible survey strategies assuring consistent agricultural statistics in developing countries. It starts both from current advancements in survey sampling theory and from the existing (and recommended) survey methods adopted in FAO. Reorganizing the existing research tools in a unified framework suitable for agricultural statistics is not a trivial task and represents a specific research output which can be converted rapidly in training materials and guidelines.

B. *Advanced topics for quality, design and estimation*. The second line of research aims at finding innovative and feasible solutions to the above issues.

*Human Resources*

Given its nature, the line A can be conducted essentially by internal resources, possibly recruiting a research assistant and an intern.

For the development of line B it would be recommended to leverage the support of high level experts either from the University or from other research organizations. The main skills which should be involved are:

- quality assessment;
- finite population sampling;
- optimal sampling design;
- model assisted estimation;
- model based estimation;
- Bayesian inference;
- record matching and record linkage.

*Expected outputs*

Line A: Guidelines for consolidation of current survey practices

This line may be divided into the following sub-lines and their appropriated outputs as follow.

| SUB-LINES | OUTPUTS |
|---|---|
| **A.1. Background.** | Intermediate Report on First Reflections on a General Approach for Defining Unbiased Sampling Strategies for integrated Agricultural Surveys |
| **A.2. Overall evaluation of the existing strategies** | Advanced version of the manual A1 |
| **A.3. Possible solution for developing integrated strategies starting from a register related to specific units** | Final report which represents the first version of the guidelines. |

The following aspects are to be covered in the final report:

1. General concepts on indirect sampling techniques (Lavalle, 2006).
2. Alternative approaches for developing survey strategies for integrated agricultural statistics
   - ✓ from a population census frame,
   - ✓ from a farm register frame,
   - ✓ from an area frame.
3. Deep discussion on Pros and cons of the proposed approaches.
4. The concept of optimal sampling in an integrated survey context (Falorsi and Righi, 2008, 2012).
5. Methods for longitudinal updating of the population census frame.
6. Indirect sampling techniques to link the updated population of households to an updated population of farms.
7. Linking from the updated population of farms to the updated population of parcels.
8. Methods for integrated and consistent estimation for households, farms and land use.
9. Use of external administrative data sources for estimation.
10. Techniques for evaluation of sampling variances.
11. Optimal multivariate and multidomain sampling designs.

12. Special topics: oversampling populations with given characteristics (e.g. populations at risk of hunger or poverty).

Line B: Advanced topics for quality, design and estimation

The main focus of the research line B is the conceptual development of the Integrated Survey Framework (IFS) which entails research results and statistical tools useful for the design, implementation and coordination of the IFS.

The coordination phase involves the definition of the informative contexts related to the national survey designs adopted. Also, the quality framework is defined in order to assess the overall statistical production process for the IFS.

The design phase of the research aims at building the theoretical structure of the IFS related to the sampling design and the estimation. This is carried out according to the various informative contexts. In particular, different survey methods are proposed for ensuring appropriate solutions in cases where multiple purpose surveys and multi-surveys approaches are adopted.

The implementation phase concerns issues related to various problems that may be encountered with actual data sets, such as the use of different statistical units in the informative contexts and their integration. Also, it involves the preparation of a user manual for the R software tools developed in the project.

This research line B may be divided into the following sub-lines.

| CODE | SUBLINE | KEYWORD |
|---|---|---|
| B1 | Common research topics | Coordination. |
| B2 | Definition of a quality framework related to the overall process in the Integrated survey framework. | Coordination. |
| B3 | Optimal sampling design which best fits the various national informative contexts | Design. |
| B4 | Integrating survey data which share some common statistical units and types of units using a probabilistic approach. | Implementation. |
| B5 | Multi-surveys approach: guidelines on integrated estimation combining different surveys through dynamic weighting and/or projection estimator | Design. |
| B6 | Multiple purpose approach in periodic surveys: longitudinal estimation in integrated framework via parametric and non-parametric modelling. | Design. |
| B7 | Estimation methods accounting for the uncertainty in the integration of sources related to different units and types of units | Design. |
| B7a | Development open source R software tools for the estimation methods accounting for the uncertainty in the integration of sources related to different units and types of units. | Implementation. |

As far as concerns with line B, most of the research sub-lines focus on specific statistical methods for achieving the integration. Conversely, there are other sub-lines that cover cross-cutting issues.

*Cross cutting issues*:

**B1**: Common research topics (i.e. definition of the different informative contexts and mapping the related different methods).

**B2**: Definition of a quality framework related to the overall process in the integrated survey framework (definition of a common quality framework for integrated statistics and production of guidelines on indicators for driving the choice of a feasible strategy).

*Specific Statistical Methods for achieving the integration in various informative contexts*

**B3:** Optimal sampling design which best fits the various national informative contexts.

The research output of this line defines methods and software implementation for:

- ✓ The selection of the samples for the three target populations which represents the minimum cost solution.
- ✓ Assuring that the expected sampling errors are lower than given thresholds of accuracy, defined for the different variables and for the different target populations.

This is a multi-population, multivariate and multi-domain type of problem.

This line of action could be useful in the cases that correspond to the ***informative context 4*** (***a*** and ***b***). In fact, when an updated MSF exists, the optimal sampling methods provide the maximally efficient solution. In the other situations, i. e. when the MSF is outdated or unavailable, the developed solutions are less efficient than the optimal solution but the most efficient in the specific informative context. The line B3 can also be an input for line A in the ***informative context 2 a*** and in the ***informative context 3 a***

**B4:** Integrating survey data which share some common statistical units and types of units using a probabilistic approach.

The research output of this line define methods and software tools for linking data-sets related to different populations which share some common statistical units. Using a capture-recapture approach, if each data-set doesn't covers the whole populations, the linkage of different data-sets allows one to improve the estimation of unknown population size, and to increase the informativeness of the integrated dataset.

Each potential link is considered together with the corresponding probability of being a true match. This research activity proposes methods that represent the necessary inputs for the subsequent steps of the statistical process like:

- ✓ Sampling from an integrated archive,
- ✓ Estimation: for instance, the estimation of the unknown population size may represents a benchmark for calibrated estimation.

Hence, this line of action could be useful in ***informative context*** of type ***b)***.

**B5**: Multi-surveys approach: guidelines on integrated estimation combining different surveys through dynamic weighting and/or projection estimator.

Usually Statistical Institutes carry out productive processes on the basis of a *stovepipe* (*sp*) model, where all the surveys are carried out independently from each other. This means that separate sampling designs and estimation techniques are applied. If the surveys are aimed to produce similar information about the same populations, each of them can be considered as a component of a unique statistical information system. This approach is characterized by many types of inefficiency, for instance: 1) figures can be inconsistent because of the lack of coherence among surveys; 2) if the estimation strategy does not exploit similar information observed in the other surveys in the system, accuracy can be lower with respect to a model in which the available information are used in an integrated way.

In case of an integrated system, instead of building a *Separate Data Base* (SDB) for each survey, as done through the *sp* model, it is useful to create an *Integrated Data Base* (IDB) merging the target units and the information arising from the different surveys. The structure of the IDB is, therefore, a rectangular table in which the rows identifies the target units included in at least one survey of the system, while columns include the statistical information from the different surveys. The traditional way to build the estimates is to use a separate set of weights for each survey and to perform separate calibration processes (Deville and Särndal, 1992). This approach guarantees internal coherence among the estimates within the same survey but coherence among different surveys is not assured. More recently, Knotterus and Coen (2006) and Knotterus et al. (2003) propose weighting techniques allowing to produce coherent and unbiased (in the model assisted paradigm) estimates for different surveys on the same population. Other methods - proposed by Kim and Rao (2012) under the model-assisted paradigm and by Elbers et. al. (2003) under the model-based approach - which allow producing coherence on different surveys on the same population are based on micro-integration using auxiliary variables common among surveys.

Starting from the methodologies listed above, this line of research is focused on finding suitable strategies for adapting these techniques in the case of surveys on different populations which is the focus of the overall activity of *Integrated Survey Framework.*

This line of action could be useful in ***informative context 1*** (***a*** and ***b***).

**B6**: Multiple purpose approach in periodic surveys: longitudinal estimation in integrated framework via parametric and non-parametric modeling.

The research output of this line defines methods and software tools for dealing with longitudinal and spatial data that may characterize a lot of the informative context for agricultural statistics in developing countries. In fact, in some circumstances the phenomena related to different units have correlation structure that could be taken into account through specific models. For instance, we can suppose that the value of the variable of interest on a person *j* living in the household *i* and working in the farm *k*, $y_{j(ik)}$, can be explained by a variance component model as the sum of an overall mean and two specific factors $\alpha_i$ and $\alpha_k$ related to the household and the farm respectively plus an *iid* random error. In symbols:

$$y_{j(ik)} = \acute{y}_0 + \alpha_i + \alpha_k + \varepsilon_{j(ik)}.$$

The components $\alpha_i$ and $\alpha_k$ may have time/space correlation structures which, if properly predicted, could be used for producing more efficient model based (or model assisted) prediction at unit level. Therefore, the estimate obtained by aggregation of the different predictions, are more efficient and automatically internally integrated. Furthermore, the estimates of the components $\alpha_i$ and $\alpha_k$ could be used for the quality assessment of the integration of independently computed statistics.

This line of activity could be useful in the ***informative context 1*** (***a*** and ***b***).

**B7 :** Estimation methods accounting for the uncertainty in the integration of sources related to different units and types of units.

The research output of this line defines estimation methods which account for the uncertainty in the integration process. As discussed above, the research lines B5 and B6 develop integration through models designed on linked data.

When the linking process is performed without errors, the standard statistical parametric or semi-parametric methods can be used in the estimation. However, in many situations the linkage process is inevitably affected by errors which are managed in a probabilistic way in line B4; in these situations the standard methods cannot be applied and methods which explicitly take into account the uncertainty on the linkage process have to be used in order to avoid the production of biased statistics. This context is further complicated since, in the integrated context, the linkage is performed on common units of populations of different types (households, farms and land parcels). Hence, this line of research can represent both a necessary input for the methods developed by lines B5 and B6 and, at the same time, it can autonomously produce estimates in an integrated survey framework.

Furthermore, the output of this line of research can represent an input also for line B3, when the optimal sampling is performed on a register in which the target variables are estimated.

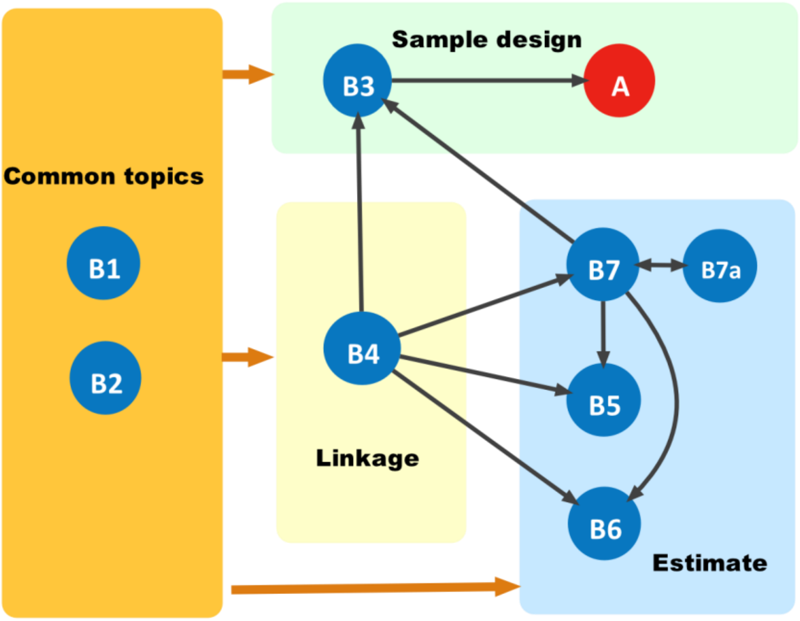
Hence, this line of action could be useful in ***informative contexts*** of type ***b)***.

**B7a:** Development of open source R software tools for the estimation methods accounting for the uncertainty in the integration of sources related to different units and types of units.

In proposing and developing methods and tools that allows the statistical integration in developing countries an important effort should be devoted by each line of activity in providing easy to manage software.

The line of research B7 forecasts the development of sophisticated computational tools for the estimation methods that need a specific commitment and high level programming skills.

The relationships among the different lines of research and their roles in the different informative contexts are illustrated in the figures below.



Relationships between research lines

**Role of each research line in the different informative contexts**

| **Informative context**<br>Defined by (B1, B2) | **Integration of the statistical system**<br>Defined by (B1, B2) | |
|---|---|---|
| | **A : CISS** | **B : CNISS** |
| **1 : Only some surveys** | B5, B6 | B4, B5, B6, B7, B7A |
| **2: Only population census + Surveys** | A, B3 | B4, A, B7, B7A |
| **3: Only agricultural census + Surveys** | A, B3 | B4, A, B7, B7A |
| **4: Agricultural and population census** | A, B3 | B4, A, B3, B7, B7A |

The links between the research questions (see section 1) and the research activities are shown in the following table.

| QUESTIONS | RESEARCH ACTIVITY | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | A | B1 | B2 | B3 | B4 | B5 | B6 | B7/B7a |
| Which could be reasonable formal definitions of the concept of consistency which can be operationalized for the integrated statistical surveys? | * | x | | | | | | |
| How to assure consistency among estimates of different surveys considering both units of the same types and units of different types? In particular: How to interlink the different statistical units in integrated survey strategies? How to deal with imperfect and not updated sample frames? | * | | | | | x | x | |
| How to assess the quality of data for producing integrated agricultural statistics such as to ensure a continuous improvement of the integrated statistics, considering the context of different statistical models (for record linkage and for predicting the variables of interest) and different approaches to inference (model assisted, model based and Bayesian)? | | | x | | | x | x | x |
| How to deal with the main key questions affecting statistical quality of integrated agricultural statistics, such as non-sampling errors (linkage error; nonresponse, etc, non-coverage), optimal sampling designs, and integrated estimation of small domains and of rare characteristics | | | | x | x | | | x |

* = Solution from the available research results
x = Innovative solution

*Timetable of the research activities and expected outputs*

The summary of the expected main outputs or the two research lines is reported in the following table.

| RESEARCH LINE | TIMETABLE OF EXPECTED OUTPUTS BY MONTHS | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 2013 | | | | | | | 2014 | |
| | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 1 | 2 |
| **Line A**<br>Guidelines on existing practices for observing units of different populations in integrated framework | | | | | | | First version | Revised version | End |
| **Line B**<br>Advanced topics for quality, design and estimation | | | | 1 | | 2 | 3 | 4 | End |

1: Report on ongoing or already completed research activities, on the review of the relevant literature, on gaps analysis and remaining methodological issue
2: Output Empirical studies designed and tested on FAO data sets and Open software identification/development.
3: Technical report on the findings and recommendations for possible solutions to methodological issues prepared, peer reviewed, and validated by experts
4: Report on Methodological basis for the guidelines for advanced technical assistance and training.

The specific outputs and the timetable of each sub-line are as follows.

| RESEARCH ACTIVITY CODE | TIMETABLE OF EXPECTED OUTPUTS BY MONTHS | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 2013 | | | | | | | 2014 | |
| | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 1 | 2 |
| A - Support research activity | S* | | | A | | | B | C | End |
| A - Intern | | | | S*** | | | | | End |
| B1 | S* 1_0 | 2_0 | | 3_0 | | 4_0 | End | | |
| B2 | S* | | | 1 | | 2 | 3 | 4 | End |
| B3 | S** | | | 1 | | 2a | 3 | 4 | End |
| B4 | S** | | | 1 | | 2a | 3 | 4 | End |
| B5 | S** | | | 1 | | 2a | 3 | 4 | End |
| B6 | S** | | | 1 | | 2a | 3 | 4 | End |
| B7 | S** | | | 1 | | 2 | 3 | 4 | End |
| B7a | S** | | | 1 | | 2 | 3a | 4 | End |

S* = 17 June Starts the contract
S** =30 June Starts the contract
S*** = 15 September starts the contracts

| OUTPUT CODE | OUTPUT |
|---|---|
| 1 | Report on ongoing or already completed research activities, on the review of the relevant literature, on gaps analysis and remaining methodological issue |
| 2 | Output Empirical studies designed and tested on FAO data sets and Open software identification/development |
| 2a | Output Empirical studies designed and tested on FAO data sets (and Open software identification/development ) |
| 3 | Technical report on the findings and recommendations for possible solutions to methodological issues prepared, peer reviewed, and validated by experts |
| 3a | User manual on the software use with findings and recommendations for dealing with problems that may be encountered with actual data sets |
| 4 | Report on Methodological basis for the guidelines for advanced technical assistance and training |
| 1_0 | Report on the Identification of the statistical mathematical setting of the diverse informative contexts characterizing the Integrated Survey Framework for Agricultural Statistics |
| 2_0 | Selection criteria of data bases for experimentation |
| 3_0 | Choice and validation of the data bases |
| 4_0 | Assessment of survey costs functions in the different informative contexts |
| A | Report on the support to methodological and empirical studies for sampling strategies for the Integrated Survey Framework for Agricultural Statistics |
| B | First version of the guidelines |
| C | Revised and final version of guidelines |
| I | Report on Actual survey practices |
| II | Report on Assessment of Actual survey practices |

*Management of the research*

For the management of the research activities the following actions are defined:

- ✓ *Monthly internal meetings for monitoring the activities.*
- ✓ *Monitoring Reports.* The main decisions of the monthly meetings are briefly summarized in a *report* for the ESS |Director

✓ *Validation of intermediate results* .The validation of results is done by the FAO responsible of the research

✓ *The overall validation of the research results will be* Workshop using peer reviews of experts.

✓ Two *workshops/seminars* for discussing the research outputs with FAO experts are planned.

| TIMETABLE OF THE MEETINGS BY MONTHS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **MEETING** | **2013** | | | | | | | **2014** | |
| | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **1** | **2** |
| First internal team meeting for starting activities | x | | | | | | | | |
| Seminar at FAO for illustrating/discussing the project | | x | | x | | | | | |
| Second internal team meeting for assessing the results of the first phase of the project of activities | | | | | x | | | | |
| Third internal team meeting for discussing the ongoing activities of the empirical stuidies | | | | | | x | | | |
| Fourth internal team meeting for assessing the results of the empirical study | | | | | | | x | | |
| Workshop and peer review of experts | | | | | | | | x | |
| Fifth internal team meeting for assessing the indication of guidelines | | | | | | | | x | |
| Seminar at FAO for illustrating the methodological indication for the guidelines | | | | | | | | | x |

*Note on empirical studies*

Each of the research lines, from B2 to B7, has to develop empirical studies on actual data sets, made available by FAO, which are somehow representative of the situations which may characterize the developing countries The FAO regional offices as well as the FAO Headquarter, may provide the above data sets

The empirical studies are aiming to:

✓ Assess the feasibility of the proposed methodology

✓ Evaluate the pros and the cons of the methodology with respect to the traditional existing methods

✓ Identify the key relevant aspects to be underlined in the guidelines and in the training materials

✓ Develop a prototypal version of the software tool implementing the method

Depending on the specific objectives of each research line, the data set considered in the empirical study may refer to an Country with Integrated Statistical System (CISS) or to a Countries with Non Integrated Statistical System (CNISS) (see section 1)

More in detail, for CISS it is useful to take into account the different operational way with which the Population and the Agricultural Census are interlinked. A lot of good practices illustrating how Population Census can be used for collecting agricultural variables are described in FAO (2012, Chapter, 4). Three main different approaches are proposed.

✓ Inclusion of agricultural items in the population and housing census (e.g. Albania, Nepal, Mozambique, …).

✓ Extended agricultural module in the population and housing census (e.g. Burquina Faso).

✓ Conducting the population and the housing census jointly with the agricultural census (Canada).

The definitions of farm holder and that of head of rural households are different (FAO, 1986, 2010).

The farm holder may be individuated making a specific screening to the head of a rural household.

The farm holders represent a subset of the heads of rural households. The screening is performed in Burkina Faso, while in Mozambique the additional screening is avoided and thus for all statistical purposes the head of a rural household coincides with the farm holder.

In other countries, as for instance Belize and Grenada, the Population Census and the Agricultural Census are conducted independently.

For stating activities, it should be useful to have relevant microdata from the last Population Census and Agricultural Census. Data will be needed for two or three districts for which all records will be required. According to the above analysis, the data could be requested to the following countries:

- ✓ Mozambique,
- ✓ Burkina Faso,
- ✓ Belize,
- ✓ Tanzania.

Other data could be available at sample level, considering the data of the international survey on *Living Standard Measurement Study* carried out by the World Bank (2013).

Finally, in order to assess the methodological proposals with respect to different contexts, could be useful to leverage from simulated data.

The main characteristics of the empirical studies for each research line are summarised below

| RESEARCH ACTIVITY CODE | AIM OF THE EMPIRICAL STUDY | TYPE OF DATA SETS (DS) |
|---|---|---|
| B2 | To calculate and evaluate the proposed quality indicators | One MSF and some survey data selected from it<br>Survey data and metadata of two or more CNISS countries |
| B3 | Evaluate the gain in efficiency of the proposed design | One or two MSF from CISS |
| B4 | Integrate two or more data sets, with some common linkage variables, also referred to different statistical units | Two or more DS from CNISS |
| B5 | To calculate sampling weights which allow to obtain consistent estimates for different statistical units | One MSF and different survey data (selected from it) related to different statistical units The sampling design variables (stratum, PSU, SSU, inclusion probability, sampling weight) have to be available in the survey data |
| B6 | To study the possibility of building:<br>✓ Synthetic DS for the different statistical units with imputed variables<br>✓ DS with microaggregated estimates for Data Warehouses | Two or more surveys with some common variables for estimating models<br>Some data with the geospatial reference |
| B7 | To study how estimating<br>✓ The unknown size of the populations of interest by linking DS referred to different statistical units<br>✓ Quality measures of the estimates accounting for the linking process | Two or more DS from CNISS |
| B7A | To evaluate the software performance and problem with actual data sets from developing countries | Two or more DS from CNISS |

## 5. Budget

The overall budget for this research activity is planned to be 189,460 USD.

The cost for the main research lines are illustrated below.

| Research line | DAYS | TOTAL (USD) |
|---|---|---|
| **Line A**<br>Guidelines on existing practices for observing units of different populations in integrated framework | 170 | 22,760 |
| **Line B**<br>Advanced topics for quality, design and estimation | 590 | 136,700 |
| Workshop and peer review | | 30,000 |
| **Total** | **760** | **189,460** |

## Main References

Antal, E, Tillé, Y (2011) A Direct Bootstrap Method for Complex Sampling Designs from a Finite Population *Journal of the American Statistical Association, 106*(494), 534-543 http://amstattandfonlinecom/doi/abs/101198/jasa2011tm09767

Bramati, MC (2012) Robust LH stratified sampling strategy *Survey Research Methods*, **6** (3)

Bramati, MC (2013) Integrating Business Surveys: guidelines and principles based on the Belgian experience In: N Torelli et al (eds), *Advances in Theoretical and Applied Statistics*, Studies in Theoretical and Applied Statistics, Springer-Verlag, Berlin Heidelberg

Chambers, R; Chipperfield, J; Davis, W; and Kovacevic, M, (2009), Inference Based on Estimating Equations and Probability-Linked Data, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 18-09, 2009, 36p http://rouoweduau/cssmwp/38

D'Aló M, Di Consiglio L, Falorsi S, Ranalli MG and Solari F, (2012), Use of spatial Information in Small Area Models for Unemployment Rate Estimation at Sub-provincial Areas in Italy, Journal Of The Indian Society of Agricultural Statistics **66(1)**, pp 43-53

D'Orazio M; Di Zio M And Scanu M (2006), *Statistical Matching Theory and Practice* Wiley

Droesbeke, J-J, Lavallée, P (1996) La non-réponse dans les enquêtes *Methodologica*, **4**, pp 1-41

ESSnet on Small Area Estimation (2012), Guidelines for the application of the small area estimation methods in NSI sample surveys, *http://wwwcros-portaleu/sites/default/files//WP6-Reportpdf*

Falorsi P D, Righi P (2008) A Balanced Sampling Approach for Multi-way Stratification Designs for Small Area Estimation, *Survey Methodology*, 34, 223-234

Falorsi P, Righi P (2012) Optimal Survey Strategies in the Multivariate Multidomain Context With Multiple Sources of Administrative Information Covering Different Population Subsets In: *Electronic proceedings, of the Seminar on New Frontiers for Statistical Data Collection*, http://wwwuneceorg/stats/documents/201210collhtml, 31 October - 02 November 2012, Geneve: UNECE

Falorsi P, Righi P D (2013) A Generalized Framework for Defining the Optimal Inclusion Probabilities of One Stage Sampling Designs for Multivariate and Multi-domain Surveys *Paper in the refereeing process of Survey Methodology*

FAO (1986), Food and Agricultural Statistics in the context in the Context of a National Information System

FAO (2010), A System of Integrated Agricultural Censuses and Surveys

FAO (2011), Global Strategy to Improve Agricultural and Rural Statistics

FAO (2012a), Actions Plan of the Global Global Strategy to Improve Agricultural and Rural Statistics

FAO (2012b), Guidelines for Linking Population and Housing Censuses with Agricultural Censuses

Gonzalez J M, Eltinge J L, (2010), Optimal Survey Design: A Review *Section on Survey Research Methods JSM October 2010*

Kim J Kand Rao J NK (2012), Combining data from two independent surveys: a model assisted approach *Biometrika*, **99**, 1, pp 85-100

Kott PS and Liao D (2012) Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine *Survey Research Methods* **6**, No2 pp105-111

Kott PS (2006) Using Calibration Weighting to Adjust for Nonresponse and Coverage Error *Survey Methodology*, **32**, No32 pp133-142

Knotterus P and van Duin C (2006), Variances in repeated Weighting with an Application to the Dutch Labour Force Survey *Journal of Official Statistics*, **22**, 3, pp565-584

Langel, M, Tillé, Y (2013) Variance estimation of the Gini index: Revisiting a result several times published *Journal of the Royal Statistical Society - Series A*, *176*(2), 521-540

Lavallée P and Rivest LP (2012) Capture-recapture sampling and indirect sampling *Journal of Official Statistics*, 28, 1-27

Lavallée P (2007), *Indirect Sampling* Springer

Liseo, B ; Torelli N, and Montanari GE (2006), *Metodi statistici per l'integrazione di dati da fonti diverse* Franco Angeli

Liseo, B: Tancredi A (2011) Bayesian estimation of population size via linkage of multivariate Normal data sets *Journal of Official Statistics*, **27**-3, pp 491-505

Lundstrom, S, Särndal, CE (2001)Estimation in the presence of nonresponse and frame imperfections Statistics Sweden, http://eppeurostateceuropaeu/portal/page/portal/pgp_ess/0_DOCS/se/ESTIMATIONpdf

Massiani, A (2012) Estimation of the variance of cross-sectional indicators for the SILC survey in Switzerland *Accepted in Survey Methodology*, 1-13

Matei, A, Ranalli, G (2011, August) *Adjusting for nonignorable nonresponse using a latent variable modeling approach* Paper presented at the Proceedings of the 58th congress of the International Statistical Institute, Dublin

Mecatti, F, (2007), A single frame multiplicity estimator for multiple frame surveys, *Survey Methodology*, December 2007 151, Vol 33, No 2, pp 151-157, Statistics Canada, Catalogue No 12001XPB

Nedyalkova, D, Tillé, Y (2008) Tirages coordonnés d'échantillons stratifiés : méthodes basées sur des microstrates In P Guilbert, D Haziza, A Ruiz-Gazen & Y

Tillé (Eds), *Méthodes d'enquêtes : applications aux enquêtes longitudinales, à la santé et aux enquêtes électorales* (pp 309-314) Paris: Dunod http://wwwdunodcom/sciences-techniques/sciences-fondamentales/mathematiques/master-et-doctorat-capes-agreg/methodes-de-sondage

Nedyalkova, D, Tillé, Y (2012) Bias-robustness and efficiency of model-based inference in survey sampling *Statistica Sinica, 22*(2), 777-794 http://www3statsinicaedutw/preprint/SS-10-238_Preprint%202pdf

Nedyalkova, D, Tillé, Y, Pea, J (2008) Sampling Procedures for Coordinating Stratified Samples: Methods Based on Microstrata *International Statistical Review, 76*(3), 368-386 http://onlinelibrarywileycom/doi/101111/j1751-5823200800057x/abstract

Nedyalkova, D, (2009), Evaluation and development of strategies for sample coordination and statistical inference in finite population sampling, *Thèse de doctorat*,Université de Neuchâtel, 2009, http://docreroch/record/12285?ln=en

Rivest, L-P (2002) Une généralisation de l'algorithme de Lavallée et Hidiroglou pour la stratification dans les enquetes auprès des entreprises *Techniques d'enquètes*, 28 (2), 207-214

Rivest, Land P, Baillargeon, S (2008) Intégration de deux sources de données dans les inventaires forestiers du Québec*Méthodes de sondage : applications aux enquètes longitudinales, à la santé, aux enquètes électorales et aux enquètes dans les pays en développement*, Y Tillé editeur, Dunod, Paris,344-348

Särndal, CE, Swensson, B, Wretman, JH (2003) *Model assisted survey sampling* Springer Series in Statistics

Singh, AC, Mecatti, F (2011), Generalized Multiplicity-Adjusted Horvitz-Thompson Estimation as a Unified Approach to Multiple Frame Surveys, *Journal of Official Statistics*, Vol 27, No 4, 2011, pp 633–650

Tillé, Y, (1994) *Utilisation d'informations auxiliaires en théorie des sondages sans référence à un modèle*, Dir Prof J-J Droesbeke, Institut de Statistique, ULB, Bruxelles

Tillè Y (2006), *Sampling Algorithm*, Springer Series in Statistics

World Bank (2013), http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/EXTLSMS/0,,contentMDK:21588800~menuPK:4196952~pagePK:64168445~piPK:64168309~theSitePK:3358997,00.html .