Food and Agriculture Organization
of the United Nations

Technical Report on
the Integrated Survey Framework

Publication prepared in the framework of the
Global Strategy to improve Agricultural and Rural Statistics

June 2014

Technical Report on the Integrated Survey Framework

Contents

Acknowledgments

The Integrated Survey Framework report was prepared by the FAO Statistics Division as a research activity, performed by the Global Strategy to Improve Agricultural and Rural Statistics under the general direction and encouragement of Pietro Gennari, *FAO Chief Statistician and Director of Statistics Division*, and Christophe Duhamel, *Global Office Coordinator of Global Strategy*. Especially, this Report would not have been possible without the contribution of Naman Keita, *Senior World Bank/FAO Consultant and Interim Research Coordinator of the Global Office*, who kindly allowed us to benefit from his invaluable expertise in guiding, developing and improving the Research.

Piero Demetrio Falorsi, with the coordinative support of Giorgio Alleva and Brunero Liseo, supervised the research team. The Report's parts were drafted by the following authors:

Giorgio Alleva, Piero Demetrio Falorsi, Angela Piersante – *Purpose and scope of the Report*
Angela Piersante, Giorgio Vasselli – *Experiences of integration in FAO*
Giorgio Alleva, Caterina Bramati – *Approaches and contexts for integration*
Brunero Liseo, Piero Demetrio Falorsi – *A common notation for integration*
Giorgio Alleva, Angela Piersante – *Quality standard and guidelines for data integration*
Giorgio Alleva – *Operational aspects*
Brunero Liseo – *Theory of record linkage techniques*
Andrea Tancredi – *Application of record linkage techniques*
Aida Khalil, Piero Demetrio Falorsi – *Direct and integrate observation of different populations*
Paolo Righi – *Optimal sampling*
Aida Khalil – *Simulation study for the assessment of the integrated sampling of different populations*
Paolo Righi – *Experiments on optimal sampling*
Michele D'Aló – *Estimation methods for combining different sources of information*
Fabrizio Solari – *Design-based methods for integrating information*
Michele D'Aló, Fabrizio Solari – *Model-based methods for integrating information*

Additional technical support was generously provided by the Mozambique National Institute of Statistics (INE) and the Tanzania National Bureau of Statistics (NBS), who kindly supplied the empirical data for carrying out the experimentation; Amade Camilo of INE and Titus Mwisomba of NBS, who presented the Country experiences in carrying out the integrated Census, during the Workshop on Methods and Applications for Developing an Integrated Survey Framework (held on 20 January 2014 at the FAO Headquarters); Gaury Datta (University of Georgia), Pierre Lavallée (Statistics Canada), Fulvia Mecatti, who suggested the initial idea of the estimator proposed in the fourth part of this Report; and Caterina Bramati, who fostered obtainment of the research results by organizing the data relating to the experimental results.

Preface

The collection of statistical data by integrating information from different sources, including both censuses and surveys, is increasingly a requirement for the production of statistics, with a view to expand knowledge capacity and to guarantee a higher quality of statistical data.

At the national level, the need to cope with a growing demand for ever more information, interlinked by demo-socio-economic factors, is fundamental for decision and policy making at national and international levels, and entails a greater effort in terms of organization and budgeting for the conduct of large-scale statistical surveys. The efficient use of all available information to produce accurate and timely statistics, in the most appropriate and necessary areas, is therefore a fundamental challenge.

Moreover, several recent development studies acknowledge the importance of agriculture for the national economies of developing countries and its key role for overall economic growth, increased incomes, poverty reduction and the fight against hunger.

Agricultural censuses are not the only source of statistical information on agriculture; however, it remains the largest source in terms of geographical coverage, since it includes most national farms and holdings, and is therefore often used as a main point of reference. Indeed, the main objective of agricultural censuses is to collect or update structural indicators, while other types of survey – such as population censuses, administrative reports, etc. – focus on subjects that do not fall within the agricultural sector itself, but that provide significant information upon it. Therefore, the information is the product of a statistical system, of a combination of various interlinked sources, that must share a conceptual and methodological basis or at least feature mechanisms cabale of fostering complementarity.

Through the Global Strategy to Improve Agricultural and Rural Statistics, the FAO Initiative aims to enable developing countries to build a sound and comprehensive agricultural statistical system, one that is sustainable, well-integrated into the overall national statistical system and capable of responding to current information needs.

This manual presents some contributions produced as part of a research project on data integration, and a proposal for a master sample frame for agriculture, based on FAO's experience in the linkage of population and house censuses and agricultural censuses, and on the experience gathered by countries following the guidelines of WCA 2000.

We were aware that data integration is a **multidisciplinary issue** involving at least three disciplines: *Statistics*, *Management* and *Governance*. Therefore, we chose to limit the scope of the research, to be able to examine some key statistical methods in detail. This should enable the development of relevant means for achieving integration in the various circumstances that characterize developing countries.

PART 1

Introductory Concepts

1

Purpose and Scope of the Report

1.1 BACKGROUND

Framework

Strengthening the capacity of Member Countries' national agriculture statistics systems to produce good quality food and agriculture statistics is one of the FAO Statistics Division's major missions: indeed, in accordance with its mandate, the Division "*assembles, analyses and disseminates statistical data on world food and agriculture; cooperates with member countries in improving the coverage, consistency and quality of the data; provides advice and assistance to Member Governments to develop & improve food & agricultural statistics[…]".*

Pursuant to this mandate, the FAO Statistics Division has conducted a multitude of capacity building activities over the years, through a vast range of channels. These include preparation and dissemination of methodological and technical guidelines; delivery of formal training courses as components of country-level field projects, study tours, and group training through seminars and workshops at international, regional and national level; provision of expert consultations; organization of regional commission meetings, etc.

The FAO Statistics Division deals with national methods for collecting data for agricultural statistics. The main sources are censuses, surveys and administrative records. Member Countries conduct population and agricultural censuses, economic surveys on household income and expenditure, surveys on costs of production, food consumption, nutrition, *ad hoc* topics, etc. These collections can vary greatly among different countries in terms of their objectives, contents, periodicity and the methodologies employed.

These activities can, at times, present inefficiencies, such as scattered data, due to the multiplicity of bodies responsible for statistics production; the same kind of statistics being produced by different bodies; or incomplete statistics and weak technical documentation (metadata) accompanying the production data.

Most surveys are carried out on an *ad hoc* basis with separate budgets and specific objectives, and in response to the most urgent needs; moreover, they are executed by different agencies, without coordination of data collection activities. A number of methodological pilot studies are carried out in different countries in addition to diagnostic, evaluative, or other agro-economic studies, undertaken with specific objectives and, often, restricted geographical coverage.

In addition, in many developing countries, the agricultural sector is very complex and undergoing rapid evolution, due to the simultaneous presence and interlinkages of several different farming systems. The typology of farms ranges from small subsistence family farms to large, modern, market-oriented and highly mechanized systems. Exacerbating this complexity, particularly as regards subsistence farming, mix-cropping, continuous planting and harvesting, combined with a lack of recording systems, pose serious challenges to collecting data on the sector.

In 2010, the United Nations Statistical Commission developed and adopted a new conceptual framework proposed by the *Global Strategy to Improve Agricultural and Rural Statistics*, under FAO's initiative. The framework broadens the concept of agriculture to include forestry, fisheries, and aquaculture and is based on three interlinked dimensions: economic, social and environmental.

The *Global Strategy to Improve Agricultural and Rural Statistics* seeks to enable the retrieval of coherent and comparable data through the use of standards and in-depth analyses across sectors/collections, while at the same time avoiding duplication of efforts, preventing release of conflicting statistics, ensuring the best use of the resource and reducing the burden of responding. For this purpose, it comprises the following three pillars:

- **First Pillar**: Establishment of a minimum set of core data relevant to national and international policy makers, and a framework for the addition of specific national indicators.

- **Second Pillar**: Integration of agricultural statistics into national statistical systems, through implementation of a methodology set that includes a Master Sampling Frame (MSF) for agriculture and an integrated survey framework, and the availability of results in a data management system.

- **Third Pillar**. Ensuring the sustainability of the agricultural statistics system through governance and statistical capacity building.

This manual falls within the second pillar, related to integration. It describes the results of the research activity aimed at developing an *Integrated Survey Framework for Agricultural Statistics*. The strategic objective is to produce statistics that interlink farm characteristics with their relevant households, and then connect these to the land cover and use dimensions. The three different target populations to be interlinked are:

- Agricultural plots (for the environmental dimension);

- Households (for the social aspect);
- Farms (for the economic dimension).

1.1.1 DEFINITION AND PURPOSES OF INTEGRATION

Data Integration is "*the process of combining data from two or more sources to produce statistical outputs*" (UN Glossary). In statistics, *it is a set of methods and procedures aimed at the production and dissemination of joint statistical information at the level of units, or groups of units, arising from a plurality of sources*, such that "*each statistical collection is carried out, not in isolation, but as a component of the national statistics system"* (FAO, 2005).

From a statistical point of view, three main purposes of data integration can be identified:

1. The *estimation of one or more phenomena* based on multiple sources. The advantage lies in the availability of a greater number of observations and/or a greater number of auxiliary variables, and an improved quality of data.

2. The *estimation of interpretative models* of relations between phenomena detected through a variety of sources. The process of bringing together information from different sources paves the way for answering a broader range of questions; through integration it becomes possible to examine the underlying relationships between various aspects of society, thus improving our knowledge and understanding of a particular subject.

3. The construction, updating, or completion of reliable *frames of statistical units of interest*, aimed at the selection of samples, the monitoring of phenomena over time, and the estimation of changes in the composition or survival of the units in the population.

1.1.2 BENEFITS OF INTEGRATION

Experience with Member Countries has shown that the integration of different data sources provides better coverage of certain statistics, for which a suitable solution cannot be found through one population of a single kind of survey only.

For instance, if the estimated data on crop and livestock production originates from surveys based on separate samples, it is impossible to analyze the economic characteristics of farms involved in both crop and livestock production. Likewise, these farms cannot be compared to farms specializing in either crops or livestock exclusively; it is impossible to assess how agricultural production activities affect the farms' well-being and rural households, and to evaluate their economic activity.

An Integrated Survey Framework can enable the comparison of sample unit data across time and across sectors, thus providing a major validation tool for the improvement of data quality in the following ways:

1. **Adding value to the entire statistical data collection and management system**, developing richer databases for detailed and extensive analyses,

meaningful comparative studies and better interpretation of relationships among different phenomena (Hwang et al, 2005). Linking administrative data and statistical data collections from different sectors creates an invaluable source of information for statistical and research purposes, as previously undiscovered relationships can be examined. For example, the integration of datasets, within a Geographic Information System (GIS) framework, enables the combination of agricultural and socio-economic data to reveal disparities emanating from variations in household well-being (Rodgers, Emwanu & Robinson, 2006), or to display spatial-temporal patterns.

2. **Reducing the costs of statistical collection and the burden placed on respondents**, especially the costs of planning and executing regular field data collection. The reduction of the statistical burden of detailed and constant interviews is also significant; this can be reduced or eliminated if surveys are integrated with archives or administrative registries. Different methods for investigating relationships of particular interest, such as conducting a survey, do exist; however, data integration can offer a less time-consuming and less costly alternative.

3. **Increasing the consistency and accuracy of statistical outputs**. Statistical integration promotes common standards in enumeration such as sampling designs, common definitions of variables and data classification and sample frames (Kiregyera, 2001). These improve the quality of National Statistical Systems' data production processes and advance the integration process of different databases, thus yielding the following collective benefits:

    - Increased data coherence, enabling repeated collection of comparable data, in addition to opportunities for data cross-analysis, exchange and re-use;

    - Increased data accuracy by promoting data scrutiny, to ascertain their contents and quality;

    - Improved data quality, in the processes of data restructuring, cleansing, reconciliation and aggregation (White, 2005);

    - Enabling comparative analysis between the same units between time scales;

    - Promotion of good experiences or practices of institutional cooperation among data producers, which can be considered as deriving from data integration.

4. **Better exploitation of common technology, analytical methods, tools and processes**. Integration promotes the use of common tools and processes in statistical analysis (e.g. data matching and record linkage), data storage, and data dissemination. In a statistical organization, integration facilitates the economical use of existing human capacity to prepare, analyze and interpret data for the common good. This in itself reduces the cost of any specialized training required.

1.1.3 LIMITATIONS

There are some limitations that must be considered carefully when conducting a process of data integration.

1. **Variations connected with changes in the data needs or in the survey plan.**

   - *Variation of the type of enumeration units considered in the data collection.* Variation in the needs and variation of data sources lead to changes in the definition of the type of enumeration units. This makes it difficult to decide on the best unit of enumeration to use for data integration. This variation can concern the same country, over time, or different sources with different objectives. For example, if there is information on the income or living conditions of families in a certain African country, recorded over the same period in three separate surveys of different enumerative units (e.g. a survey on agriculture holdings, a census on households in the population and housing, and a survey on dwellings), which enumeration units are brought to the entire ex post information? How can the best, coordinated, integration planning of the three separate surveys be achieved, especially in terms of sample selection and interview protocol (choice of respondents, definitions, etc.)?

   - *Variation of variables collected and variation of the coverage*. Even within the same unit of enumeration, a variation in needs (sometimes a mere variation in the survey's organization) can cause variations in the questions posed (variables) or in the survey's coverage (e.g. sample size). In particular, the variation of variables hinders the process of record linkage (a reduced number of key variables to identify the same unit in the different sources). In addition, the variation in coverage changes the number of units that can be linked.

2. **Variations connected with different definitions of the enumeration units.**

   - *Use of different definitions of local units: agricultural holding and agricultural household.* In developing countries, agriculture is characterized by the existence of a large number of small subsistence farms, with areas cultivated depending upon the manpower available. Although *agricultural holding* has been recommended as the statistical unit, surveys consider *agricultural household* (household in which at least one member operates an agricultural holding) rather than the holding itself, ignoring the aspect of size. This has led to an absence of uniform standards, as varying units such as hectares, acres or other local units are employed. Similarly, data on production is estimated on the basis of local units, without standardization (Gutu, 2001) making it difficult to compare data from different households or regions.

   - *No clear definition of household.* Although the household has been preferred as a unit of enumeration for agricultural censuses or surveys, in most African economies a major challenge remains, as no clear definition of the term has been given. This is irrespective of the fact that a household in the African context can be single, nuclear or extended. Surveys using different definitions are likely to generate inconsistent data that are difficult to integrate.

3. Factors that make it difficult to compare units in agriculture over time and space.

- *Variation in production methods and farming techniques.* The bulk of agricultural production depends on a wide variety of farming techniques (Gutu, 2001). Hence, no two holdings can be correctly compared and integrated. As a result, sampling units based on factors other than ownership, e.g. agricultural zonation, could be devised for comparison purposes.

- *Choice of timing of data collection.* Annual surveys are usually based on recalling the power of interviewees, despite significant cases of illiteracy (Keita, 2004). The time lapse between crop growth (seasons) and annual surveys (CBS, Kenya) has led to inaccurate estimations and subjectivity on the part of respondents even when considering similar sampling frames; this complicates statistical comparisons.

- *Varying sampling methods*. African institutions are faced with resource limitations, which make it impossible to conduct regular agricultural censuses. Instead, sample enumeration is common (Keita, 2004). However, different sampling designs or methodologies (random, stratified and multi-stage sampling) are applied in surveys depending on the research goals and available resources. Varying methodologies lead to varying sampling units, as well as differences in obtaining data within the same geographic region or population.

- *Difficulty in tracing target households for Migration and Nomadism*. Like agricultural systems, pastoralism is a major economic activity in many African countries and involves nomadism or migration in certain periods of the year, depending on resource availability. However, although migration is a major productive technique used to overcome numerous environmental constraints (Tadingar, 1994), it constantly disrupts data collection, since target households cannot be reached.

- *Impact of diseases on data collection*. Several diseases, including AIDS, have devastating effects on African populations over short timeframes. These diseases have great adverse impact on agricultural activities and data collection, especially when a household constituting a sampling unit for a panel/time series study is wiped out, such that no further data can be obtained. Change in the sampling unit thus becomes inevitable, to avoid the arising of data gaps that are unsuitable for integration.

- *Impact of cultural barriers.* Low levels of literacy and lack of formal education create communication barriers with outside society. Even in settled conditions, some societies follow customs that bar outside men from addressing women (Mogoa and Nyangito, 1999); therefore, when the heads of the household (men) are not available, it is difficult to use these households for data collection.

- *Impact of insecurity and conflicts.* On one hand, situations of insecurity such as armed wars and conflicts drive people out of their living areas; on the other, these situations make data collection dangerous for the people involved. These issues impel adjustments, because these factors involve targeted units of enumeration which can sometimes not be reached.

4. Limitations posed by institutional factors.

- *Policy and legal constraints.* Concerns relating to integrated data systems often regard legal constraints and copyright issues. Problems frequently arise in connection with access to datasets that contain personal information, as this may infringe national confidentiality regulations. The commercial value of data can also represent a constraint, which can sometimes be overcome with the payment of a fee (Jones and Taylor, 2004). Moreover, there are various different national data policies: some countries allow access to their data freely, while others enforce high-confidentiality or stringent data access policies that discourage casual and unqualified data requests (FAO, 2003).

- *Limited skilled manpower of National Statistical Systems (NSSs)*. Many African countries have weak NSSs, characterized by limited skilled manpower. This poses a challenge for integration, as there is no effective capacity to generate new statistical data from existing data or to collect, analyze and manage data from different sources for use in decision-making; this is particularly true when dealing with quickly-evolving scientific methods and technologies for data storage and management (FAO, 2003).

- *Weak coordination and collaboration between producers of statistical information.* Gutu (2001) notes that in many African countries, agencies collect, compile and disseminate statistical information with and without reference to the NSI. Agricultural censuses, surveys and other statistical inquiries are undertaken in isolation, without understanding and co-ordination between statistical data producers. As a result, data inconsistencies among producers are common, due to the different methodologies employed or units of enumeration used. This creates the need for massive work to refine such data, if they are to be integrated.

5. Limitations posed by technical factors.

- *The choice of specific integration technologies.* The limited availability of skilled manpower is complicated by the existence of several data integration technologies, which makes it difficult to decide which one is the most appropriate (Meta Group, 2004), or which data integration technology must specialists use.

- *Poor IT infrastructure.* Many African countries have poor infrastructure for statistical integration, owing to ineffective IT connectivity (White, 2005). This limits data sharing and communication of results. Hence, countries face challenges in adopting the rapidly evolving techniques for data collection, processing and analysis, and storage and dissemination technologies, especially in areas such as GIS, satellite data processing and modelling (FAO, 2003).

- *Data inconsistency and poor quality*. Since most existing data were collected in the absence of common standards, definitions or classifications, there is great inconsistency among different data sources (FAO, 2003), and others are of poor quality (White, 2005) because of the methods used for their collection. Other data yet were compiled in inadequate formats and presentations, which limits integration. For example, most data lack the metadata necessary for providing

summary information, such as the sources and origin of data, methods of access and processing, fitness for use, adjustments made and their impact on data integrity (Polach and Rodgers, 2006; Guptill, 1999). Moreover, due to the absence of standards for data collection, some data are not classified according to standard categories (numerical, qualitative, discrete, continuous, categorical).

In conclusion, the most important requirement for the success of an integration process is the use of the most suitable unit of enumeration, standards, definitions, and classifications. Furthermore, it is necessary to devise prudent ways to overcome existing constraints, especially those related to technological development and associated capabilities.

1.2 SCOPE OF THIS TECHNICAL REPORT

The **success of a data integration project** depends on the particular methods for integrating data sets, but also involves:

- the integration of metadata
- efficient organization
- effective cooperation between organizations.

Furthermore, integration is related to the overall chain of the statistical production process and can occur **at any stage**:

- before data collection
- during field operations or data processing
- during the data dissemination process.

In other words, data integration is a **multidisciplinary issue** that involves at least three disciplines:

- Statistics
- Management
- Governance

In this context, the compilation of a technical report aiming to cover thoroughly all topics and methods related to data integration would be a massive task. Therefore, we have decided to limit the report's scope, to be able to examine some key statistical methods in detail. This should enable the development of some relevant cases for achieving integration in the different circumstances that can arise in developing countries.

We have considered the following three cases of data integration, which represent increasing levels of statistical maturity:

- **Case A**: Integration through the introduction of a multi-purpose survey (low level of statistical maturity)
- **Case B**: *Ex post* integration of data from different sources (medium level)
- **Case C**: Planned data integration (highest level)

A detailed description of these cases is available in Part 1, Chapter 3. For current purposes, it will be briefly noted that:

- Case A is typical of many developing countries that have the statistical capacity to conduct not more than one survey on the agricultural sector, such that it would be reasonable to use that survey to obtain information on all target populations of interest (agricultural plots, households, and farms). An example of this strategy is the *Agriculture Module* of the World Bank's *Living Standards Measurement Survey*.
- The most frequent situation in developing countries is Case B, in which each survey is carried out on an *ad hoc* basis, with separate budgets and specific objectives, to cater for the most urgent needs. However, these surveys are executed by different agencies without any coordination of data collection activities. In these circumstances, it is possible to achieve integration only *ex post*, attempting to link (or match) the surveys and to attain integration by using specific estimation methods.
- Case C denotes the most mature stage of a statistical system, and represents a benchmarking theoretical situation. Surveys can be conducted separately, having been designed (and planned) to ensure the integration of the overall statistical system. The statistical methods used in this case are not different from those used in Cases A and B. However, they are planned and combined with a view to achieve the consistency of the statistical information disseminated.

For example, a feasible planned strategy could be based on:

1. conduction of a limited single multipurpose survey (Case A) to establish statistical models that link the variables of the different target populations; and
2. execution of the different surveys, taking care during the planning phase to collect, in each, some of the variables that enable use of the statistical models examined in Step 1 (according to methods useful for Case B), and to achieve integration through their judicious application.

There are certainly many other strategies that can be adopted to achieve a planned integration. It is not theoretically possible to identify *a priori* the best strategy, or even all solutions possible.

The aim of this Report is to examine in detail, with specific regard to integration, the statistical methods that constitute the basic tools for achieving integration in each of the three cases seen above (A, B, C). These methods concern:

- record linkage
- sampling

- estimation

The focus will be on Cases A and B, since these represent almost all developing countries. These Cases will be seen in further detail in specific Parts of this Report. However, the methods illustrated for Cases A and B enable a planned integration, if they are organized into a suitable strategy.

The study will focus on the following issues/questions:

1. How can different statistical units be interlinked in integrated survey strategies?
2. How can different statistical units be linked, taking into account the uncertainty on the appropriate linkage for producing unbiased and robust inferences?
3. How can imperfect and not updated sampling frames be dealt with?
4. How can consistency among estimates from different surveys be assured, when considering both units of the same type and units of different types?
5. How can the quality of data for producing integrated agricultural statistics be assessed? That is, how can a continuous improvement of integrated statistics be ensured, considering the contexts of different statistical models (for record linkage and for predicting the variables of interest) and the different approaches to inference (model-assisted, model-based and Bayesian)?

Finally, we emphasize that each process of data integration entails the harmonization and reconciliation of multiple sources. This requires feedback loops between the existing documentation of variables, the data analyses and the methodology. Furthermore, it highlights the need for a specific ***Quality Framework*** for integration, that must take into account several critical factors: the quality and coherence of sources, the explanatory power of common variables, the matching/imputation methods applied, the methods used to compute estimates based on the datasets matched, etc.

1.3 GUIDELINES FOR READERS: STRUCTURE OF THE TECHNICAL REPORT

The Report discusses how to deal with the main questions affecting the statistical quality of integrated agricultural statistics, such as non-sampling errors (linkage error; non-response, non-coverage, etc.), optimal sampling designs, and integrated estimation.

The Report is organized in five main Parts. Parts 1 and 2 investigate common topics, while Parts 3, 4 and 5 provide a detailed analysis of the key statistical methods. In particular:

- **Part 1** provides a general overview of the statistical issues concerning data collection, the benefits of an Integrated Survey Framework, and a description of the FAO initiative through the Global Strategy and its experiences with countries (Chapter 2). Chapter 3 focuses on the different contexts and Cases of data integration, detailing the benefits and risks of each. Chapter 4 introduces the symbology used in the final three technical parts of this report.

- **Part 2** illustrates the quality framework for integration. Chapter 1 presents quality standards and guidelines for data integration, in terms of revisited best proposals from NSIs or international organizations. Chapter 2 identifies general aspects of the quality assessment of data integration and operational factors.

- **Part 3** discusses record linkage methodologies. After an introduction to record linkage and a presentation of the classical and Bayesian approaches, some solutions for making inferences using linked data are illustrated. Finally, methods for linking different statistical units are shown, taking into account the uncertainty surrounding linkage for the production of unbiased and robust inferences. Record linkage enables data integration for Case B, in which the various surveys can be linked through certain matching variables.

- **Part 4** examines Case A (the single multipurpose survey) in depth. Some new methodologies for the integrated observation of different populations are presented. This part discusses how to interlink different statistical units in integrated survey strategies and how to deal with imperfect and not updated sampling frames. Furthermore, the problem of Optimal Sampling is addressed.

- **Part 5** deals with Case B (*ex post* integration). This Part shows how to achieve integration by means of estimation techniques, starting from separate surveys, and resorting to model-assisted and model-based approaches. Essentially, this Part demonstrates how to assure consistency among estimates from different surveys, considering either units of the same type or units of different types.

Case C (planned integration) can be attained through a strategy based on a joint use of the statistical methods illustrated in Parts 3, 4 and 5.

The Parts are connected as illustrated in Figure 1 below. Specifically, **Parts 1 and 2** discuss the common topics and introduce the basic concepts for the development of **Parts 3, 4** and **5**.

The **Record Linkage** presented in **Part 3** is crucial for:

- **Sampling (Part 4)**, since: (*i*) it enables construction of the Sampling Frame in which to record the linkages among the different target statistical units; (*ii*) the construction of an optimal sampling design must adequately consider the *matrix with the probabilities of correct linkage* that is a standard output of the linkage procedures and (*iii*) the linkages recorded in the Sampling Frame and derived from the record linkage process foster checks of the linkages, among the different statistical units, that can be collected during the data collection phase.

- **Estimation (Part 5)**, because: (*i*) the statistical models for imputing the target variables can be applied to the auxiliary variables of the Sampling Frame (obtained as a result of the record linkage process); (*ii*) to assure the robustness and un-biasedness of the statistical models used for estimation, it is necessary to estimate the model parameters, taking into account the matrix with the probabilities of correct linkage.

Finally, there is a relation between **Sampling** (Part 4) and **Estimation** (Part 5), since the inclusion probabilities that constitute the basic tool for estimation in the design-based approach are defined in the sampling design phase.

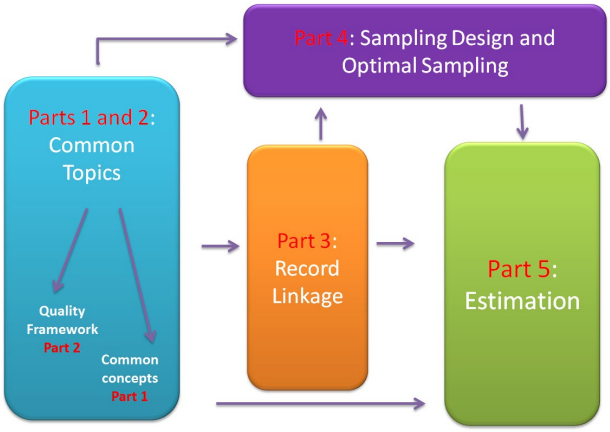

FIGURE 1.1. Relations among the Parts of the Manual

In addition to the theoretical work, some applications to existing data have been performed, to assess and demonstrate the potential of the statistical methods suggested. Thus, some countries were asked to collaborate by providing both survey and census data, where available. Three National Statistical Institutes expressed their interest to assume an active role in the exercise: INE (Instituto Nacional de Estatistica – Mozambique), NBS (National Bureau of Statistics – Tanzania) and ISTAT (Istituto Italiano di Statistica – Italy).

All NSIs were asked to provide the most recent data from Population Censuses and Agricultural Censuses, where available, or from Agriculture Surveys. The data include variables relating to individuals: the name-code of residence, household identifier, demographic characteristics (gender, age, identifier of the head of household), professional status (employment, education, etc.) and identification as a potential agricultural production unit.

These data constituted the basis of the experiments conducted during the research activity. To ensure uniformity in presenting our results, and to limit the size of this report, we focused on the data released by the INE, which provided two datasets: one relating to the 2007 Population Census as restricted to three districts of the Gaza province; and the second to the 2009-2010 Livestock and Agriculture (CAP) Survey, restricted to the same districts.

Therefore, Parts 3, 4 and 5 the Methods are presented and tested according to the above, and the results of the empirical studies, mostly related to the INE datasets, are discussed. The software (and related tutorials) used for the applications and data are available on request.

2

Experiences of Integration in FAO

2.1 INTRODUCTION

In food and agricultural statistics, data collection methods vary from country to country. However, among the methods most commonly adopted, are periodic agricultural censuses and surveys based on sampling methods.

The agricultural sector in developing countries is characterized by the existence of several farming systems within the same country and great differences between countries. The 2010 FAO World Census of Agriculture defines two categories of agricultural holdings:

- holdings in the household sector, operated mainly by household members, and
- holdings in the non-household sector, which include a wide variety of units: corporations, governments and semi-governmental institutions, etc.

However, this dual categorization does not fully reflect the complex reality of the farming systems. Indeed, both types of holdings usually comprise highly diverse units in terms of:

- size and level of production: from small subsistence farms to very large farms
- purpose of production: self-consumption and cash crops
- level of mechanization: simple tools and modern equipments
- type of management: no record keeping and modern methods
- highly diverse holder characteristics, especially the level of literacy

There is a wide range of possible holdings, each of which presenting a different combination of these characteristics. The bulk of agricultural production comes from the household sector, with a large number of smallholder farms scattered across the country operating small parcels for self-consumption.

Data collection systems must adapt to the evolution of agriculture, and exploit emerging technologies to respond to the increasing data needs in a cost-effective manner.

One of the main objectives of the Global Strategy research activity is to identify the most appropriate master sample frame for conducting an integrated survey, which will enable integration of agriculture into national statistical systems and will be the foundation for all data collections based on sample surveys or censuses.

Moreover, experience shows that collecting limited and well-defined agricultural data during population and housing censuses can make a substantial contribution to the construction of an efficient sample frame for agricultural censuses and surveys in many developing countries. Indeed, agriculture is often one of the most important sectors of national economies, and in some regions the majority of households are engaged in agricultural production activities such as cultivating crops, raising livestock or growing vegetables on small plots surrounding the house. The close relationship between these agricultural activities and the population characteristics recorded in population and housing censuses means that, in many countries, there is a strong case for including agricultural items in population and housing censuses.

In 2012, in collaboration with UNFPA, FAO prepared the Guidelines for Linking Population and Housing Censuses with Agricultural Censuses, in line with the Global Strategy's objectives. This technical document provides countries with practical guidelines on how to coordinate and link the two censuses, and can play a key role in a cost-effective census strategy.

WCA 2010's new modular approach is based on two main components:

- a complete enumeration of only 16 data items, conducted as a core module; this provides a restricted range of key structural items, important for national policy-making, international comparisons, the construction of sampling frames (in turn used to collect more detailed data in specific and country-relevant modules) and the analysis of data at narrower geographical or other levels.

- one or more census supplementary modules, conducted on a sample basis at the same time as or immediately after the core census module, to provide more detailed structural data or data not required at lower administrative levels. The sample for the census supplementary modules is selected on the basis of sampling frames from the core census module. The supplementary data items correspond to the core data items of the WCA 2010 census, and, if relevant and feasible, they are always collected in the agricultural module. Figure 2.1 below illustrates the relationship between the two censuses.

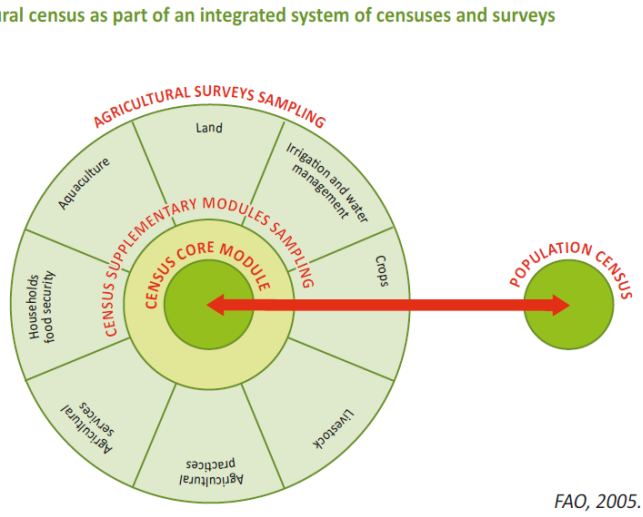

FIGURE 2.1. The Linkage Approach, with Population and Agricultural Censuses

In creating the linkage between the basic units of the two censuses, the fundamental aspect is the identification of **farm households**. This enables a one-to-one correspondence between an agricultural holding and a household with own-account agricultural production activities. These own-account agricultural activities, performed by members of a given household, are usually undertaken under a single management. Therefore, in many countries, agricultural holdings are defined as equivalent to households with own account agricultural production. This approach is considered to have several benefits, including simplifying the identification of the holding, facilitating the linkage with population and housing censuses, and enabling analysis of household characteristics.

The relationships between the concepts of household, farm and agricultural holding are defined as follows:

- The **agricultural holding** of the agricultural census is "an economic unit of agricultural production under single management comprising all livestock kept and all land used wholly or partly for agricultural production purposes, without regard to title, legal form, or size".

- The **household** of the population census "is based on the arrangements made by persons, individually or in groups, for providing themselves with food or other essentials for living. A household may be either (a) a one-person household, that is to say, a person who makes provision for his or her own food or other essentials for living without combining with any other person to form part of a multi person household, or (b) a multiperson household, that is to say, a group of two or more persons living together who make common provision for food or other essentials for living. The persons in the group may pool their incomes and may, to a greater or lesser extent, have a common budget; they may be related or unrelated persons or constitute a combination of persons both related and unrelated" (UN, 2008, p. 102).

- The **farm household** is the fundamental item that must be included in the population census through a questions model, which includes minimum core data items for identification. These items are used to develop a frame for agriculture censuses and surveys, and to create tabulations linking agricultural activities to household characteristics.

*A **farm** is where one or more of the household members are engaged in agricultural production. What is of importance is the household, rather than the dwelling unit. On one hand there may be more than one household in a single dwelling. On the other hand one household may also consist of extended families making common provision for food and occupying more than one dwelling. In other cases, different family units live in separate dwellings but have a common head, as in polygamous unions.*

The relationship between the household and the number of holdings is determined by the **management units** within the household. A management unit consists of the individuals who make the management decisions concerning particular activities. To qualify as a separate holding, the household (or the individuals within the household) must be managing the activities.

However, it is difficult to apply the integration approach in linking household data to holding data when this has not been planned in advance (case of one-to-many and many-to-one); therefore, it is necessary to define the items to be included during the census planning stage. The identification of farm households in the population census can thus provide a framework for selecting households during the agricultural census (list of agricultural households).

2.2 AGRICULTURAL SAMPLE FRAME

A major innovation introduced by the WCA 2010 is the recommendation that a complete enumeration of a very limited number of data items be conducted as a core module, with the use of sampling to collect more detailed data in specific and country-relevant modules. Sampling is a key element in this new approach, in which the availability of an effective sampling frame becomes crucial.

Data collected during the population census provide information for a list frame and supplementary information for area frames, enabling the construction of an effective sampling frame for agricultural censuses and surveys. The area frame can also be expanded with a list of enumeration areas, indicating the number of agricultural holdings.

The WCA 2010 recommends including the questions for identifying farm households in population and housing censuses. It is still possible to use the household frame from the population and housing census as a starting point for the list frame of the household component of the agricultural census.

The case of Nepal presents a good example of frame building for agricultural censuses: "*The proposed sampling design was a multiple frame (two frames, in fact), with a list*

frame for the Private Large Scale and Institutional Farms (PLS&IF) and an area frame for all other household-based holdings. The list frame, which was only part of the totality of all Private Large Scale and Institutional Farms, was to be completely enumerated.

For smallholder farms, the sampling procedure designated districts as strata. In each stratum, a sample of Enumeration Areas (EAs) was then selected as primary sampling units (PSUs) in the first stage. A sample of agricultural households was subsequently selected from each sample EA as second stage units (SSUs)." The Guidelines provide details on the method used, as well as information from other country experiences on using the agricultural information collected in population censuses to choose the survey design (see Mozambique) or to determine the sample size (see Uganda).

2.3 FAO RECOMMENDATIONS ON COORDINATING LINKAGE

In the Guidelines, FAO encourages countries to examine all aspects of the planning of both population and housing and agricultural censuses, such as:

1) **Use of common concepts, definitions and classifications.** This was featured in previous agricultural census programmes, and is again strongly recommended in WCA 2010.

2) **Sharing field materials**. The field systems for the two censuses can usually be coordinated, for example by using the same enumeration areas (EAs) and maps for field work. It is recommended that countries fully explore these possibilities when planning their census operations.

3) **Using the data from population and housing censuses as a frame for agricultural censuses.** FAO encourages countries to use the household lists from the population and housing censuses as a frame for agricultural censuses, if appropriate. Problems with lists becoming obsolete and differences in the statistical units applied for the two censuses (households, farm households and agricultural holdings) are discussed in detail in the Guidelines.

4) **Data related to agriculture found in population and housing censuses, and their possible use in agricultural censuses.** FAO demonstrates how standard population and housing census data on occupation, industry and employment status can be used to identify farm households. The conceptual shortcomings are highlighted, and countries are advised to consider the actual extent to which these data are useful in the agricultural context.

5) **Collecting additional agricultural data in the population and housing census**. It is suggested that countries consider including additional agricultural topics in the population and housing census, to enable development of frames or compilation of tables.

6) **Linking data from agricultural censuses with population and housing censuses.** FAO encourages countries to link data from population and housing censuses with data from agricultural censuses, wherever possible.

7) **Conducting the two censuses as a joint field operation**. The FAO Guidelines outline how data collection for the population and housing and agricultural censuses can be carried out as a joint field operation.

3

Integration: Approaches and Contexts

3.1 INTRODUCTION

Data integration is broadly defined as the *combination of data from different sources on* the *same or a similar individual or unit*. This definition includes linkages between data arising from statistical surveys, or from administrative sources.

Other terms used to describe the process of data integration include '*record linkage*' and '*data matching*'.

An important aspect is the level of data integration. It is possible to distinguish between *micro-level* and *macro-level integration*.

- *Micro-level* integration aims to link the information from one unit:
    i. to a different set of information on the *same unit*; or
    ii. to information on another unit having the same characteristics.
- *Macro-level* integration aims to use and compare collective statistics on a group of units (individuals or regions).

The literature focuses on micro-level data integration of type (i), i.e. the linkage of records likely to belong to the *same* individual or unit.

Correct definition of the statistical unit is crucial. The *Statistical Unit* is the unit of observation or measurement for which data are collected or derived.

The following list provides examples of standard statistical units that have been defined by National Statistical Organizations (NSOs):

- Person; Economic Active Person; Head of Family; Worker; Agriculture Holder and Sub-Holder;
- Census Family; Economic Family; Household; Farm Household;
- Dwelling;

- Land parcel; Location; Area; Census Enumeration Area Unit (EA);
- Enterprise; Establishment; Agriculture Holding and sub-holding; Land tenure.

Integration can occur at any stage of the statistical process:

- **Before data collection**: for example, by adopting international standards for the definition of variables and using common frames of reference, or adopting sample designs created for data integration.
- **During field operations or data processing**: e.g. by sharing field materials, adopting international protocols for interviews, linking data, or using sound methodologies to estimate variables and models based on different sources.
- **During data dissemination**: e.g. by taking account of common interpretative schemes, domains and indicators for the comparison of data over time and space; or building informative systems that add and link information from different statistical or administrative sources.

3.2 TYPES OF DATA INTEGRATION AND SCENARIOS

Integration is generally based on a procedure that merges information originating from multiple surveys or archives. This leads to an increase of information in terms of:

- units of analysis
- variables
- temporal occasions (panel data or rotated samples)

A standard case is the study of a specific population, with information collected at the same level of aggregation in different surveys over time. Here, the information to be integrated may or may not overlap. The units and variables of independent surveys on the same type of enumeration units, and therefore in the integration, feature, respectively, a certain number of repeated units and variables. Figure 3.1 illustrates an example of a simple integration of only one of the three elements.

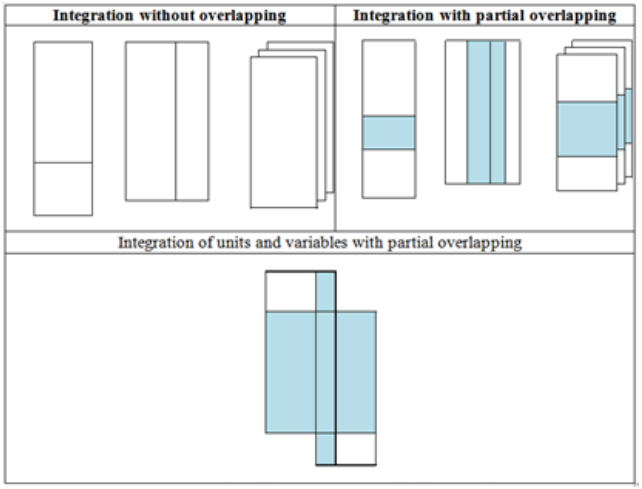

FIGURE 3.1: Example of integration of different types of information

Some questions arise. What is the relationship between the two databases? Why are we interested in the results of the integration process? Should we consider the units of both databases, all units, or only those belonging to one of the two databases?

Considering two datasets, from a theoretical point of view it is possible to examine three different relationships, i.e. three different ways in which the datasets to be integrated relate to each other (New Zealand Statistics, 2006).

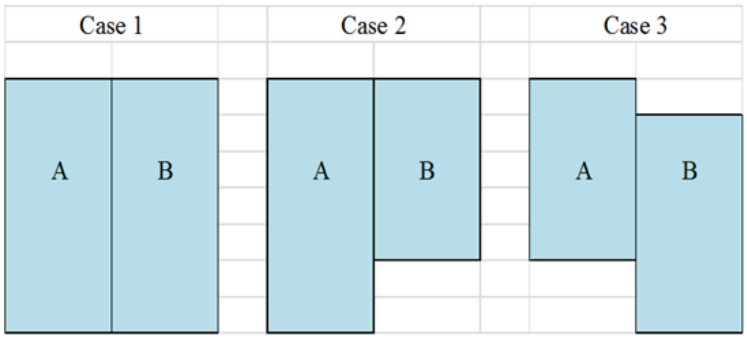

FIGURE 3.2: Cases of integration of two data sets

Case 1

Every unit in dataset A is also in dataset B and vice versa.

Case 2

Every unit in dataset B is in dataset A, but there are individuals that appear only in dataset A.

Case 3

Some units appear in both datasets A and B, and other units appear in only one of the two datasets.

However, the information available is not perfect, and in both datasets there is an unknown fraction of:

- duplicated units
- omitted units
- missing data
- errors in the data
- timing differences between the two datasets

To mitigate the above issues, the following principles must be considered:

a) the input files (the dataset) should be carefully prepared before commencing the linking procedure.

b) the choice of the integration procedure to be applied should be based upon the reliability of the quality of the datasets.

Furthermore, it must be considered that a process of integration may yield several different outputs. For example, in relation to the above cases, the following integrated information is worth considering (see Figure 3.3 below):

- *the intersection of the two databases*: in Case 1, we expect to obtain combined information on most of the units; in Case 2, the extent of the intersection depends on the ratio between the number of units in B and the number of units in A; in Case 3, the extent of the intersection is a function of the number of units that belong to only one of the two datasets.

- *the union of the two databases*: in Case 1, the union is expected to match the intersection of the two databases. In Cases 2 and 3, the union can take into consideration many more units than the intersection; this implies that in certain units of the integrated dataset, some information is missing.

- *only database A, or only database B*: in this case, the additional information arising from the integration of the two databases, especially from the units of one of the two databases, is of relevance. Thus it is possible to obtain information on the units that represent the intersection; however, data for the other units of interest are missing.

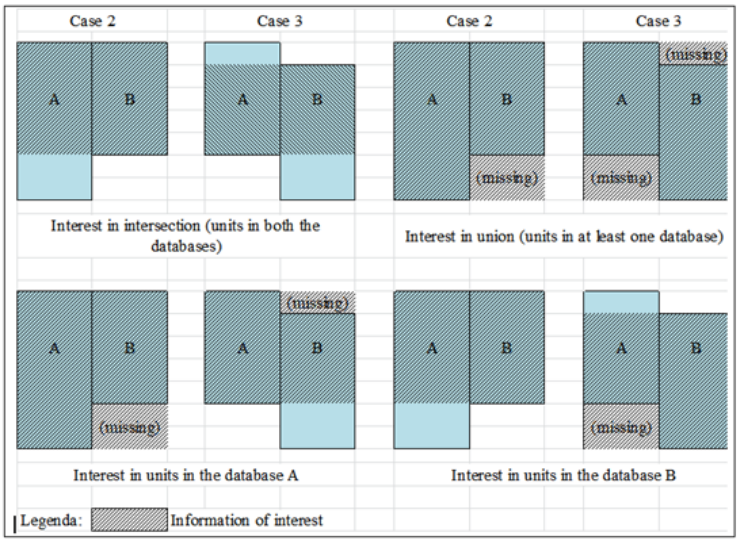

FIGURE 3.3. Different interests in the outcome of integrating two data sets

3.3 LINKING THE INFORMATION COLLECTED ON DIFFERENT TYPES OF UNITS

This research seeks to address the issue of linking statistical units that are not necessarily of the same type or level of aggregation. This is the case of data integration performed on the basis of different surveys considering different types of units, that can therefore be considered a case of integrating units belonging to different populations and different clusters of units.

In particular, the issue addressed concerns the integration of standard units of investigation in the food and agricultural field, especially:

1. *Individuals*, represented by
   - Members of the household
   - Farm holders
   - Workers in agricultural holdings, enterprises or establishments
2. Households or Farm households
3. Land parcels or other geographical or administrative units (Census Enumeration Areas, localities, provinces, etc).
4. Businesses, represented by:
   - Agricultural holdings
   - Farm Households and Non-household Farms
   - Cooperatives, Companies, Governmental companies
   - Enterprises

- Establishments

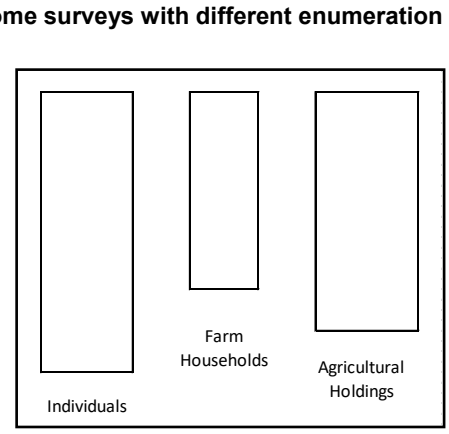

FIGURE 3.4: Example of Datasets with Different Types of Units

The units involved in data integration are illustrated in Figure 3.5 below.

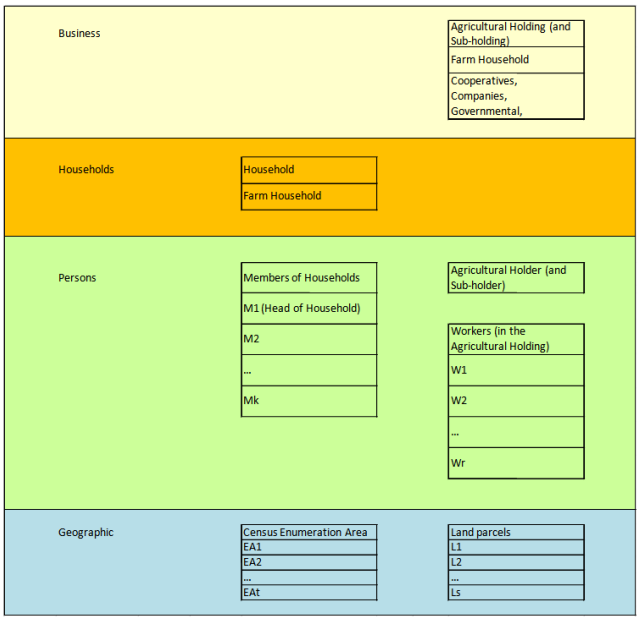

FIGURE 3.5: Units and populations involved in a process of data integration

A data linkage process between different types of units must consider the following cases:

- One-to one matching
- Many-to-one matching
- Many-to-many matching.

One-to-one matching

In this type of integration, one record of a database links to only one record of another database. This is the integration applied when one seeks to link information collected by a census, sample survey or administrative archive on the same individual or the same agricultural holding.

Many-to-one matching

In a many-to-one match, a record from one database is linked to more than one record in another database. A common example of the use of many-to-one matching is geocoding (grouping of addresses of companies or households according to regions or provinces), or the classification of persons according to their households, or students according to their schools.

Many-to-many matching

This case is similar to many-to-one, but it is possible for records on both databases (and not only for one) to be linked to more than one record on the other. This is rarely applied by NSOs.

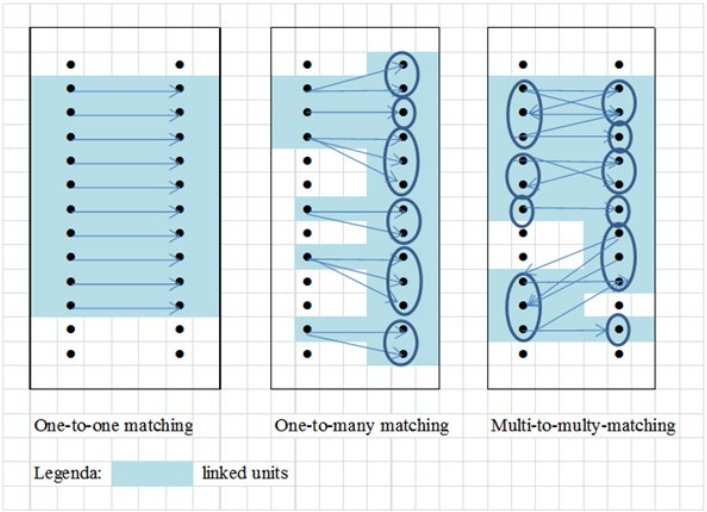

FIGURE 3.6: Examples of Different Forms of Data Linkage

The relations between the units of each pair of populations can be represented by *linking matrices*. This is the strategy considered in this research, and is presented in further detail in other parts of this report.

An important case is the link between statistical units in agricultural and population censuses. In FAO's conception, these sources constitute the pillars of an *integrated agricultural statistics system*.

The primary statistical unit for population censuses is the household, while for agricultural censuses it is the agricultural holding. This is the main issue that must be considered when coordinating the activities of the two censuses; the problems that arise have different solutions, depending on whether they occur in the household sector or in the non-household sector.

In the household sector, the agricultural holding and household units are usually closely related, but there are differences that can hinder coordination between the two censuses (for example, two or more holdings within a single household, or two or more households operating a single holding).

It is much more difficult to link information from the two censuses for the non-household sector, which comprises cooperatives and governmental enterprises. Equating the agricultural holding unit with the household unit is not essential for relating the two census activities. Nevertheless, countries sometimes define the agricultural holding as equivalent to the household to simplify the agricultural census field procedures (FAO, 2005).

3.4 SOME ADDITIONAL REQUIREMENTS FOR DATA INTEGRATION

It should be recalled that data integration also implies the following actions:

- **Integration of metadata**. It is necessary to consider, for example, the definitions and classifications adopted, and the methods for collecting information (in terms of unit selection and measurement tools), and for information processing. The availability of metadata for all sources to be linked is an essential prerequisite for the creation of an integrated database, capable of assessing source quality and therefore gain benefits;

- **Efficient organization, trained personnel, and secure budgetary allocations over a period of years**. Efficient organization also requires solid cooperation between the users and producers of statistics.

- **Cooperation between organizations responsible for surveys and archives**. Integration will ensure compliance with the applicable legislation, in particular that concerning data protection and privacy. The various statistical activities are not always all within the competences of a single government institution: for example, the NSO is often responsible for agricultural censuses, whereas ongoing agricultural production surveys are carried out by the relevant ministry. In these circumstances, establishing coordination among the various agencies is paramount. This is sometimes difficult because each agency may have different mandates regarding the purpose, scope and timing of their work (FAO, 2005).

Before carrying out the integration, it is necessary to verify the existence of minimum conditions for its feasibility and, especially, its appropriateness.

In this research report, we focus on data integration methodology. However, it should be noted that the success of an integration project depends to a great extent on the relevant *institutional aspects*. Second, the quality of the integrated database or of the integrated processing depends on the *quality of the individual statistical sources used*. In particular, we are fully aware that an assessment of the quality of integrated statistics, and a measurement of the benefits of integration, requires *documentation on the quality of integrated sources* and, therefore, of the related metadata.

3.5 THE DIFFERENT CONTEXTS OF DATA INTEGRATION

Data integration has been used by several NSOs since the 1990s, in a variety of ways, but recently interest in this complex system of data management has revived. The most significant early contributions to record linkage were made in the 1950s, in the field of medical research. Newcombe *et al*. (1959), and Fellegi and Sunter (1963), authored two of the most influential early papers.

To distinguish between different contexts and approaches to integration in statistics, it is first necessary to consider the objects of integration and the source and nature of the integrated information.

Indeed, integration can make reference to one or more information sources, which may correspond to one or more holders, and can concern data of different nature.

The data objects of the integration process may include:

- *statistical data*, collected through total or sample surveys, with the adoption of statistical standards;
- *administrative data*, collected through archives or registries created for administrative purposes, or in compliance with laws or regulations.

We can distinguish between Official Statistics, statistics produced by bodies of the National Statistical Systems (NSS), and information produced by other private or public bodies.

The difference lies in the institutional mandate and data production responsibilities of NSS subjects, and the consequent obligation to produce statistical information according to international and national quality standards.

As for data information from administrative sources, it is necessary to distinguish between the sources deriving from regulations, and those established by the law for statistical purposes.

The dual purpose of "administrative and statistical" archives or registers is important, because it enables administrative information:

- to be used for statistical purposes, and disseminated, together with other statistical information;
- to be planned and organized in a coherent and consistent manner, in accordance with the definitions, classifications and methodologies established for official statistics.

In particular, it is considered *good practice* that when a new archive or register is established by law, the NSS should at least monitor the collection and organization of the administrative information.

As for different contexts for the integration processes, we can consider:

1. integration defined *ex ante* or *ex post* with respect to the data collection;
2. integration of information collected at the same or at a different level of aggregation, i.e. with same type or different types of enumeration units.

Three different cases are presented below. These correspond to three approaches, described in terms of the contexts of application and of the principal effects on the quality of the information produced:

- **Case A**: Integration of surveys through the introduction of a single multi-purpose survey
- **Case B**: *Ex post* integration of data from different surveys or archives
- **Case C**: Planning data integration on the basis of different surveys or archives

Case A: Integration of surveys through the introduction of a single multi-purpose survey

A first type of integration is the introduction of a multipurpose survey having a single subject holder. This survey should detect a plurality of information recorded previously in two or more surveys, that were designed separately to meet different cognitive needs and possibly with different holders.

Therefore, a multipurpose survey is an example of integration only if it is considered as an alternative, or an evolution, with respect to the conduction of further surveys on the same population, through contact with the same types of units (individuals, households, businesses).

Similarly, an agricultural census conducted in conjunction with the population and housing census (at the same time and with the same field operations) can be considered as a case of this type of integration.

Case A integration is therefore the result of a *joint project of collection of several phenomena on the same type of unit* (or on different types, but with known relationships), and is therefore a case of *ex ante integration*, with respect to data collection.

It must be noted that in this case, the integration of surveys designed to capture different phenomena must also address the fact that different respondents were provided in the survey protocol, and therefore the different ways to reflect these changes.

Integration through a multipurpose survey enables us to make with certainty, in relation to a single unit, a joint reading of phenomena that had previously been recognized separately; therefore, it is possible to study interpretative models of the phenomena of interest.

In addition to this enrichment of information, other important advantages of this type of integration are the reduction of the sample size, compared to the total sum of the sample sizes of the individual surveys. This decrease in the number of interviews entails a lower overall cost of detection and statistical burden.

The entity of this reduction depends on the sampling design adopted, and therefore on the information available on the population and the methodology used. The cost reduction may free additional resources, which could be employed to ensure a better quality of the investigation in terms of coverage, accuracy, or timeliness.

Given these benefits, possible disadvantages of integration can arise from a non-optimal choice of timing, which can hinder the detection of various events with different surveys (Keita, 2004). This negative effect is related to cyclical or seasonal factors. If surveys were repeated frequently during the year, or if a continuous plan of investigation were adopted, these effects could be kept under control. However, in many countries, most surveys are carried out only once a year.

Furthermore, to ensure an acceptable amount of time for the interviews, thereby guaranteeing the attention and collaboration of the respondent for the duration of the interview, sections of the questionnaire dedicated to aspects investigated previously through specific surveys could be reduced in terms of level of detail (loss of specificity). In relation to repeated surveys to study the variation in time of given phenomena, effects of sample fatigue may also arise, thus leading to biased estimates.

It should also be noted that the complexity introduced by a survey having a multipurpose interview protocol that provides for more than one respondent, requires systematic monitoring of its actual compliance, to properly deal with and limit the increase in case of non-response or measurement errors.

Positive effects	Negative effects
Ability to study the relations between different phenomena, previously investigated through different surveys on different units (Increased information available for the construction of interpretative models).	Non-optimal timing for the detection of various phenomena may lead to bias in the estimates, especially as regards changes of phenomena over time.
Reduction of the overall sample size and consequent reduction of the overall cost and statistical burden.	Sample fatigue and greater statistical burden for individual respondents. Possible adverse consequences in terms of accuracy (measurement errors, total and partial non-response.
Additional resources freed and can be employed to improve the quality of the survey, in terms of coverage, accuracy, or timeliness.	Simplification of the questionnaire and subsequent loss of information.
	Difficulties with the interview protocol (change of respondents) and consequent possible non-sampling errors (non-response and measurement errors).

FIGURE 2: Potential positive and negative effects on quality in Case A (The transition from multiple surveys on the same population to a single multi-purpose survey)

Case B: *Ex post* integration of data from different surveys or archives

To study phenomena collected through different surveys or archives, it is possible to combine different datasets *ex post*, selecting a specific record linkage procedure, instead of adopting a single multi-purpose survey (see Case A above).

The procedure of record linkage can be considered:

- conceptually simple, if the different datasets present the same type of enumeration units (individuals, households, business, etc);
- technically simple, if the units are identified by the same unique identifier (UID) or by a combination of variables uniquely defined and available in the different datasets (key variables or linking variables);
- operationally feasible, if the linkage is in compliance with the policies governing the dissemination of the results of the various surveys, and the owners have a common goal.

The quality of linking, however, will depend on the quality of information within the individual datasets, and especially on the quality of the UID and of the linking variables.

Depending on the context, this *ex post* integration can be based on the linking or on the aggregation of information of the different datasets. The following cases can be identified.

- ***Case B1****: The different surveys or archives present the same type of enumeration units*. In this case, the integration of information through the record linkage is conceptually possible; therefore, a single data set can be constructed, at the same level of unit. This is the standard example of *micro-level integration* mentioned above. In this way, it is possible to obtain a single dataset at the level of individuals, households or enterprises. This context arises frequently; for this reason, data integration in statistics is often equated with record linkage. Linkage can be achieved by means of various procedures; therefore, in this case, the crucial issue is the choice of the optimal procedure.

  This choice must take into account the specific objectives of integration (in terms of the prospective use of integrated data) and the characteristics of the information to be integrated. The choice of procedure for record linkage must therefore take into account the fact that integration is aimed at the estimation of phenomena or at the estimation of relationships between phenomena, or at the construction or updating of sample frames. Generally, the fundamental issue is the decision on how to combine the statistical units, how to evaluate errors, and how to take them into account to draw inferences concerning the population (Istat, 2008).

- ***Case B2****: Different types of enumeration units used in the different surveys or archives. In this case, two strategies can be adopted for the integration process.*

  - ***Case B2.1****: Aggregation of each survey's units, for the same domains of study*. These domains must have two features: a) they must be uniquely

defined by the different types of enumeration units; and b) they must hold relevant informational content, coinciding with planned estimation domains in the different surveys. A typical example is that of geographical domains (EA, Provinces, Regions, etc.).

This is the classic case of *integration at the macro level*. If this strategy is chosen, it is important to assess whether the benefits from the availability of joint information are significant at an aggregate level, with respect to the case in which individual blocks of separate information are available at the disaggregate level.

Certainly, integration for aggregation can also be conducted in the case of the same enumeration units. This is the case when one of the sources, even if collected at the micro level, is available for integration only at the macro level.

- ***Case B2.2****: Combination of different enumeration units according to rules based on their logical relationship*. This strategy can be seen, for example, in the combination of individuals of the same family, employees of the same company, land parcels of the same area, etc). The complexity of this combination depends on the type of relationship existing between the different units. As mentioned above (Par. 1.3), this relationship may be one-to-one, one-to-many, or many-to-many. The relationship between the units of pair datasets can be formalized by linking matrices. This specific case is studied in this research; its theoretical development is discussed in later parts of this report.

Finally, it should be emphasized that *ex-post* integration may be adopted as part of a trial, not scheduled in advance, seeking to assess the technical feasibility of integration for subsequent use; or to integrate surveys or archives that were difficult to plan jointly, due to the presence, within the relevant institutional framework, of several different ownerships.

In terms of the quality of the information produced with *ex post integration through matching the units* (Case B1), the following effects can be achieved:

- The direct estimation of the characteristics of a variable (average, variance, etc.), on the basis of a greater number of observations, which result from the union of several sources, generally leads to greater consistency in the estimates. The degree of this improvement depends on the quality of the results of the linking procedure, and therefore on both the fraction of matched units, and the entity of the matching errors.

- The estimation of a phenomenon based on its relationship with other phenomena (auxiliary variables) can improve in terms of efficiency, when these are selected from a larger number obtained by combining the units of various databases having different content (variables). Naturally, the advantage of the availability of a greater number of potential auxiliary variables should not be affected by a significant decrease in the number of observations, due to the non-matching of

a relevant part of the units; in this case, mismatch errors influence the quality of the estimates (ref .....).

- The estimation of the parameters of the relationship between phenomena that have not been jointly detected by any of the integrated sources is possible only with the integration; therefore, its contribution on knowledge is purely additional. The accuracy of these estimates is greatly dependent on the quality of the results of the linkage procedure; the literature has studied the strong influence that even a fe pairs of outliers can have on estimating the parameters of the regression line, or on the goodness of fit (Cook, 1977, Freund and Wilson, 1998).

- The construction of a sample frame and its updating or enrichment, through the combination of units of different surveys, is another traditional objective of integration, and enables multiple possible advantages:

    - the possibility to use sampling techniques that allow greater efficiency of the estimates and/or a reduction of the sample size, with a consequent reduction of costs (for example through introduction of *optimal sampling*).

    - the ability to select subpopulations within a frame of interest, on which to conduct a census or a sample survey. This is made possible by integrating the agricultural census with the population and housing census, when the latter is used to identify which farm households are among the households, and the corresponding agricultural holdings. In this case, the advantages consist of a significant reduction of costs, deriving from this additional information.

Positive effects	Negative effects
Greater consistency of the direct estimates of variables present in the various integrated databases.	Effects of mismatches on estimates.
Increased efficiency of the estimates of the variables present in the various integrated databases, thanks to the increased availability of auxiliary variables.	
Ability to estimate parameters of the relationship between phenomena not jointly collected in any of the integrated surveys.	
Availability of new or richer sample frames, from which more efficient sampling designs can be defined.	

FIGURE 3.8. Positive and negative effects on quality in Case B1 (Record linkage of data from different surveys or archives, collected with the same type of enumeration unit)

As mentioned above, in the case of surveys or archives relating to a different type of unit, a solution may be *ex post integration through aggregation of the units* (*Case B2.1*).

Depending on the parameter of interest, the aggregation may be based on different operations. For example, the integration of surveys each focusing on a specific population of individuals, families and businesses, can be achieved by providing estimates, at the territorial domain level (provincial or otherwise), of the number or fraction of individuals, families and businesses presenting certain characteristics of interest.

With reference to the quality of the estimates obtained by aggregation, it can be observed that:

- The gain in information achieved with data integration arises at the study domain level, and not the unit of analysis level; if the domains of study are relevant aggregates, such as those planned for the estimates, the estimates obtained are still relevant.

- The domain of study is a compromise between the different types of enumeration units, and the level of detail that is lost depends on the extent of this compromise; depending on the importance of the phenomena and the specific goals of the integration, the domains of study can be defined in different ways, which are not necessarily common to all data sets.

- The additional information from this type of integration is much greater when the surveys cover different aspects, such as when each variable is collected from only one of the sources to be integrated; a typical case of integration for aggregation is the construction or renovation of a geographical information system, powered by a plurality of surveys and archives that identify different phenomena detected through different types of units.

- A relevant case is the estimation of the relationship between phenomena detected only in different surveys; here, the relationship can be examined only at the level of domains of study, and not at the level of statistical units. However, if the aggregation concerns geographical units, it must be noted that the level of aggregation can affect the estimate of the relationship between the phenomena – this is the so-called *modifiable areal unit problem*, or MAUP (Openshaw, 1983).

- The non-necessity for matching units based on common variables in different databases makes this type of integration better able to deal with differing definitions of these variables.

- The availability of variables collected in most surveys is not necessary, but could enable the provision of more accurate estimates, taking into account any constraints on marginal distributions (for example, calibration).

Positive effects	Negative effects
Potential exploitation of the entire body of information, regardless of the outcome of a linking procedure.	Comparison of observed phenomena in different surveys only at the level of the domains of study permitted by the aggregation.
Studies of relations between phenomena, even if detected in separate investigations.	Dependence, of the correlation between phenomena, on the level of aggregation (MAUP).
Construction of geographical information systems based on surveys collecting information on different types of units.	
No need for common definitions of the variables.	

FIGURE 3.9. Positive and negative effects on quality in Case B2.1 (Aggregation of data from different surveys or archives, collected at different types of enumeration units)

Case C: Planning data integration on the basis of different surveys or archives

In this case, integration is a process designed ex ante, and each survey or archive is designed bearing in mind this common goal of integration. For example, if it is sought to link the information of a database, established for administrative purposes, with the data of one or more statistical surveys, these surveys will be designed with proper consideration of the database.

Compared to the design of a unique multipurpose survey, in this case each survey maintains its own ownership and autonomy in response to specific needs. However, each survey is designed on the basis of a common project, from both a methodological and a political point of view.

This common project allows each individual holder:

- to achieve his specific goals at a lower cost, thanks to the joint exploitation of the data collected in the different surveys;
- to take advantage of an overall survey design based on a wider set of skills and experience. Integration is generally a factor supporting an increase in the quality of outputs in terms of consistency and accuracy, and promotes common standards, common definitions and sample frames.

The benefit for individual parties must be sufficient to overcome the disadvantage of losing autonomy in designing its survey. The agreement between the holders of the investigation will naturally be integrated with the rules established to meet any new or different informational needs of individual parties.

From a collective point of view, these forms of *ex ante* integration constitute good practices, especially when different institutional stakeholders are involved.

From a methodological point of view, the advantages of *ex ante* integration over *ex post* integration are significant. First, the integration of different sources does not occur episodically, but in a structured and programmed manner. In addition:

- A single unified survey design enables an optimization of the use of the overall technical, organizational and financial resources available to carry out various surveys, especially in terms of costs; in particular, an overall sample design will enable the planned level of the quality of estimates to be achieved, establishing a coherent sample size, a sample selection technique and an estimation method of the target parameters;
- An integrated *ex ante* design of the surveys maximizes the quality of matching units sampled in the different surveys, due to a clear definition of the UIDs and the key variables for record linkage, and to greater accuracy in data collection for these variables. This greater detail is the result of specific actions aimed at reducing non-sampling errors such as non-response and measurement errors.

Problems can arise from the need for a minimum time gap between the various data collections, and a possible loss of specificity and flexibility can ensue from the need to link the units of the different surveys.

Positive effects	Negative effects
Compared to the design of a unique multipurpose survey, each survey maintains its autonomy in response to specific needs.	Need for a minimum time gap between the various data collection exercises.
Cost reduction, thanks to the joint exploitation of the data collected in the different surveys.	Loss of specificity and flexibility due to the need to link the units.
Optimization of the use of the overall technical, organizational and financial resources available (overall design based on a wider set of skills and experience).	
Good quality of record linkage, thanks to the special attention given to the adoption of common definitions, UIDs and to the data collection for the linking variables planned.	
An overall sample design will enable achievement of the planned level of quality of the estimates, establishing a coherent sample size, a sample selection technique and an estimation method of the target parameters.	

FIGURE 3.10: Positive and negative effects on quality in Case C (Planning data integration from different surveys or archives)

In conclusion, this type of integration maximizes the traditional benefits of integration, reducing the costs of data collection and the statistical burden, and also helps to ensure attainment of a planned level of quality of the estimates, in relation to both the single sample design and to the linking procedure.

Such *ex ante* integration, resulting from the joint efforts of the holders of different surveys or archives, can be considered the evolution of the experimentation with *ex-post* integration. It is the duty of National Statistical Systems and international organizations to promote and support these projects.

The Global Strategy to Improve Agricultural and Rural Statistics (World Bank, 2010) is an example of a program having this aim, as are the recommendations of the UN, UNECE and UNECA for coordination between agricultural censuses and the population and housing censuses. The FAO Guidelines for linking population and housing censuses with agricultural censuses are a good example of how to best leverage the experiences and practices of different countries, circulating and interpreting them to improve the global system of agricultural statistics.

An example of *ex ante* planned integration of different sources is the integrated data system for agricultural statistics considered in the Global Strategy.

The integrated survey framework, shown in Figure 4 below, provides an overview of how annual and periodic surveys are connected in the data system. The survey framework also takes other data sources into account:

- Administrative data (e.g. registers of agricultural holders, land ownership, inspections, trade data, etc.,)
- Remotely-sensed data (land cover, land use, vegetative indices)
- Agri-business (consumptions and prices)
- Expert evaluations
- Community surveys (village-level data on infrastructures, the environment, services to households and agricultural holdings).

The framework provides the capacity for a longitudinal analysis of the core data, as well as linkages to data collections for economic, environmental, and social issues. The use of the master sample frame ensures that the data collections are also connected to land use.

The remaining pillar of integration is the management of data capable of maximizing its use for data analysis.

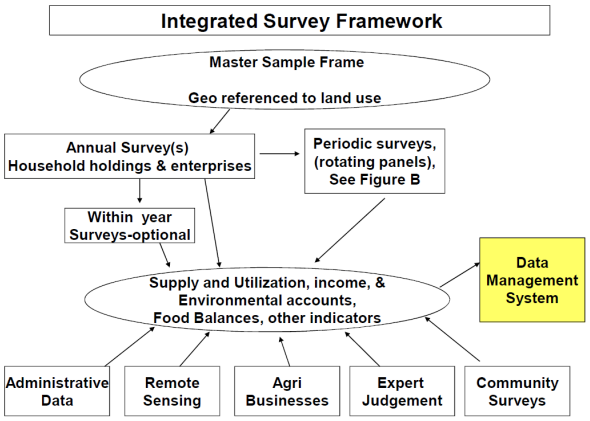

FIGURE 4. The overall integrated data system (World Bank, 2010)

4

A Common Notation for Integration

The last chapter of this Part introduces a basic notation, a common toolbox of symbols and concepts used in the different parts of this Technical Report and that simplifies the combined use of the various methods that can be involved in a process of statistical integration.

This symbology overcomes the problem posed by the fact that the notation for describing a standard statistical problem refers to a single population, whereas an integrated context takes into account several different populations with several statistical units, with complex linking structures among them.

4.1 POPULATIONS

Integrated agricultural statistics examine phenomena concerning several statistical populations. Each population is a collection of specific statistical units, capable of providing a given subset of the information of interest for the integrated statistics.

Four different populations must be considered: *individuals, land-parcels, households and farms*.

The first two populations comprise elemental statistical units, and the latter two include clusters of elemental units. Households are clusters of individuals, while farms may be viewed either as clusters of individuals or as clusters of land-parcels.

Populations of elemental units

NOTE

Unless expressly stated otherwise, here and in the following Chapters, we adopt the convention that a symbol used for denoting a *set* also indicates its *cardinality*[1]. This avoids the introduction of an excessive number of symbols. Generally, the overall context (of a sentence or of an expression) clarifies the use of a given symbol.

[1] The *cardinality* or *size* of a set indicates the number of its units.

Individuals. Let I denote the population of individuals. The symbol I represents both the population and the number of individuals within it.

The subscript $i \in I$ designates a generic person.

Common individual enumeration units in agriculture are:

- households members
- agricultural holdings
- workers of rural or economic farms or agricultural companies.

Land parcels. The land of a given country can be considered as a set L of land parcels.

The specific area of a land parcel l is denoted with $a_l, l \in L$.

In this Technical Report, the term *land parcel* generically indicates the elemental unit of land. Depending upon how geographical information is prepared in a given country, this term may denote either:

- A field that is a given portion of land operated in a uniform manner (e.g. the seeding of a certain type of plant, or that has undergone a specific phyto-pharmacological treatment) by one farm. In this case, a_l denotes the area of the field.
- A definite point in a geographic grid with given latitudinal and longitudinal coordinates. In this case, area a_l is the area around the point (for example, a circle having a radius of two meters) on which the survey enumerator must collect the target information pertaining to the land (e.g. type of cultivar, number of plants, etc.).

It is important to recall that land parcels may be aggregated at different geographical levels: Census Enumeration Areas, provinces, etc.

Population of clusters

Households. Let H denote the population of households.

The symbol $u_h, h \in H$ is used to represent a generic household that can also be considered as a set of individuals. Therefore, households are also clusters within the population I.

The number of individuals of a generic household is $n_h, h \in H$.

Farms. Let F denote the population of farms.

The symbol $u_f, f \in F$ is used to indicate a generic farm.

In this manual, the term “farm” is used to denote objects that can be different in certain respects, such as:

- rural farms
- cooperatives
- economic farms or companies/farming businesses

NOTE

We assumed that one individual belongs to only one farm. If an individual is related to more than one farm, s/he is assigned to only one farm, according to a criterion of prevalence.

Similarly, a land parcel belonging to more than one farm is assigned to only one farm, according to the criterion of prevalence.

Each farm $u_f, f \in F$ can be viewed as a set of individuals (workers or farm holders). Taking into account the note above (on prevalence), farms are also clusters within the population *I*.

The number of individuals of the generic household is $n_f, f \in F$.

In addition, each farm can be viewed as a set of land parcels. Thus, farms are also clusters within the population *L*.

The overall area of a farm is given by:

$$\tilde{a}_f = \sum_{a_l \in u_f} a_l.$$

In the following pages, we will generically denote a population or its statistical unit, without specifying the type of population or unit under consideration. In these cases, the symbol *J* (*J*=*I* or *L* or *H* or *F*) is used to indicate the generic population (and its size), and the symbol *j* (*j* =*i* or *l* or *h* or *f*) is used to denote its (generic) statistical unit.

4.2 LINKING MATRICES

The linking matrices enable formal definition of the relationships between the units of the different populations.

Since each individual belongs to a household or can work in a farm, it is necessary to define correspondence matrices that enable identification of the clusters to which the individual belongs. Likewise, a specific correspondence matrix allows definition of the link between the land parcel and the farm to which it belongs.

Other relationships can also be described, considering that an individual (in a given household) works on a farm with a specific land parcel. Thus, it is possible to establish the links between households and farms, and land parcels and households.

The relationships between the population J (J=I or L or H or F) and the population J'($J' \neq J$ and J'=I or L or H or F) are represented by linking matrices $\Lambda^{JJ'}$, which are $J \times J'$ matrices, the generic element of which is given by:

$$\lambda_{jj'} = \begin{cases} \alpha, \text{if the unit } j \in J \text{ has a relationship with the unit } j' \in J' \\ 0, \qquad\qquad\qquad\qquad \text{otherwise} \end{cases},$$

with $\alpha > 0$. Generally, we set $\alpha = 1$.

Thus, let Λ^{IH} an $I\times H$ matrix, with a generic element given by

$$\lambda_{ih} = \begin{cases} 1, \text{if} \quad i \in u_h \\ 0, \text{otherwise} \end{cases}.$$

Analogously, let Λ^{IF} an $I\times F$ matrix, with a generic element given by

$$\lambda_{if} = \begin{cases} 1, \text{if} \quad i \in u_f \\ 0, \text{otherwise} \end{cases}.$$

The ***standardized linking matrix*** $\tilde{\Lambda}^{JJ'}$ ensures that the sum of the column of its generic elements, $\tilde{\lambda}_{jj'}$ equals 1. A standardized linking matrix can be derived from the original matrix $\Lambda^{JJ'}$ by setting

$$\tilde{\lambda}_{jj'} = \lambda_{jj'} / \sum\nolimits_{j\in J} \lambda_{jj'} .$$

Thus, for example, let $\tilde{\Lambda}^{HF}$ be an $H\times F$ standardized matrix, with a generic element given by

$$\tilde{\lambda}_{hf} = \begin{cases} n_{hf}/n_f \text{ if } u_f \cap u_h \neq 0 \\ 0 \qquad\quad \text{otherwise} \end{cases},$$

n_{hf} being the number of individuals of the household u_h who work in the farm u_f .

The standardized matrix is a natural way to express the relationships between land parcels and farms. Let $\tilde{\Lambda}^{LF}$ be an $L\times F$ matrix, with a generic element given by

$$\tilde{\lambda}_{lf} = \begin{cases} s^{lf} = a_l/\tilde{a}_f \text{ if } l \in u_f \\ 0, \qquad\quad \text{otherwise} \end{cases}.$$

4.3 VARIABLES OF INTEREST

Let Y denote the generic variable of interest.

In the **standard approaches** to statistical inference (design- or model-based), adopted in Parts 4 and 5 of this Report, the symbols

$$Y_i, i \in I;\ Y_h, h \in H;\ Y_f, f \in F\ ; \text{and } Y_l, l \in L$$

denote the value of the variable *Y* observed on a generic individual, household, farm or land parcel.

In the **Bayesian approach to inference** (adopted in Part 3), the symbols

$$Y_i, i \in I;\ Y_h, h \in H;\ Y_f, f \in F\ ; \text{and } Y_l, l \in L$$

indicate random variables. The values measured on specific units are denoted by lower case letters, as follows:

$$y_i, i \in I;\ y_h, h \in H;\ y_f, f \in F\ ; y_l, l \in L.$$

To simplify the description, here and in the rest of this Chapter, we will adopt the notation of the standard approach.

In certain situations, it will be necessary to analyze several variables concerning the same population. Suppose that v different variables are observed for the same set of individuals. To distinguish between them, we denote with $Y_{i,v}$ the value of the v-th variable of interest (with $v=1,...,V$), of the i-th unit of the population I. Analogously, $Y_{l,v}$, $Y_{h,v}$ and $Y_{f,v}$ represent, respectively, the v-th variable evaluated on a land parcel, a household or a farm.

In the following pages, without loss of generality, an original categorical variable with M different modalities Y_j referred to the j-th unit (with $j=i$ or $j=l$ or $j=h$ or $j=f$), is expressed in terms of M dummy variables Y_{jm}, (m=1, ..., M), where Y_{jm} = 1 if the unit j is characterized by the m-th modality of the original variable Y_j. Thus, in the manual, only quantitative and dichotomous variables will be considered.

4.4 PARAMETERS OF INTEREST

A parameter of interest θ_J of the population J can be defined as the value of a function f , defined on all the J values Y_j of the population

$$\theta_J = f(Y_j; j = 1, \cdots, J). \qquad (4.1)$$

Thus, the symbols θ_I θ_L, θ_H and θ_F denote the parameters of interest referring respectively to populations I, L, H and F. These parameters are functions of the values $Y_i, i \in I, Y_l, l \in L, Y_h, h \in H$ and $Y_f, f \in F$.

NOTE

The variables referring to elemental units ($Y_i, i \in I$ and $Y_l, l \in L$) may originate new variables at cluster level ($Y_h, h \in H$ and $Y_f, f \in F$), by simple algebraic operations (generally, aggregation) upon the cluster.

Example 1. If Y_i is a dummy variable that equals 1 if the individual *i* is a female, then $Y_h = \sum_{i \in u_h} Y_i$ and $Y_f = \sum_{i \in u_f} Y_i$

denote, respectively, the number of females in the household u_h and in the farm u_f.

Example 2. If Y_l denotes the total production of crops of the land-parcel *l*, then

$$Y_f = \sum_{l \in u_f} Y_l.$$

denotes the total production of crops of the farm u_f.

The inverse operation (from the cluster to elemental units) cannot be performed easily. Indeed, the variables $Y_h, h \in H$ and $Y_f, f \in F$ are related to characteristics that refer to the entire cluster; thus, some assumptions are necessary if a specific value is to be assigned to the individual elemental units of the cluster itself.

It follows that the main quantities of interest for this Report will be the parameters θ_H and θ_F. Indeed, the set of parameters θ_H includes the parameters θ_I of the population *I*. Likewise, the set of parameters θ_F includes the parameters θ_L of the population *L*.

Unless expressly stated otherwise, the specific parameter θ_J will refer to a total, e.g.:

$$\theta_H = \sum_{h=1}^{H} Y_h, \quad \theta_F = \sum_{f=1}^{F} Y_f.$$

In certain situations, where several variables concerning the same population must be analyzed, we denote with $\theta_{H,v}$ the value of the parameter related to the *v*-th variable of interest (with *v=1,...,V*) for the population *H*. Analogously, $\theta_{F,v}$ will represent the parameter related to the *v*-th variable of interest for the population *F*.

The definition (see Para. 4.1) can be extended to the case of an *integrated parameter,* $\theta_{JJ'}$, which refers to the two different populations *J* and *J'*. A general definition is given below.

An *integrated parameter,* $\theta_{JJ'}$ is the value of a function *f,* defined on all the values of the two different populations *J* and *J'*, and may take into account the links between the two populations. In the box below, we develop some examples.

EXAMPLES OF INTEGRATED PARAMETER

Example 1: Ratio of totals. An example is the ratio between the income of the households (total θ_H) and the total revenue of the farms (total θ_F):

$$\theta_{HF} = \frac{\theta_H}{\theta_F}.$$

In this case, the integrated parameter is a function of the values of the variables of interest on the units of the two populations, and does not take into account the links between the units of the populations.

Example 2: Total of a subclass. θ_{HF} may correspond to the total of a subclass, where the quantitative variable is related to a population (e.g. *H*) and the subclass is defined in relation to the other population (e.g. *F*). An example is the total income of households of people working in farms specializing in animal breeding:

$$\theta_{HF} = \sum_{h=1}^{H} \sum_{f=1}^{F} \sum_{i=1}^{I} \lambda_{ih} \lambda_{if} Y_{(hf)} Y_f,$$

being $Y_{(hf)} = Y_h / n_{hf}$ in which Y_h is the income of household u_h and Y_f is a dummy variable equal to 1, if the farm u_f is specialized in animal breeding.
In this case, the integrated parameter is a function of the values of the variables of interest on the units of the two populations, and of the links existing among the units of the populations.

Note. In Example 1, the integrated parameter can be measured with two separate and non-integrated surveys, the first for θ_H and the second for θ_F. However, in Example 2, it is necessary to use an integrated observation, such as that described in Part 4 of this manual. In an integrated observation, once a given unit of interest (e.g. a farm) is observed, the linked units of the other population (e.g. the households) are also contextually observed. Otherwise, it is possible to use record linkage techniques (such as those described in Part 3) or micro-level estimation (as described in Part 5 of this Manual).

PART 2

A Quality Framework for Integration

This Part presents a new set of quality standards and guidelines for data integration, based on revisited best proposals from NSIs and international organizations. It identifies general aspects of quality assessment of data integration and operational factors.

The main issues to be addressed are:

1. To what extent should the quality reporting of integrated data and the process of integration differ?
2. Does the peculiarity of the data integration process justify the adoption of a specific set of requirements, quality indicators and guidelines?

The research highlights that the answer depends on the specific context of the data integration. Three such scenarios are identified in the present research.

- In the case of multipurpose surveys (Case A), the issues are those typically arising in a statistical survey. The integration does not require a specific quality assessment framework.
- In the case of *ex post* integration (Case B), it is necessary to use record linkage to establish specific requirements, indicators and guidelines for the quality of the integration process. Therefore, a framework of rules is required, to guide the decisions on (*i*) how to conduct the match, (*ii*) how to measure errors, and (*iii*) how to take these into account in making inferences with regard to the population.
- In the case of survey integration through the joint planning and coordination of the activities (Case C), all surveys are to be conducted in accordance with guidelines for statistical surveys. Their added value consists in the record linkage of the data, which requires the establishment of a specific quality assessment framework.

In conclusion, when record linkage is the central process of data integration, its quality should be assessed according to a specific quality framework. Currently, appropriate conditions exist to promote shared guidelines and best practices for data integration in agricultural statistics. We also note that the Annex to this Part contains a brief glossary on data integration.

1

Quality Standards and Guidelines for Data Integration

1.1 GENERALITY

Quality and "reliable statistics are vital for describing the reality of people's lives and providing the evidence required to develop and monitor effective development policies (Klosterman, 1995)", especially for developing countries.

The design and implementation of integrated statistical surveys is an important strategy in improving the quality of information produced by these countries, and may lead to a more sustainable collection and organization of statistical information.

From the literature on the quality of official statistics (statistical information produced by national institutions responsible for compiling statistical data), it is clear that a great part of the work carried out in the last two decades by NSIs and international organizations has enabled accessing a modern theoretical framework for the assessment of statistical quality, one that is now widely accepted at the international level.

The following section examines the current theoretical quality framework, and identifies some critical issues and shared features.

1.2 INTERNATIONAL REQUIREMENTS ON DATA QUALITY

The International Organization for Standardization (ISO) defines quality as: "the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs" (ISO 8402 – 1986, 3.1). Therefore, the most important characteristics of quality depend on users' perspectives, needs and priorities; naturally, these vary among different groups of users.

The *FAO Statistics Quality Assurance Framework* (FAO SQAF) defines quality in statistics as the degree to which statistical outputs fulfil requirements (ISO standard 9000, 2005) and respond to international standard principles such as *relevance, accuracy and reliability, timeliness and punctuality, accessibility, comparability and coherence*. In the SQAF, each principle features corresponding good practices. The principles are defined as follows:

- *Relevance* is the degree to which statistics meet current and potential user needs.
- *Accuracy and reliability*
  - *Accuracy* refers to the estimates' closeness to the true values that the statistics sought to measure;
  - *Reliability* refers to the closeness of the initial estimates to the subsequent or final estimates.
- *Timeliness and Punctuality*
  - *Timeliness* is the speed of dissemination of statistical outputs, the period of time between the end of a reference period (or a reference date) and the dissemination of the statistical outputs.
  - *Punctuality* refers to the possible time lag between the actual delivery date of statistical outputs and the target date for their delivery – for example, dates announced in an official release calendar or previously agreed between partners.
- *Coherence and Comparability*
  - *Coherence* is the suitability of statistical outputs to be combined meaningfully, in different ways and for various uses.
  - *Comparability* refers to the extent to which differences between different geographical areas, non-geographical domains, or over time, can be attributed to differences between the true values of the statistical characteristics.
- *Accessibility and Clarity*
  - *Accessibility* is defined as the ease, the set of conditions and the modalities with which users can obtain data.
  - *Clarity* refers to the availability of adequate documentation: whether data are accompanied with appropriate metadata or illustrations such as graphs and maps, whether information on their quality is also available (including limitation in use), and the extent of additional assistance.

As evidence of the various dimensions of the quality of statistical products, and of the individual principles on which they are based, Figure 1.1 below compares the components of quality determined by select international organizations.

FAO	UNECE	OECD	EUROSTAT	IMF
Relevance	Relevance	Relevance	Relevance	Pre-requisites of quality
				Methodological soundness
Accuracy and Reliability	Accuracy	Accuracy	Accuracy	Accuracy and Reliability
Timeliness and Punctuality	Timeliness	Timeliness	Timeliness and Punctuality	Serviceability (part)
	Punctuality			
Accessibility and Clarity	Accessibility	Accessibility	Accessibility and Clarity	Accessibility
	Clarity	Interpretability		Assurance of integrity
Comparability and Coherence	Comparability	Coherence	Comparability	Serviceability (part)
			Coherence	
	Considered more relevant at the level of the organization	Credibility		Pre-requisites of quality (part)
				Pre-requisites of integrity (part)

FIGURE 1.1. ***Mapping of quality components used by international statistical organizations***

In addition, quality management system play a fundamental part in producing statistics. EUROSTAT provides an interesting analysis of data quality assessment, reporting the existence of a relationship between producers/users and the corresponding standard principles. However, the scope of this Report is limited to statistical products; it will not cover areas such as support processes, management systems or leadership, or the institutional environment of statistical production.

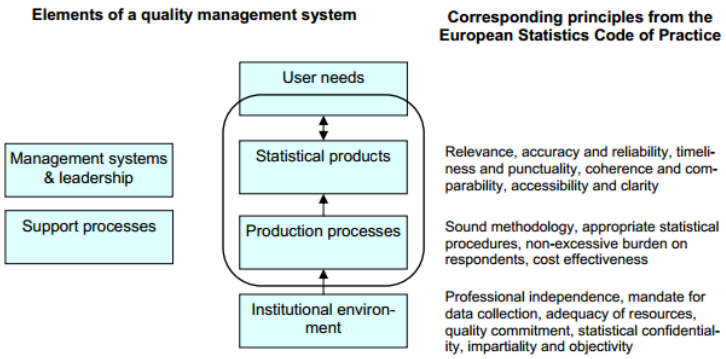

FIGURE 1.2: ***Elements of a Quality Management System.*** **Source: EUROSTAT**

1.3 QUALITY AND DIFFERENT CASES OF DATA INTEGRATION

An important outcome of NSIs' and international organizations' experiences with data integration is the awareness of a need for specific guidelines (manuals) to manage

processes. These guidelines should be consistent with the principles and requirements relating to the quality of data integration.

Although integration can be performed in various contexts, it is necessary to distinguish three cases especially:

- **Case A:** integration is performed by replacing several surveys with a single multipurpose survey. In this case, the issues that may arise are those typically relating to the design, implementation and evaluation of a similar statistical survey. It can be assumed that integration does not require specific principles, requirements or guidelines. As mentioned above, this substitution of several surveys with a single multipurpose survey has several critical aspects, which should be taken into account as appropriate.

- **Case B:** *ex post* integration, through record linkage between data collected from different surveys or archives. This should be governed by a specific quality assessment frame. Indeed, if the integration is based on record linkage, regardless of whether the sources are surveys or administrative data, there should be a framework of rules guiding decisions on (1) how to conduct the match (2) how to measure errors and (3) how to take these into account when making inferences concerning the population. These decisions should cover both record linkage of the same type of enumeration units, as well as record linkage of different enumeration units.

- **Case C:** the integration of various surveys through the joint planning and coordination of the activities of each survey. This is a complex process, designed as a statistical project. All surveys that contribute to the outcome of the project must be conducted in accordance with relevant guidelines for statistical surveys. However, in this case, the added value lies in the record linkage of the data. Thus, the use of record linkage makes it necessary to establish specific principles, requirements and indicators of the quality of this integration process. Record linkage is the central and peculiar process of integration of data from different sources, and its quality should be assessed in a specific manner.

In conclusion, when record linkage is the central process of data integration (as occurs in Cases B and C) its quality should be assessed against a specific quality framework.

1.4 GENERAL ASPECTS FOR STRENGTHENING THE QUALITY OF DATA INTEGRATION

The quality of data integration should respond to international requirements; the main principles from the "Guidelines on Integrated Economic Statistics" are outlined below. The Guidelines provide practical guidance on enhancing the consistency, coherence and reconciliation of statistical information by applying the methodology of integrated economic statistics, with the System of National Accounts 2008 as the overarching conceptual framework.

The Guidelines also feature case studies and other practical material, to share experiences on the implementation of integrated statistical production within national statistical systems **(Guidelines on Integrated Economic Statistics, 2013, UNSD).**

From an examination of the good practices described in these Guidelines, it emerges that the main topics for consideration are:

- Principles of statistical integration
- Guidelines for maximizing added value through data integration
- Guidelines for promoting common statistical frames, definitions and classifications
- Requirements for Record Linkage
- Guidelines for the Record Linkage of administrative data

Some examples of good practices, from Statistics Canada, the U.S. Census Bureau, Statistics New Zealand and National Statistics Great Britain, are reported below.

1.4.1 PRINCIPLES OF STATISTICAL INTEGRATION

In a data integration project, it is crucial to ensure that the principles and norms governing personal data integration for statistical and related research purposes are observed. In particular, it is necessary to verify whether the principles *allowing* integration have been observed, and, if so, *how* integration should be performed.

Some examples of principles on statistical integration affirmed by National Statistics are illustrated below. These principles apply to both data integration and data classification. It is interesting to note that *common standards* and *administrative data* assume a pivotal role.

a) **United Kingdom: Principles of statistical integration and the classification of National Statistics as stated in the 2002 Code of Practice and embodied in the 2008 Protocol:**

1. Statistical systems will be designed in ways that maximize the potential to add value through data integration.
2. Common statistical frames, definitions and classifications will be promoted and used in all statistical surveys and sources.
3. The value of administrative data in producing National Statistics will be recognized, and statistical purposes should be promoted in the design of administrative systems.
4. All producers of National Statistics will, wherever practicable, adopt common geographic referencing and coding standards.
5. All those involved in the production of National Statistics will promote the adoption of international standards and guidance.

6. Responsibility for the adoption, development, management and application of corporate statistical frames, definitions and classifications will be clearly defined, and the details made widely accessible.
7. Information about the application of statistical classifications will be widely disseminated.
8. Classification decisions will be based on professional considerations; comply with standard guidance on use and interpretation; and be reached and disseminated using established and transparent procedures.

b) Data integration approach as stated by Statistics New Zealand, 2005

Each project of data integration should be articulated into two main sub-phases: (*i*) Evaluation and Approval and (*ii*) Operational Phase.

In evaluating the need for integrating personal data for statistical or research purposes, the following recommendations should be considered.

1. NSIs must only undertake data integration if integration can produce or improve official statistics.
2. Data integration should be considered when it can reduce costs, increase quality or minimize compliance load.
3. Data integration benefits must clearly outweigh any privacy concerns about the use of data and risks to the integrity of the official statistics system.
4. Data integration must not occur when it materially threatens the integrity of the source data collections.
5. Data must not be integrated where any undertaking has been given to respondents that would preclude this.
6. Data integration must be approved at an appropriate level by all the agencies involved.

The following suggestions should guide the operative implementation.

1. Integrated data must only be used for approved statistical or related research purposes.
2. The size and data variables of the linked dataset must be no larger than necessary to support the approved purposes.
3. Integrated data will be stored apart from other data.
4. Names and addresses can only be kept in an integrated dataset while necessary for linking.
5. Unique identifiers assigned by an external agency must not be retained in an integrated dataset.
6. Data integration must be conducted openly.

c) Policy Statement on Record Linkage adopted by the U.S. Census Bureau (2009)

This document contains further examples of principles governing statistical data integration, listed in the table below. It is noteworthy that the first two principles are "Mission necessity" and "Best Alternative", relating to the expected quality of the data.

N.	Principle	Description
1	Mission Necessity	The linkage must be necessary and consistent with the Census Bureau's legal authority and mission.
2	Best Alternative	The Census Bureau will examine alternatives for meeting the project objectives and determine that record linkage is the best alternative given considerations of cost, respondent burden, timeliness, and data quality.
3	Public Good Determination	The Census Bureau will weigh the public benefits to be gained by the information resulting from the record linkage against any risks to individual privacy that may be created by the linkage and determine that the benefits clearly outweigh any risks. In addition, the Census Bureau will proactively implement procedures to mitigate any risks. The confidentiality of the resulting information is protected […].
4	Sensitivity	The Census Bureau will assess the public perception of the level of risk to individual privacy of a particular linkage and create an appropriate level of review and tracking.
5	Openness	The Census Bureau will communicate with the public about its record linkage activities, how they are conducted, and the purpose and benefits derived from them.
6	Consistent Review and Tracking	Record linkage activities will undergo a consistent review process using the criteria set forth in this policy and be centrally tracked by the Census Bureau.

1.4.2 GUIDELINES FOR MAXIMIZING THE POTENTIAL TO ADD VALUE THROUGH DATA INTEGRATION

As mentioned above, according to the **Guidelines on Integrated Economic Statistics**, the aim and benefits of integration consist in the enrichment of available information, and the reduction of costs and of the statistical burden. National Statistics Institutes should set and promote a statistical system that is designed to maximize the potential to add value through data integration, in accordance with the following recommendations:

- When a new statistical collection is requested, existing data sources should be analyzed to evaluate whether they can provide the required data, individually or through integration.
- Data collection should be coordinated with a view to maximize the potential use of each dataset and minimize the potential respondent burden.
- Data processing should be integrated as much as possible, to exploit the benefits of common technology, methods, tools and processes.
- Statistical analysis should examine all relevant data sources, for a coherent comprehension of the subject matter.
- Data and metadata should be made available on the website on a timely basis, to facilitate users' access and make a more effective use of the material available.

1.4.3 GUIDELINES FOR THE PROMOTION OF COMMON STATISTICAL FRAMES, DEFINITIONS AND CLASSIFICATIONS

Statistical agencies seeking to perform data integration should first adopt international concepts, classifications, and methods. The current general framework for statistical quality was established pursuant to the need to adopt international standards. These standards were then defined and updated, by NSI representatives and experts from several different countries.

National Statistics (NS) have produced guidelines for the implementation of the principles of statistical integration and classification. The guidelines reported below promote and use common statistical frames, definitions and classifications in all statistical surveys and sources.

- NS and other subjects of the National Statistical System should develop and promote the *use* of common statistical frameworks, questions, definitions and classifications, to encourage harmonization across official statistics.
- Standards for the harmonization of NS should aim to *cover*:
  - classifications (geographical, social, and economic)
  - statistical units: e.g. business units (enterprise, company, establishment), and social units (family, census family, economic family, household, dwelling)
  - definitions, standard concepts and variables
  - harmonized questions and question modules
  - common frameworks, such as National Accounts, and others that constitute a basis for consolidating statistical information
  - variable names
- The harmonization of data *over time*, and across different sources, will aim to:
  - minimize the time invested in, and the costs of, developing data collections and the respondent load, by avoiding unnecessary duplication across sources;
  - maximize the quality and value of information gained from any single source, taking full advantage of opportunities for cross-analysis, exchange and re-use of data users' understanding of statistical information.
- The harmonization of geographic referencing and coding standards should aim to:
  - ensure that all producers of official statistics apply the geographic standards, unless there are strong operational reasons for doing otherwise;

- encourage the development and promotion of further geographic standards, a harmonized approach to geographic referencing and data visualization, and the use and sharing of standard digital boundary data sets, referencing data and gazetteers.

All standards and classifications adopted for the NS should be made available on the relevant *website*. Guidelines on the adoption and use of standards and classifications, and a series of practical guides on the application of standards should be developed, and should also be available on the website.

Information on the use of statistical classifications within each statistical resource should be incorporated in the *metadata* accompanying that resource.

NS quality assurance procedures will include an assessment of the extent to which each statistical product and its associated data source(s) comply with the standards outlined in this protocol, and harmonizes with data obtained from other sources. This assessment will include an evaluation of the degree to which this compliance addresses user needs.

To promote the requirement of conformity to common frameworks or classifications, and to ensure the consistency and comparability of statistical series, it also necessary to identify the cases of non-compliance with agreed standards.

1.4.4 REQUIREMENTS FOR RECORD LINKAGE

In the majority of cases, data integration is achieved through the procedure of record linkage. This procedure is based on a common unique identifier of the units (UID), or on a combination of variables that identify the units in a unique way (linking variables). The U.S. Census Bureau gives a good example of the detailed requirements of record linkage.

Three of the four requirements are general and typical of other statistical processes:

- Requirement 1, *confidentiality*, is also typical of the use of administrative data;
- Requirements 2 and 4 – respectively, the need to *plan* for the process of record linkage, and to provide the *documentation* needed to evaluate this process –are standard requirements for any statistical project of data collection or processing.
- Requirement 3 concerns the accuracy of record linkage and represents the specific value added by statistical data integration to the quality framework of the statistical process. In turn, the accuracy of this process is detailed by 4 sub-requirements.

The descriptions and examples of each requirement are the following.

Requirements	Description
Req. 1	Throughout all processes associated with linking, unauthorized release of protected information or administratively restricted information must be prevented by the following laws, NSI policies and additional provisions governing the use of the data (*Protecting Confidentiality*) […].

Req. 2	A *plan* must be developed that addresses: • Objectives for linking the files • Data sets and files to be linked • Verification and testing of the linking systems and processes • Training for staff involved in the clerical record linkage operations • Evaluation of the results of the linkage (e.g. link rates and clerical error rates)
Req. 3	Record linkage processes must be developed and implemented to link data records *accurately.*
Sub-Req. 3.1	Specifications and *procedures* for the record linkage systems must be developed and implemented.
	Examples of issues which automated record linkage systems might address: • Criteria for determining a valid link. • Linking parameters (e.g. scoring weights and the associated cut-offs). • Blocking and linking variables. • Standardization of the variables used in linking (e.g. state codes and geographic entity names are in the same format on the files being linked). Examples of clerical record linkage systems include: • Criteria for determining that two records represent the same entity. • Criteria for assigning records to a specific geographic entity or entities (i.e. geo-coding). • Linking variables. • Guidelines for situations requiring referrals. • Criteria for sending cases to field follow-ups.
Sub-Req. 3.2	Record linkage systems must be *verified and tested* to ensure that all components function as intended.
	Examples for automated record linkage systems include: • Verifying that the specifications reflect system requirements. • Verifying that the systems and software implement the specifications accurately. • Performing a test linkage to ensure systems work as specified. Clerical record linkage systems include: • Verifying that the specifications reflect system requirements. • Verifying that the instructions will accomplish what is expected. • Testing computer systems that support clerical linking operations.
Sub-Req. 3.3	*Training* for the staff involved in clerical record linkage (as identified during planning) must be developed and provided.
	Examples of training activities include: • Instructing clerks on how to implement the specifications. • Providing a training database to give clerks a chance to practice their skills. Assessing error rates of clerks and providing feedback.
Sub-Req. 3.4	Systems and procedures must be developed and implemented to *monitor and evaluate* the accuracy of the record linkage operations and to take corrective actions if problems are identified.
	Examples of monitoring and evaluation activities for automated record linkage operations include: • Evaluating the accuracy of automated linkages by a manual review.

	• Monitoring link rates and investigating deviations from historical results, and taking corrective action if necessary. Examples of monitoring and evaluation activities for clerical record linkage operations include: • Establishing an acceptable error rate. • Establishing quality control sampling rates. • Monitoring clerks' error rates and referrals, and taking corrective action if necessary (e.g., feedback or retraining).
Req. 4	*Documentation* needed to replicate and evaluate the linking operations must be provided. The documentation must be retained, consistent with applicable policies and data-use agreements, and must be made available to Census Bureau employees who need it to carry out their work.
	Examples of documentation include: • Plans, requirements, specifications, and procedures for the record linkage systems. • Programs and parameters used for linking. • Problems encountered and solutions implemented during the linking operations. Evaluation results (e.g. link rates and clerical error rates)

1.4.5 GUIDELINES FOR RECORD LINKAGE OF ADMINISTRATIVE DATA

As mentioned above, data integration is often based on sources that include administrative archives. For this reason, guidelines on the treatment of administrative data for statistical purposes are also useful in setting requirements and guidelines on data integration. Statistics Canada (**Statistics Canada, 2009**) has developed specific guidelines for the record linkage of administrative data:

Guidelines	Description
Conform to the Agency's Policy	When Record Linkage of administrative records is necessary (e.g. for tracing respondents, for supplementing survey data, or for data analysis), conform to the Agency's Policy on Record Linkage.
Privacy concerns	Privacy concerns that may arise when a single administrative record source is used are multiplied when linkage is made to other sources. In such cases, the subjects may not be aware that information supplied on two separate occasions is being combined. The Policy on RL is designed to ensure that the public value of each record linkage truly outweighs any intrusion on privacy that it represents.
Select the type of linkage methodology in accordance with the objectives	When a common matching key for both sources is not available and RL techniques are used. In this case, select the type of linkage methodology in accordance with the objectives of the statistical program (e.g. exact matching or statistical matching).
Exact matching	For frame creation and maintenance, or data editing, exact matching should be used. In the case of imputation or weighting, exact matching should be used, but statistical matching can be also sufficient.
Statistical matching	For performing some data analyses that are otherwise impossible, consider statistical matching, e.g. the matching of records with similar statistical properties (see Cox and Boruch, 1988; Kovacevic, 1999).
Make appropriate use of existing software	Statistics Canada's Generalized Record Linkage Software is but one example of a number of well-documented packages.

Reconcile potential differences	When data from more than one administrative source are combined, pay additional attention to reconcile potential differences in their concepts, definitions, reference dates, coverage, and the data quality standards applied at each data source.
Special attention with longitudinal data	Some administrative data are longitudinal in nature (e.g. income tax, goods and services tax). When records from different reference periods are linked, they are very rich data mines for researchers. Remain especially vigilant when creating such longitudinal and person-oriented databases, as their use raises very serious privacy concerns.
Detect change of ID over time	Use the identifier with care, as a unit may change identifiers over time. Track down such changes to ensure proper temporal data analysis. In some instances, the same unit may have two or more identifiers for the same reference period, thus introducing duplication in the administrative file. If this occurs, develop a non-duplication mechanism.

2

Operational aspects

2.1 INTRODUCTION

This Chapter emphasizes that the quality of a data integration process also depends on certain relevant operational aspects. The main actions to be considered are:

1. preparation of the integration project;
2. ensuring adequate data protection;
3. preliminary investigation of the likelihood that the data source is of sufficient quality for the objectives of data integration;
4. performance of an adequate procedure for obtaining external data;
5. preparation of data for record linkage and cleaning of the linking variables;
6. preparation of the documentation for quality assessment.

These actions are relevant for all Cases (A, B and C) of data integration and are good practices for developing a record linkage procedure (described in the third Part of this report).

2.2 PREPARATION OF THE INTEGRATION PROJECT

The quality of the data integration process is linked to the development of a specific project, aimed at verifying its feasibility from a technical and institutional point of view. In the box below, we mention the *key steps* of a data integration process, as proposed by Statistics New Zealand.

Key steps in a data integration project (Statistics New Zealand, 2006).

Through experience gained over the various data integration projects, Statistics NZ have identified the following key steps that must be undertaken for a successful outcome.

1. Develop clearly defined objectives
2. Address legal, policy, privacy, and security issues
3. Define governance structures and establish relationships with data providers and data users
4. Gain a thorough understanding of data sources
5. Decide how to carry out the matching
6. Define and build information technology (IT) data storage and processing requirements
7. Obtain the source data
8. Carry out the matching
9. Validate the matching and provide quality measures
10. Consider provision of access to micro-data and confidentiality of published outputs
11. Carry out the analysis and disseminate results.

It is essential to highlight the significance of some of these steps.

- First, *the definition of the purpose and stakeholder consultation.* The data integration must explain how these purposes will produce or improve official statistics, in terms of quality. As part of delivering a data integration project, it is necessary to consult with all stakeholders.

- A key component of a data integration project is the *Privacy Impact Assessment (PIA)*, to be carried out in accordance with national laws. PIAs for integration projects generally include a description of the procedures to be followed for collection, use, disclosure and retention of personal information. This is related mainly to steps 2 and 10 of the box above, but all others (governance, the matching procedure, the transmission and storage of data, the IT solutions, and dissemination) are also involved.

- Generally, data integration projects involve data from external agencies, and produce outputs that are of wide interest outside the organization. Therefore, in a data integration project, *relationships with external groups* are therefore a significant source of possible criticism. In a feasibility study, these issues would be developed in steps 3 and 7.

2.3 ENSURING ADEQUATE DATA PROTECTION

When data are linked for *statistical purposes*, the units are identified only to the extent necessary for making the link. When the linkage is complete, the unit's identity is no longer of statistical interest. The linked dataset is used to report statistical findings on the population or the sub-populations.

On the other hand, when data are linked for *administrative purposes*, individuals are identified not only to enable the link, but also for administrative use.

This may result in adverse actions against individuals, such as prosecutions, mandatory payment of fees, licensing of commercial activities, historical court sentencing and other penalties. Generally, a National or International Organization that produces and disseminates statistics undertakes data integration for statistical purposes only. In the box below, we quote the Guidelines to Ensure Protection of Linked Data proposed by Statistics New Zealand. This institution also uses a security checklist to ensure the protection of linked data, developed by Maxwell (2006) on the basis on international practice.

Guidelines to Ensure Protection of Linked Data, Statistics New Zealand

- Information about individual records cannot be sent to data providers.
- Names and addresses can only be retained in an integrated dataset for a limited period (if this is approved in the data integration project).
- Unique identifiers assigned by an external agency must be removed immediately after integration.
- Unique identifiers assigned by an external agency cannot be used for longitudinal linking.
- All data integration projects must have exclusive use of their own physical servers for processing and exclusive use of their own physical disks for storage, and be accessible only to the smallest practical number of the internal employees.

2.4 PRELIMINARY INVESTIGATION OF THE LIKELIHOOD THAT THE DATA SOURCE IS OF SUFFICIENT QUALITY FOR THE OBJECTIVES OF DATA INTEGRATION

It is often remarked that the actual process of *performing* the record linkage is only a small part of the overall data integration project. Gill (2001) estimated that in the implementation of record linkage:

- 75 percent of the effort consists in preparing the input files
- 5 percent of the effort in carrying out the linkage itself
- 20 percent of the effort in checking the results of the linkage.

A thorough understanding of the source data is fundamental, if meaningful results from the analysis of integrated data are to be obtained. It is commonly noted that the

comprehension of new data sources is time-consuming and resource-intensive. The investigation can be undertaken in several phases.

The investigation's initial phases could focus on the following topics:

- Statistical units, population, definition and classification of variables, time and geographical references of data collection. These can be derived from available documentation and contact with the producer, without access to actual data files.
- Metadata and documentation for quality assessment. This should be compiled on the basis of the documentation available. Metadata quality is crucial for good data integration.

If external data is required, *meetings with data providers and field visits* to data collection agencies or data entry points should be planned. At this early stage in the project of data integration, information can be considered as the starting point, *because it is* used to determine whether the data source is likely to be of sufficient quallity for meeting the objectives of data integration.

2.5 PERFORMANCE OF AN ADEQUATE PROCEDURE FOR OBTAINING EXTERNAL DATA

Once the preliminary investigation on the data source quality (see above, para. 2.3) has been performed, it is possible to plan the procedure for obtaining data. According to Statistics New Zealand (2006), this procedure should include the following steps: a) Request for supply of data; b) Data Transfer; c) Data verification d) Feedback to Provider. These activities are described below, mainly with reference to the experience of Statistics New Zealand.

Request for supply of data. The specifications to be included in the formal request are: the variables requested, the corresponding format, periodicity and timing of delivery, the transport mechanism, guidance on handling missing data, information on data quality and on the responsibility for cleaning the data.

In particular, the request for a data extract should include the following items:

- content of the file: population, time period, fields required
- how the file will be formatted
- checklist for the extract prior to its delivery (valid data values, range checks, etc.)
- means of data delivery

The source agency must:

- ensure that all variables required are identified, specified and supplied as agreed

- provide the relevant documentation relating to the dataset
- provide information on who currently, or potentially, has access to the dataset (to establish confidentiality requirements)
- advise the NSI of any changes in its collection mode or classifications

The data request and supply processes may need to be iterative, with modifications or corrections made to the data supplied as required. It is recommended that the specification be tested first, by transferring a brief version of the full dataset.

Data transfer. In data integration projects, various transmission modes can be considered: by email, other telematics, courier or carried by hand. The means to ensure data security for these media have been variable.

Data verification. Upon receiving a data extract, a number of checks can be performed to verify that:

- the number of records extracted is equal to the number received
- there are no duplicate unique identifiers
- numerical fields contain numbers, and text fields are predominantly text
- all variables requested are present, and whether any extra variables have been provided by mistake
- the range of values in each field is appropriate, and there are no unusual or surprising values
- the distribution of values in each field is as expected
- there is consistency with other fields within the data
- the relationship between files is as expected (only relevant if more than one file was supplied).

Feedback to provider. If the data was successfully validated, the data custodian should inform the provider that the data transfer and validation were successful. If the data transfer or data validation failed, the provider must be informed as to the reasons for the failure, and a new set of data must be requested.

2.6 PREPARING DATA FOR RECORD LINKAGE AND CLEANING THE LINKING VARIABLES

When preparing data for record linkage, a number of issues must be addressed. Often, data is recorded or captured in different formats and classifications, and data items may be missing or contain errors. A pre-processing phase for editing and standardizing the data is therefore an essential first step in all linkage processes. Datasets may also contain duplicate entries, in which case linkage may need to be applied within a dataset, to de-duplicate it before attempting linkage with other files.

The key actions in record linkage are:

- choice of linking variables
- improvement of the quality of the data sets to be integrated
- standardization: editing, parsing, formatting, concordance
- de-duplication
- anonymization of the unique identifiers (UID)

Choice of linking variables. The linking variable is a common unique identifier (UID), where one exists or is available; when no UID is available, the decision to link each pair of units is made on the basis of the combination of variables available in both databases.

The UID should be:

- *universal* (all units must respond to these variables),
- *permanent* (or unchangeable in time),
- *accurate* (even though quality problems are often unavoidable),
- *not sensitive* (they do not infringe the units' right to privacy).

It is not always possible to satisfy all of these requirements.

Data quality improvement in the databases to be integrated. To avoid matching errors, every effort must be made to ensure that the databases available are extremely accurate. In particular, it is necessary to promote the accuracy and completeness of the linking variables.

Establishing whether a UID is available is an essential first step in choosing the linkage method. As noted in the literature, there are two key methods for record linkage (Istat, 2008):

- *Exact linkage* involves using a unique identifier (e.g. a tax number, passport number or driver's license number) that is present in both files for linking records. This is the easiest and most efficient way to link datasets, and standard statistical software can be used.

- *Probabilistic linkage* is employed when a unique identifier is not available, or is not of sufficient quality or coverage to be relied on alone. This requires use of other variables common to both files, i.e. linking variables for linkage (for example names, addresses, date of birth and sex). Probabilistic linking is more complex; sophisticated data integration software is required to achieve high-quality results.

If using probabilistic linkage, the quality of linking variables is essential. Errors in linking variables may occur during the capture and processing of these variables. Sources of error in the linking variables include: variation in spellings, data coding and preparation, use of nicknames, Anglicization of foreign names, use of initials, truncation or abbreviation of names and addresses, use of compound names, missing words and extra words (Gill, 2001).

The *errors* recurring in the most commonly used linking variables are illustrated below (Statistics New Zealand, 2006).

Unique numeric identifiers (Numeric UID). Unique numeric identifiers, when available, can be excellent linking variables. However, very strict control over the issue of new identifiers and recording in the data file is necessary, if high-quality linkage is to be produced on the basis of the numeric identifier alone. Typical errors include: missing identifiers (particularly important where links are longitudinal); transcription errors in the recording of data, such as the transposition of digits; use of the same identifier for more than one unit; assignment of more than one identifier to the same unit (duplicates); 'units' may refer to different identities in different files. Numerical identifiers that include a 'check digit' are much less likely to be recorded incorrectly.

Surnames. Name changes due to marriage or divorce pose perhaps the main difficulty. Certain ethnic groups may use many surnames, and the order of their recording may vary. Concatenation of the birth surname and the marriage or partnership name into a compound (or hyphenated) name is common, so that both parts are required for linking purposes. Spelling variation is quite common in surnames, due to the effects of transcription of the names in various systems. In some cultures, there is no exact equivalent of a surname (Gill, 2001).

First names. There are wide variations in the spellings of first names due to recording and transcription errors. Common problems include the use of nicknames and contractions. Some are readily identifiable (Will for William, Liz for Elizabeth), but others are not (Ginger for Paul, Blondie for Jane).

Address. This is an excellent variable for confirming otherwise questionable links. However, disagreements are difficult to interpret, because of address changes, address variations and differences in mailing and physical addresses (Gill, 2001).

Sex. Sex is generally well-reported and, except for transcription and recording errors, is a very reliable variable. The main difficulty is that in some administrative records, sex may not always be available. For example, some databases do not collect this variable,

such that it can be generated only by recording the first name, which cannot be done with complete accuracy (Gill, 2001).

Date of birth. Date of birth is generally well-reported. Problems may occur when the date of birth is filled in for others (i.e. by proxy), for example for children and the elderly, when an approximation may be given. Typical transcription errors arise when only the day and the month, or when only two digits for the year, have been transposed. Other common errors are when the current date is entered mistakenly in the date of birth field, or the current year in the birth year field.

Another problem encountered in using linking variables is the *swapping of first names with surnames.* Occasionally, surnames and first names are exchanged. Also, *titles may be embedded in the name*: surname and first name fields may contain titles such as 'Mr', 'Mrs', 'Dr', 'Jr', etc. Before the names can be used for linking, they should be parsed and their various components identified and separated (Gill, 2001).

Standardization: editing, parsing, formatting, concordance. The success of a data integration exercise depends upon the use of standardized data fields.

Because of potential quality problems, some variables may not be suitable for linking.

Rigorous editing, parsing and formatting of linking variables and creation of concordances must be undertaken to minimize errors. These terms can be described briefly as follows:

- Editing is the process of detecting and dealing with erroneous or suspicious data.
- Parsing a field means separating the entities within that field to facilitate comparison.
- Formatting is necessary when the fields are recorded in different formats in the various files, such as the date of birth.

Creating a concordance means to achieve consistent coding across files.

Deduplication. Duplicate records are common in administrative datasets. The impact of duplicates upon integration depends on the frequency of duplicates, how duplicates are generated and the type of integrated dataset being created. False positives will occur if the duplicates are linked to the wrong unit. If the resulting integrated dataset consists of the intersection of the source files, then unlinked duplicates will appear as false negative links. Great care should be taken where the integrated data consists of the union of the source files, as the unlinked duplicates will inflate the number of cases in the final integrated file.

Anonymization of unique identifiers (UID). The use of unique identifiers (UIDs) assigned by other agencies must meet the requirements of the NSI's Data Integration Policy. Any UID assigned by other agencies and passed to an NSI as part of an integration project will be converted to an internally-assigned unique identifier (IUID) as soon as is practicably possible. External UIDs will be retained within the NSI systems (servers, databases and applications) only for the time strictly required for performing

validation, editing and integration. They will then be wholly replaced and removed. An externally-assigned UID will not be used for longitudinal linking. The IUID provides the capacity to create a consistent longitudinal link to the same unit, without the need for the NSI to store the original UID in any of its production databases.

2.7 PREPARING THE DOCUMENTATION FOR QUALITY ASSESSMENT

Any matching exercise should be accompanied by full documentation of the method used for record linkage.

This 'technical description' of the matching methodology has two main uses:

- to enable peer review of the methodology;
- to record what has been done for the future.

Peer review of the record linkage methodology is required to confirm that the work has been performed soundly. Ideally, the review should be done before the linked data are consigned to the clients, so that any improvements in methodology suggested by the reviewer can be made. This implies a two-stage process, in which the first results obtained are essentially a trial. If it is not possible to perform the review beforehand, the reviewer's suggestions can be considered for future processes. The peer review of the matching methodology, including the time and resources allocated, should be included in the initial project planning.

All output data of integration should be supported by adequate metadata, to enable users to adequately comprehend the data. This may include information on:

- the context of the source data;
- data processing;
- quality indicators, including match rates;
- the format of the output data, including detailed information on variables;
- advice on how to obtain the technical description of the matching.

According to Belin and Rubin (1995), a record linkage procedure is an algorithmic technique that can identify which pairs of records from two databases correspond to the same unit. *Methods for record linkage can be divided into two main groups:*

- *heuristic methods*, typically prepared by the computer, requiring the intervention of specialized personnel for the manual control of records. This method is often very expensive.
- *statistical methods*, which allow the quality of the results to be assessed. Decisions are based on procedures where the possibility of making linking errors is checked. This aspect leads to a formalization of the problem of record linkage

and its discussion in terms of decision-making and statistical methods (Falorsi *et al*, 2005; Liseo *et al.*, 2006).

Therefore, it is the choice of statistical methods that enables a quality assessment of linked data to be conducted.

2.8 THE TWO PHASES OF THE RECORD LINKAGE PROCEDURE

The first phase provides a solution to the decision-making problem, which involves the use of a tool for ascertaining whether two records refer to the same unit. Fellegi and Sunter (1969) defined an optimal rule. This tool also includes the calculation of the level of error (probability), which is essential for decision-making.

The second phase is a statistical phase, and concerns the estimation of the elements required to construct decision rules.

False positives, false negatives and match rates

In the literature on data integration, the errors typically occurring in the operation of record linkage are well known. Two records are considered a link when it is determined that they refer to the same unit. Not every match is a link, and not every link is a match, as outlined in the following table.

	True Match	True Non-Match
Link	Correct Outcome	False Positive Link
Non-Link	False Negative Link	Correct Outcome

Both linking methods can result in two types of errors: false positive matches and false negative matches.

- A false positive match is where two records are linked together, but they are not, actually, the same unit (person, household, farm, etc).
- A false negative match is where two records are not linked together, but they actually do belong to the same person or unit.
- Generally, there is a trade-off between the two types of errors since, for example, reducing the rate of false positives may increase the rate of false negatives. Therefore, it is important to consider the consequences of each type of error, and to determine whether one is more critical than the other.
- An assessment of the size of each of these sources of linkage error should be undertaken as part of the integration, and its results made available. Analysis of an integrated dataset should take into account the possible impact of the linkage error.

Generally, there is no good method for ensuring an automatic estimation of error rates; therefore, false positive rates are estimated by manually checking samples of linked records. In large datasets, an analysis of false positives can be time-consuming, and it is often advisable to group the linked data prior to selecting a sample.

If at least one of the files is expected to match completely, and the false positive rate is low, then the false negative rate may be calculated simply as one minus the match rate (where the match rate for a given file is the number of matched records over total records). However, in other situations, such as when the integrated dataset consists of the union of two files, expected matches are unknown and the false negative rate is difficult to estimate.

A clerical review of these groups of records can be useful for understanding problems, but necessarily involves the subjective view of the reviewer. If it is possible to understand where errors in the datasets are most likely to occur, it may be necessary to target the sample in these areas, with a view to improving the quality of the match. Several iterations of a clerical review and adjustments of matching criteria may be required before a linked dataset is confirmed, and final false positive error rates are calculated.

If at least one of the files is expected to match completely, and the false positive rate is low, then the false negative rate may be calculated simply as one minus the match rate (where the match rate for a given file is the number of matched records over total records). However, in other situations, such as when the integrated dataset is the union of two files, expected matches are unknown and the false negative rate is difficult to estimate.

The *rule for determining cut-off thresholds* is important in relation to the trade-off between the level of false positive and false negative matches. When determining cut-off thresholds, it is also important to consider the objectives of the matching exercise. For example, if it is essential to avoid false matches, then the cut-off threshold can be higher, bearing in mind the fact that some true matches will be missed.

The (non-negative) cut-off threshold is the composite weight value that distinguishes the links analysts consider to be matches from those that are not considered matches. All record pairs whose composite weight is greater than or equal to the cut-off are considered links. Deciding on the cut-off value is one of the most difficult tasks that analysts face in a data integration project, because the boundary is not straightforward. It is commonly acknowledged that even experienced analysts can produce significantly different linked outputs.

2.9 MEASUREMENT ERROR IN DATA INTEGRATION

Measurement error affects inferences and can lead to severe bias in estimation. In data analysis, best practice procedures examine the data used for measurement errors, and properties of known measurement errors are incorporated into the analysis (Chesher and Nesheim, 2004).

The measurement error processes that arise when there is probabilistic record linkage are complex and non-standard. Chesher and Nesheim (2004) list possible causes of measurement error in data linking:

- units incorrectly linked, such that data from one unit is incorrectly associated with another unit (false positive links)
- in many-to-one linking, statistics computed using only a few sub-units are used to measure characteristics of all sub-units
- in many-to-one linking, characteristics of sub-units are inferred from features of major units (and vice versa).

Chesher and Nesheim also state that from a practical perspective, measurement error is inevitable. Since the potential effects are so damaging, data-linking procedures that are likely to generate large amounts of measurement error should be avoided.

The first step in estimating the quality of linked datasets is often the estimation of the rates of false positives and false negatives. In record linkage projects carried out by NSIs, quality measurement has focused on these two dimensions of quality, with the aim of minimizing false positive links.

Annex

Brief Glossary on Data Integration

Microdata	A file consisting of a record for each unit (Unit record data). This is the lowest level of data available.
Reporting unit	Level at which the data source is provided. It may not be the same as the *integration unit*.
Integration unit	Level at which the data is integrated.
Integration input dataset	A dataset containing data that has been edited, parsed and standardized in view of integration.
Integrated dataset	The dataset resulting from record linkage.
Unique identifier (UI or UID)	A variable that uniquely identifies a person, place, event or other unit.
Linking variable	Variables used to compare two records, including both blocking variables and matching variables.
Matching variable	Variables used to compare two records that fall within the same block, to examine the likelihood that the two records belong to the same unit.
Blocking variables	Variables used to divide a file into blocks.
Record linkage	The combination of data, from different sources, on the same individual or unit, or on a similar individual or unit, at the level of individual unit records. (record linkage is synonymous with *data integration* at the micro level).
Probabilistic record linkage	Record linkage methodology based on the relative likelihood that two records belong to the same unit, given a set of similarities/differences between the values of the linking variables (e.g. name, date of birth, sex) on the two records.
Bayesian record linkage	Record linkage methodology based both on the likelihood that two records belong to the same unit, and on prior experience from previous record linkage operations. It is based on the posterior distribution of parameters, rather than on likelihood alone.
Statistical matching	Statistical matching occurs at the unit-record level, but does not necessarily link records of the same person. In statistical matching, a unit record for one individual is linked to one or more records for similar individuals in other datasets, on a probabilistic basis.
Stochastic matching	Matching groups from two different datasets based on similar characteristics, with the assumption that these individuals will act in the same way. This procedure is useful for creating synthetic datasets.
Link	A decision that two records belong to the same unit.
Non-link	With reference to record linkage, a decision that two records do not correspond to the same unit.
True match	Two records that truly correspond to the same unit.
True non-match	Two records that truly do not correspond to the same unit (e.g. two different people).
Linked	The status of a record that has undergone the integration process and was linked to a record from another file.
Unlinked	The status of a record that has undergone an integration process and was not linked to a record from another file.

PART 3

Record Linkage

Record linkage is a collection of techniques aimed at identifying data records, held on two different electronic files, that contain information on the same "entity". Record linkage is performed for essentially two reasons: data collection and list construction. Both of these are among the most crucial tasks of National Statistics Institutes, other National and International bodies and private users.

The use of record linkage techniques poses several interesting problems, from both methodological and computational points of view. From the methodological perspective, the very definition of a statistical model (the description of how comparisons between records should be performed) is still open to debate: see, for example, Fellegi and Sunter (1969), Copas and Hilton, (1990); Belin and Rubin (1995); Fortini et al. (1990), and Tancredi and Liseo (2011). From the computational perspective, problems become daunting once the databases reach a large size (over 100 units). One of most popular solutions in this respect is to perform comparisons only between records showing the same values, on certain "blocking variables" that are assumed to have been recorded without errors: therefore, at least partial resolution of this problem is crucial. In this Report, we review the general theory of probabilistic record linkage, with a special emphasis on a Bayesian perspective, to construct statistical models that are appropriate for linkage purposes.

While it is certainly true that the result of a statistical analysis produced by an official body "must" be objective (or should at least be perceived as such, by users), it also cannot be denied that Bayesian ideas and techniques can play an important part in official statistics in the following circumstances: (i) when important prior (or extra-experimental) information on the variables of interest exists and cannot be adequately exploited in a classical inference framework; (ii) even when prior information is lacking, a Bayesian analysis may be necessary, if only because a classical approach is not capable of providing answers unless strong assumptions, which cannot easily be tested, are introduced. In these cases, a Bayesian analysis enables at least performance of a sensitivity analysis, with the aim of quantifying the influence of assumptions on the inferences drawn.

In this Chapter, we perform a record linkage exercise on a real data set. A limited data set comprising the records observed in the same enumeration area, across two different agricultural surveys, will be used, to illustrate the different record linkage methodologies described in the previous Chapter. We will also propose an alternative strategy that is more specifically tailored to the data at hand. In the final Section, we will extend the record linkage analysis to all enumeration areas common to the two surveys, and we will illustrate the results of two statistical inference procedures based on the resulting integrated data set.

1

Theory

1.1 INTRODUCTION TO RECORD LINKAGE TECHNIQUES

Today, the need for increasingly detailed and timely statistical information is shared by several international (European Union, European Central Bank, International Monetary Fund, etc.) and national (National Statistical Service, Ministries, Regions, etc.) bodies, and private users too. In this respect, increasing computational potential provides important opportunities. It is now possible to collect and maintain massive amounts of statistical data obtained from surveys; it is also possible to retrieve data from pre-existing public and/or private archives. Examples of exploitation of the latter opportunity include the Italian Statistical Archive of Companies (ASIA) and the construction of business data archives. In this context, a significant problem consists of the need to merge various data archives, possibly in view of the fulfilment of different goals.

The main aim of this report is to review the use of new statistical tools for the production of statistical information, through the integration of non-homogeneous databases. The scientific community's awareness of this problem is testified by the numerous international symposia on "combining data from different sources", and the special sessions devoted to these topics at the Joint Statistical Meetings of recent years. The main statistical approaches employed to address these problems can be classified as:

- Record linkage
- Statistical matching

The latter technique seeks to derive integrated statistical information by combining information from different datasets, in which only some variables are observed twice, and no overlapping of observed units is necessary. In this Report, we will focus on the former approach.

Record linkage refers to the use of specific algorithms that aim to identify pairs of records, corresponding to a single statistical unit, that are present in different databases. The same problem is addressed – albeit in a more general manner – in information technology literature, as the problem of integrating non-aggregated databases. In this context, relevant issues are (i) the construction of a general framework (ii) the detection and specification of semantic relationships between non-homogeneous data sources (iii) the characterization of data quality factors and (iv) the reconciliation of datasets from

different sources, to construct a representation that is coherent with the relevant general framework and quality requirements. Statistical matching techniques aim to estimate the joint distribution of several variables, that may be observed in different databases. The following applications of record linkage methodologies deserve mention:

1) The construction and maintenance of a list of statistical units, to be used as a "reference population" in sample and/or total surveys. The ASIA archive mentioned above is an example of this methodology. In this context, it is important to identify units that feature in more than one database.

2) The merging of two or more databases to obtain a single archive, which is more informative at a non-aggregated level. This makes it possible to perform statistical analyses that would otherwise be difficult.

3) The use of several data sources for the improvement of the overall survey's "coverage". The methodological implications of these problems are not yet well-developed, but it is certain that the information provided by administrative data archives can be of great assistance in this regard.

4) The estimation of the size of a population via capture-recapture methods. A relevant example is the estimation of the under-coverage given by an overall survey: this is usually performed via a linkage analysis between total survey data and an *ad-hoc* post-enumeration survey (Winkler, 1986).

5) The evaluation of the validity of a disclosure method, to protect access to administrative data from the risk of identification of single units by an intruder (Duncan and Lambert 1989; Winkler 1998).

Important reference materials on these and other applications are the Proceedings of two international conferences held in the United States, in 1985 and 1997 (Kills and Alvey 1985; Alvey and Jamerson 1997).

The use of record linkage techniques poses several interesting problems, in both methodological and computational terms. From the methodological point of view, the very definition of a statistical model (the description of how comparisons between records are performed) is still debated: see, for example, Fellegi and Sunter (1969), Copas and Hilton (1990); Belin and Rubin (1995); Fortini *et al*. (1990), and Tancredi and Liseo (2011). From the computational perspective, problems become formidable once the databases reach a large size (over 100 units). In these cases, comparisons are performed only between records that have the same values for certain "blocking variables", which are assumed to have been recorded without errors. Overcoming this problem appears, therefore, to be crucial.

In recent years, we have experienced a great proliferation of new Bayesian methodologies and, especially, an increasing number of statistical applications performed from a Bayesian perspective. The main reason for this trend lies in the development of Monte Carlo Markov Chain methods which enable building and calibrating virtually any statistical model, regardless of complexity. This opportunity has made Bayesian methods much more appealing and visible in many areas of application, including official statistics. National Statistics Institutes (henceforth NSIs) have several

important and complex tasks; for their practical implementation, different kinds of – more or less – subjective operational decisions must be taken. For example, several important economic and social indexes are the result of procedures that at least implicitly involve the use of complex statistical models. Nevertheless, the result of a statistical analysis performed by NSIs "must" be objective or, at least, should be perceived as such by users.

For these reasons, the use of Bayesian methods in official statistics is the subject of an important debate (see e.g. the special issue of *Research in Official Statistics* (2001) entirely devoted to the topic). Bayesian concepts can be important for official statistics when (i) important prior (or extra-experimental) information on the variables of interest exist, and cannot be exploited adequately in a classical inference framework; and (ii) even when prior information is missing, a Bayesian analysis can be required, because a classical approach cannot provide answers unless strong assumptions, not easily tested, are introduced. In these situations, a Bayesian analysis enables at least a sensitivity analysis, to quantify the influence of the assumptions on the inferences made.

1.2 USES OF RECORD LINKAGE

Record linkage refers to the use of an algorithmic technique to match records from different datasets that correspond to the same statistical unit, but that lack a unique personal identification code. The need for record linkage techniques is steadily rising in various areas of statistics. For example, in official statistics, record linkage is a necessary preliminary step when estimating the size of a population through capture-recapture techniques, especially when the target population is elusive (e.g. the number of irregular immigrants into the European Union), and differences in identification variables are the rule rather than the exception.

Another particularly important example for Statistical Institutes is the use of administrative databases, to integrate information obtained from surveys, thereby relieving response burden. From a broader perspective, many Statistical Institutes and agencies use file merging to create comprehensive files from multiple incomplete data sources.

The main objective of this endeavour is to perform statistical analyses on the synthetic dataset, generated by file merging, which could not be performed by analyzing the incomplete datasets separately. In theory, the validity and efficacy of file merging could be assessed by means of statistical models representing the mechanisms that generate the incomplete datasets. However, there is as yet no complete, satisfactory theory of record linkage procedures.

In general, from the point of view of statistical methodology, merging two (or more) data files can be important for two reasons:

- *per se*, to obtain a larger and integrated reference dataset; and

- to enable performance of a subsequent statistical analysis, based on the additional information obtained, that cannot be extracted from either of the two individual data files.

We will now give a toy example of the latter reason, which will be discussed in further detail in Section 5 below. Suppose that we have two computer files, A and B, whose records relate to different units (e.g. individuals, firms) of partially overlapping populations P_A and P_B. The two files consist of several fields or variables, either quantitative or qualitative.

For example, in a file of individuals, fields can be "surname", "age", "sex", etc. The goal of record linkage is to detect all the pairs of units (a,b), $a \in A$ and $b \in B$ such that a and b actually refer to the same unit. Suppose that the variables observed in A are denoted by

$$(Z, W_1, W_2, \cdots, W_k)$$

while we observe

$$(W_1, W_2, \cdots, W_k, X)$$

in file B. Then, we might be interested in studying a linear regression analysis (or any other more complex association model) between Z and X, restricted to the pairs of records that are declared matches. The intrinsic difficulties of such a simple operation are well-documented and discussed in Scheuren and W. E. Winkler (1993) and Lahiri and Larsen (2005).

In statistical practice, it is quite common for the *linker* (the researcher matching the two files) and the *analyst* (the statistician performing the subsequent analysis) to be two different persons working separately. However, we agree with Scheuren and W. E. Winkler (1993), according to whom

"*it is important to conceptualize the linkage and analysis steps as part of a single statistical system and to devise appropriate strategies accordingly.*"

In a more general framework, suppose that file A contains variables $(\boldsymbol{Z}, \boldsymbol{W}_A) = (Z_1, Z_2, \cdots Z_h, W_1, W_2, \cdots, W_k)$ observed on ν_A units, while X_B contains the variables $(\boldsymbol{W}_B, \boldsymbol{X}) = (W_1, W_2, \cdots, W_k, X_1, X_2, X_p)$ observed on ν_B units. Our goal can be stated as follows:

- to use the key variables $(W_1, W_2, \cdots, W_k)$ to detect the true links between X_A and X_B
- to perform a statistical analysis based on vectors of variables $\boldsymbol{Z}$ and $\boldsymbol{X}$ restricted to the records that have been defined matches

To perform this task, we present a fully Bayesian analysis, which is particularly suitable for accomplishing the objectives described above. Fundamentally, in our approach, all uncertainty related to the matching process is automatically retained in subsequent inferential steps.

1.3 CLASSICAL RECORD LINKAGE THEORY: THE JARO APPROACH

We first examine the classical approach to the problems of record linkage. Consider two data files A and B, with, respectively, ν_A and ν_B units. Let us call A and B the two sets (lists) of observed units, $a = 1,\cdots,\nu_A$, $b = 1,\cdots,\nu_B$. We assume that at least some units are present in both lists. The set of all ordered pairs

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

can be logically split into two non-overlapping sets, namely

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

as the set of matches, and

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

as the set of non-matches. To decide whether a specific pair (a,b) is actually a member of M or U, we may compare variables observed in both files (e.g. surname, name, sex, address, etc. for individuals): these are called *key* variables. Let us assume that we have k key variables, $k \geq 1$, whose realizations in the two data lists are denoted by:

$$w_a = (w_{a,1}, w_{a,2}, ..., w_{a,k}), \qquad a \in A,$$

and

$$w_b = (w_{b,1}, w_{b,2}, ..., w_{b,k}), \qquad b \in B.$$

We denote with $Y_{ab}^{(j)}$, $j = 1,\ldots,k$, the result of the comparison between the values $w_{a,j}$ and $w_{b,j}$. In general, the comparison $Y_{ab}^{(j)}$ can be any function of $w_{a,j}$ and $w_{b,j}$. The most commonly assumed comparison function takes the form of a vector of k elements, $Y_{ab} = (y_{ab}^{(1)}, ..., y_{ab}^{(k)})$ with:

$$y_{ab}^{(j)} = \begin{cases} 1 & \text{if } w_{a,j} = w_{b,j} \\ 0 & \text{otherwise.} \end{cases} \qquad j = 1, \cdots, k. \tag{1.1}$$

More general and sensitive comparison functions can also be used, especially in the case of continuous key variables. However, the 0/1 comparisons are compatible with a reasonably fast and accurate matching process. A simple and not excessively expensive generalization of this dichotomy can be used, by discretizing the observed values into a small number of classes. We will discuss this issue in further detail in the final section.

Another, radically different, approach is based on actual observations taken from files A and B, rather than considering comparisons between them. Copas and Hilton (1990) address this issue. We sketch a possible extension of their ideas in Section 1.4.

In the 0/1 case, the comparison vector y_{ab} can assume 2^k different values, which we indicate with y_i, where $\boldsymbol{i} = 1, \ldots, 2^k$. To decide whether a pair (a,b) with comparison vector y_{ab} should be linked, Fellegi and Sunter (1969) suggest considering the sampling distribution of the comparison vectors in *M*, say $m(y)$, and the corresponding distribution in *U*, $u(y)$. The decision rule for the pair (a,b) is based on the likelihood ratio

$$t(y_{ab}) = \frac{m(y_{ab})}{u(y_{ab})}. \tag{1.2}$$

Fellegi and Sunter (1969) discuss several frequentist optimality properties of this decision rule. Given that neither $m(y)$ nor $u(y)$ are known, most literature on record linkage focuses on methods for their estimation. Starting with Jaro (1989), a model-based approach is advocated for this task. The usual assumption is that the status of a pair (e.g. C_{ab}, where $C_{ab} = 1$ when a pair (a,b) is a true match and 0 otherwise) is a non-observable random variable, while the comparison vector Y represents the actual data. Also, a general latent structure is assumed via the configuration matrix $C = \{C_{ab}, a \in A, b \in B\}$, so that $C = \{C_{ab}, a \in A, b \in B\}$, so that

1. C_{ab}, $(a,b) \in A \times B$, are assumed to be i.i.d. Bernoulli r.v., such that for all a, b, $P(C_{ab} = 1) = p$;

2. the comparison vectors Y_{ab}, $(a,b) \in A \times B$, are assumed to be i.i.d. replications of the r.v. Y whose marginal distribution (with respect to $\boldsymbol{C}$) has the mixture structure

$$Pr(Y = y \mid p) = pm(y) + (1-p)u(y);$$

3. for fixed p, the random vectors (C_{ab}, Y_{ab}), $(a,b) \in A \times B$, are independent and identically distributed, with distribution, for $c = 0,1$,

$$Pr(C = c, Y = y) = [pm(y)]^c [(1-p)u(y)]^{1-c},$$

The independence assumptions are rather unrealistic: if $c_{ab} = 1$, then all other elements on row a and on column b must be 0. Nevertheless, independence makes it particularly easy to compute the likelihood function, given the $n_A \times n_B$ observations (c_{ab}, y_{ab}):

$$\prod_{(a,b) \in A \times B} (pm(y_{ab}))^{c_{ab}} ((1-p)u(y_{ab}))^{1-c_{ab}}. \qquad (1.3)$$

Maximum likelihood estimates of the distributions $m(y)$ and $u(y)$ can thus be obtained, using for example the EM algorithm, where the matrix C performs the role of *missing data*. Jaro (1989) assumes that the components of the comparison vector Y are mutually independent, while e.g. Winkler (2004) and Larsen and Rubin (2001) consider comparisons of dependent key variables.

We will now describe the approach proposed by Jaro (1989). Using the above formulation, let

$$m_i = Pr(\text{ key variable } X_i \text{ agree} \mid \text{pair} \in M)$$

and

$$u_i = Pr(\text{ key variable } X_i \text{ agree} \mid \text{pair} \in U).$$

For a given record pair r, if the key variable values agree, the weight for that variable will be $w_i = \log_2 m_i - \log_2 u_i$. If the key variable values disagree, the weight will be $w_i = \log_2 (1-m_i) - \log_2 (1-u_i)$.

For any record pair, a composite weight can be computed by adding the individual component weights arising from individual key variables. In general, given a pair, the value of m is larger than the value of u; consequently, fields that agree make a positive contribution to the sum, while fields that disagree make a negative contribution.

To reach a decision, three states are defined. A record pair is classified as a "match" if the composite weight is above a given threshold, a "non-match" if the composite weight is below another threshold, and an "undecided situation" if the

weight is between the two thresholds. The thresholds must be calculated in terms of the given level of false matches rate (FMR) and false non-matches rate (FNMR)[2] that one is willing to accept in the method's specific application.

In this procedure, the most important statistical issue is the accurate estimation of vectors $\boldsymbol{m}$ and $\boldsymbol{u}$. It is significantly easier to estimate quantities u_i rather than m_i, because the cardinality of set U is much larger than the cardinality of set M. For example, if the two sets of records both have size 100, and only 10 records belong to both files, then $\text{card}(M) = 10$ and $\text{card}(U) = 100 \times 100 - 10 = 9990$. In general, the contribution from M is ignored and the u_i's from the total agreement on each key variable are estimated.

Estimating the m_i is more complex: Jaro (1989) proposes a mixture model approach. After selecting some key variables that are supposed to have been observed without error[3], we are given k key variables and a sample of N record pairs drawn from the Cartesian product $A \times B$. As in Formula (1.1) above, let

$$y_{ab}^{(j)} = \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases}, \qquad a \in A; b \in B; i = 1, \ldots k,$$

let $\boldsymbol{y}_{ab}$ be the vector containing all comparisons related to pair (a,b), and $\boldsymbol{y}$ the vector containing all y_{ab}'s. Thus, it is possible to rephrase the m_i's and the u_i's as

$$m_i = Pr\left(y_{ab}^{(j)} = 1 \,|\, (a,b) \in M\right); \qquad u_i = Pr\left(y_{ab}^{(j)} = 1 \,|\, (a,b) \in U\right).$$

As above, let p denote the proportion of matched pairs, i.e. $\text{card}(M)/(card(M) + \text{card}(U))$. Then, each individual record has a distribution that is a combination of two components with unknown parameters $(\boldsymbol{m}, \boldsymbol{u}, p)$. Estimates of p and $\boldsymbol{m}$ are obtained through a standard application of the EM algorithm (Dempster *et al.*, 77). This algorithm is very stable and simple to implement; further details may be found in Jaro (1989). However, in the following section it will be argued that a Bayesian approach is more adequate in accounting for the uncertainty that is intrinsic in all linkage procedures. The final part of the procedure is the matching algorithm. After the weights and the thresholds have been computed, a single record of file A might be matched with

[2] FMR is the probability that a record pair is classified as a match when it is actually not; FNMR is the probability that a record pair is classified as a non-match when it is actually a match.

[3] This procedure is commonly known as "blocking". It enables the creation of smaller files.

more than one record of file B. Since this possibility has been ruled out by assumption, an operational technique must be applied to select the matches. It must be noted that this problem is well-known in operational research literature, and is called the "linear sum assignment problem". The problem can be restated in the following form

$$\max \sum_{i=1}^{k}\sum_{j=1}^{k} d_{ij} X_{ij},$$

subject to constraints $\sum_{j=1}^{k} X_{ij} = 1, i = 1,\ldots,k; \sum_{i=1}^{k} X_{ij} = 1, j = 1,\ldots,k$, where d_{ij} is the weight of matching record i on file A to record j on file B, and the X_{ij}'s are 0/1 variables (Burkard and Derigs, 1981).

1.4 THE BAYESIAN MODEL

The Bayesian model should be expressed in terms of a prior distribution on the unknown parameters, and in terms of the conditional distribution of the observed data, given the unknown parameters. The observed data are lexicographically ordered in the vector $y = (y_{11},\ldots,y_{v_a v_b})$ while the parameters are represented by the configuration matrix $\boldsymbol{C}$, the vector $\boldsymbol{m} = (m_1,\ldots,m_{2^k})$ where $m_i = P(Y_{ab} = y_i \mid c_{ab} = 1)$, and the vector $\boldsymbol{u} = (u_1,\ldots,u_{2^k})$, where $u_i = P(Y_{ab} = y_i \mid c_{ab} = 0)$.

The conditional distribution of y given the parameters $\boldsymbol{C},\boldsymbol{m},\boldsymbol{u}$ can be written as

$$f(\boldsymbol{y} \mid \boldsymbol{C},\boldsymbol{m},\boldsymbol{u}) = \prod_{a=1}^{v_A}\prod_{b=1}^{v_B} f(y_{ab} \mid \boldsymbol{C},\boldsymbol{m},\boldsymbol{u})$$

$$= \prod_{a=1}^{v_A}\prod_{b=1}^{v_B} f(y_{ab} \mid c_{ab},\boldsymbol{m},\boldsymbol{u}) \qquad (1.4)$$

$$= \prod_{a=1}^{v_A}\prod_{b=1}^{v_B}\left[\prod_{i=1}^{2^k} m_i^{d(y_{ab},y_i)}\right]^{c_{ab}}\left[\prod_{i=1}^{2^k} u_i^{d(y_{ab},y_i)}\right]^{1-c_{ab}}$$

where

$$d(y_{ab}, y_{\mathbf{i}}) = \begin{cases} 1 & \text{if } y_{ab} = y_{\mathbf{i}} \\ 0 & \text{if } y_{ab} = y_{\mathbf{i}}. \end{cases}$$

In the following parts, we will assume that $\boldsymbol{m}$ and $\boldsymbol{u}$ are *a priori* independent of $\boldsymbol{C}$. In the absence of specific prior information on vectors $\boldsymbol{m}$ and $\boldsymbol{u}$, it is reasonable to adopt a conjugate Dirichlet prior distribution for both $\boldsymbol{m}$ and $\boldsymbol{u}$. In particular,

$$\tilde{\mathbf{m}}\mathsf{D}(\alpha_1,\ldots,\alpha_{2^k}); \qquad \tilde{\mathbf{u}}\mathsf{D}(\beta_1,\ldots,\beta_{2^k}).$$

In addition, a hyper-structure must be introduced over the vectors $\tilde{\mathbf{m}}$ and $\tilde{\mathbf{u}}$ to *unsaturate* the model. Following Fortini *et al*. (1990), we set

$$\log\alpha_i = (\sum_{i=1}^{k} y_i^k - \phi)\log\theta, \qquad \log\beta_i = (\phi - \sum_{i=1}^{k} y_i^k)\log\theta. \quad (1.5)$$

The introduction of hyper-parameters θ and ϕ greatly simplifies the model and makes it non-saturated. The rationale for this reparametrization is its capacity to model our beliefs on the informative power of each comparison variable. Indeed, in the Equations (5) above, the hyper-parameters hierarchically order the comparison vectors' possible values such that, for example, the prior distribution for $\boldsymbol{m}$ places greater mass around "large values" of the m_i's, for those i's $\boldsymbol{i} = 1,2,\cdots,2^k$, with a large number of 1's in the observed comparison vector. The opposite argument is true for the hyper-parameters in the second equation of (5). Fortini *et al*. (1990) show that, by introducing the hyper-parameters ϕ and θ, the marginal prior means of the m_i's and the u_i's are simple functions of θ alone, while their variances depend on both θ and ϕ. The hyper-parameters (5) also have direct effects on the statistical relationships between the comparison variables. For example, the linear correlation between two comparison variables has a null expected value for any θ and ψ, while their variance depends on both. These considerations can guide the elicitation process for the hyperparameters. It must be noted that the above elicitation of the prior structure makes a limited use of the data information; therefore, in this sense alone, ours can be viewed as an empirical Bayes approach.

To complete the model, it is necessary to elicit a prior distribution for the configuration matrix $\boldsymbol{C}$. We assume that each individual record in A can match at most one record in B; then $\boldsymbol{C}$ must satisfy the natural constraints

$$c_{ab} \in \{0,1\}(a = 1,\cdots,A; b = 1,\cdots B), \qquad \sum_{a=1}^{\nu_A} c_{ab} \le 1, \quad \sum_{b=1}^{\nu_B} c_{ab} \le 1. \quad (1.6)$$

Let $T = \sum_{ab} c_{ab}$ denote the number of true matches in $\boldsymbol{C}$; also, let $T_m = \min\{\nu_A, \nu_B\}$ be the maximum possible number of matches. Also, we denote by $T_q =$ the quantity $\max\{\nu_A, \nu_B\}$. The prior distribution on $\boldsymbol{C}$ can be built up in two stages. First, we

assume that T, the number of matches, follows a binomial distribution with parameters ξ and T_m, that is

$$P(T=t)=\binom{T_m}{t}\xi^t(1-\xi)^{T_m-t}, \qquad t=0,1,T_m.$$

In the second stage, we assume that, conditional on $T=t$, the distribution over the space of all possible matrices $\boldsymbol{C}$ – satisfying the relevant constraints of Equation (6) – is uniform.

Then,

$$P(C\mid T=t)=\begin{cases}\left(\binom{T_m}{t}\binom{T_q}{t}t!\right)^{-1} & \text{if } \sum_{ab}c_{ab}=t\\ 0 & \text{otherwise}\end{cases}.$$

Note that the hyper-parameter ξ has a specific interpretation, since it represents the probability that a generic unit in the smaller dataset will also be present in the larger dataset. We can consider ξ either known or unknown. In the latter case, we assume that ξ follows a Beta(δ_1,δ_2) distribution. It can also be proved that $E(C_{ab})=p$, where $p=\xi/T_q$. Then, the quantity p can be interpreted as the probability that a randomly chosen pair (a,b) is actually a match. Our prior assumptions thus lead to the following posterior distribution for the parameters $(\boldsymbol{C},\boldsymbol{m},\boldsymbol{u},\xi)$ (recall that $\boldsymbol{m}=\boldsymbol{m}(\phi,\theta)$ and $\boldsymbol{u}=\boldsymbol{u}(\phi,\theta)$)

$$p(\boldsymbol{C},m,u,\xi\mid y)\propto\prod_{a=1}^{\nu_A}\prod_{b=1}^{\nu_B}\left[\prod_{i=1}^{2^k}m_i^{d(y_{ab},y_i)}\right]^{c_{ab}}\left[\prod_{i=1}^{2^k}u_i^{d(y_{ab},y_i)}\right]^{1-c_{ab}}$$
$$\times\frac{\xi^{\sum c_{ab}+\delta_1-1}(1-\xi)^{T_m+\delta_2-1-\sum c_{ab}}}{\binom{T_q}{\sum c_{ab}}\sum c_{ab}!}\prod_{i=1}^{2^k}m_i^{\alpha_i-1}u_i^{\beta_i-1}.$$

1.4.1 MCMC IMPLEMENTATION

The Bayesian model proposed in this Report is too complex to be amenable to analytical calculations. Hence, we turn to Monte Carlo Markov Chain methods, in particular to a Gibbs sample algorithm. Indeed, it is easy to show that

- the full conditional posterior distributions of the vector $\boldsymbol{m}$ and $\boldsymbol{u}$ are still Dirichlet distributed, while the full conditional of ξ is still Beta distributed;
- each individual entry of matrix $\boldsymbol{C}$ has a full conditional distribution (also given the other entries of the matrix), which is either Bernoulli or degenerate.

To update each individual element of the matrix $\boldsymbol{C}$, we must first calculate the conditional prior probability that a couple (a,b) is a match, given all the other elements of the matrix c. We will indicate with the symbol $\boldsymbol{C}^{-ab}$ the matrix $\boldsymbol{c}$, without the element c_{ab}. Of course, $Pr\left(C_{ab}=1\,|\,\boldsymbol{C}^{-ab}\right)=0$ if a match is present in row a or in column b, i.e. if $\sum_{b'\neq b}c_{ab'}=1$ or $\sum_{a'\neq a}c_{a'b}=1$.

Let $t^{(-ab)}$ be the number of matches of the matrix $\boldsymbol{C}^{-ab}$; when $t^{(-ab)}=t-1$ and $\sum_{a'\neq a}c_{a'b}=0$, $\sum_{b'\neq b}c_{ab'}=0$ it can be shown that

$$P(C_{ab}=1\,|\,c^{-ab})=\left[1+\frac{1-pT_q}{pT_q}(T_q-t+1)\right]^{-1}$$

The above formula enables an easy calculation of the full conditional posterior distribution for each single $\{c_{ab}\}$. Indeed,

$$C^{ab}\,|\ldots\sim Bernoulli(P(C_{ab}=1\,|\,y,c^{-ab},m,u))$$

where $P(c_{ab}=1\,|\,y,\boldsymbol{C}^{-ab},m,u)$ can be written as

$$=\frac{P(y_{ab}\,|\,c_{ab}=1)P(c_{ab}=1\,|\,\boldsymbol{C}^{-ab})}{P(y_{ab}\,|\,c_{ab}=1)P(c_{ab}=1\,|\,\boldsymbol{C}^{-ab})+P(y_{ab}\,|\,c_{ab}=0)P(c_{ab}=0\,|\,\boldsymbol{C}^{-ab})}$$

$$=\frac{\prod_{\boldsymbol{i}=1}^{2^k}m_{\boldsymbol{i}}^{d(y_{ab},y_{\boldsymbol{i}})}P(c_{ab}=1\,|\,\boldsymbol{C}^{-ab})}{\prod_{\boldsymbol{i}=1}^{2^k}m_{\boldsymbol{i}}^{d(y_{ab},y_{\boldsymbol{i}})}P(c_{ab}=1\,|\,\boldsymbol{C}^{-ab})+\prod_{\boldsymbol{i}=1}^{2^k}u_{\boldsymbol{i}}^{d(y_{ab},y_{\boldsymbol{i}})}P(c_{ab}=0\,|\,\boldsymbol{C}^{-ab})}.$$

The full conditional distributions of the other parameters can be obtained through lengthy but simple calculations. For $\boldsymbol{m}$ and $\boldsymbol{u}$, there are

$$m\,|\ldots\sim \mathsf{D}(\alpha_1+\sum_{ab}d(y_{ab},y_1)c_{ab},\ldots,\alpha_{2^k}+\sum_{ab}d(y_{ab},y_{2^k})c_{ab})$$

and

$$u \mid \ldots, \sim \mathrm{D}(\beta_1 + \sum_{ab} d(y_{ab}, y_1)(1 - c_{ab}), \ldots, \beta_{2^k} + \sum_{ab} d(y_{ab}, y_{2^k})(1 - c_{ab})),$$

whereas the hyperparameter ξ has a beta conditional distribution

$$\xi \mid \ldots \sim \mathrm{B}(\delta_1 + t, \delta_2 + T_m - t).$$

For all these variables, a Gibbs sampling step can be used.

1.4.2 POINT ESTIMATES FOR *C* AND THE FALSE MATCH RATE

The usual output of a MCMC-based analysis is a sample of approximately independent "observations", simulated from the posterior distribution; this sample can be used to obtain a representation of the uncertainty concerning the parameter of main interest, essentially the matrix $\boldsymbol{C}$. Also, rather often, record linkage procedures are only the first step of a more complex statistical analysis: indeed, record linkage is a crucial step in creating a suitable dataset to be used subsequently. In terms of statistical theory, this is equivalent to producing a point estimate of $\boldsymbol{C}$, from which our "declared" matches can be selected. Classical inference methods can only provide plug-in estimates, based on a theory developed by Fellegi and Sunter (1969) and Jaro (1989).

First, the functions $m(\cdot)$ and $u(\cdot)$ are estimated; then, a sequence of statistical tests is performed, to decide whether each pair $(a,b) \in \boldsymbol{A} \times \boldsymbol{B}$ can be declared a match or a non-match. Usually, these tests are calibrated to obtain a specific level of False Match Rate (FMR), i.e. the ratio between the number of false matches and the total number of declared matches. Note that the FMR is exactly equivalent to the well-known False Discovery Rate, which is very popular in multiple comparison applications (wavelets theory, micro-array analysis, etc.) Also, according to record linkage procedures currently in use must complete the statistical data analysis with a reallocation procedure, that eliminates inconsistencies between the results of different tests (see Jaro (1989)) and the problem posed by Larsen (1999, para. 3.3): in a Bayesian framework, these problems are automatically solved.

Indeed, the Bayesian method for dealing with record linkage problems is completely different, and raises interesting issues from both the practical and the methodological points of view. Although in a formal Bayesian analysis the point estimate selected should be that which minimizes the posterior expected loss, in practical applications, it is common practice to use the posterior mean, or sometimes the posterior median. Clearly, these "shortcut" solutions do not appear reasonable in a record linkage context: the marginal posterior mean of each individual element of the matrix $\boldsymbol{C}$ will be a number between 0 and 1, which is not of much assistance in deciding whether the pair (a,b) is a match. In multivariate discrete settings problems, the use of the posterior median is even more complicated.

Therefore, adopting a decision-theoretic approach appears to be necessary: let $g = \{g_{ab}\} \in G, a = 1, \cdots, \nu_A$ and $b = 1, \cdots, \nu_B$, a generic matrix of size $\nu_A \times \nu_B$, with the same characteristics as $\boldsymbol{C}$, such that it represents our "action"; here, G represents the set of all possible actions. Also, let $L(\cdot,\cdot)$ a loss function defined as $L : G \times C \to \mathsf{R}^+$.

Our goal is to select, for a given loss L, the optimal decision $\boldsymbol{g}^*$: that which minimizes the posterior expected loss

$$\mathbf{g}^* = argmin_{\mathbf{g} \in G} W(\mathbf{g})$$

where

$$W(\boldsymbol{g}) = \mathrm{I} \quad E^{\pi(\boldsymbol{C}|y)} L(\boldsymbol{C}, \boldsymbol{g}).$$

Below, we will consider certain loss functions:

- **Quadratic Loss**

$$L_q(\boldsymbol{C}, \boldsymbol{g}) = \sum_a \sum_b (c_{ab} - g_{ab})^2.$$

Since the elements of $\boldsymbol{C}$ and $\boldsymbol{g}$ are either 0 or 1, the loss L_q is equivalent to the loss L_1 (absolute value): $L_1(\boldsymbol{C}, \boldsymbol{g}) = \sum_a \sum_b | c_{ab} - g_{ab} |$.

- **False Match Rate**

$$L_{FMR}(\mathbf{C}, \mathbf{g}) = \begin{cases} 0 & \text{if } \sum_a \sum_b g_{ab} = 0 \\ \dfrac{\sum_a \sum_b g_{ab} I_{c_{ab}=0}(c_{ab})}{\sum_a \sum_b g_{ab}} & \text{otherwise} \end{cases}.$$

L_{FMR} is the loss that translates the classical use of the False Match Rate into Bayesian terms, to measure the performance of the record linkage analysis.

- **Absolute number of errors**

$$L_{Abs}(\boldsymbol{C}, \boldsymbol{g}) = \sum_a \sum_b \left[g_{ab} I_{c_{ab}=0}(c_{ab}) + (1 - g_{ab}) I_{c_{ab}=1}(c_{ab}) \right]$$

The following theorem provides the optimal solution for the losses mentioned above.

Theorem 1

Under the losses L_q and L_{Abs}, the optimal Bayesian solution is given by the matrix $\boldsymbol{G}^*$, whose generic element is defined as

$$g_{ab}^* = \begin{cases} 1 & \text{if } Pr(c_{ab} = 1 \mid \mathbf{y}) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}.$$

The optimal solution under the loss L_{FMR} is given by the matrix $\boldsymbol{G}^0$, consisting of all zeroes.

Proof:

A) since

$$L_q(\boldsymbol{C}, \boldsymbol{g}) = \sum_a \sum_b [c_{ab} + g_{ab} - 2c_{ab}g_{ab}],$$

the problem is equivalent to the maximization of the posterior expected value of

$$L_q(\boldsymbol{C}, \boldsymbol{g}) = 2\sum_a \sum_b g_{ab}\left[c_{ab} - \frac{1}{2}\right].$$

Also, when considering the loss L_{Abs}, simple calculations lead to

$$L_{Abs}(\boldsymbol{C}, \boldsymbol{g}) = \sum_a \sum_b \left[g_{ab}\left(1 - I_{c_{ab}=1}(c_{ab})\right) + \left(1 - g_{ab}\right) I_{c_{ab}=1}(c_{ab})\right],$$

$$= \sum_a \sum_b \left[g_{ab} - 2g_{ab} I_{c_{ab}=1}(c_{ab}) + I_{c_{ab}=1}(c_{ab})\right]$$

The minimization of the posterior expected loss of L_{Abs} is equivalent to the maximization of the quantity

$$L_q(\boldsymbol{C}, \boldsymbol{g}) = 2\sum_a \sum_b g_{ab}\left[I_{c_{ab}=1}(c_{ab}) - \frac{1}{2}\right].$$

Also, note that c_{ab} and $I_{c_{ab}=1}(c_{ab})$ are two different ways of denoting the same random variable; thus, the quantities L_q e L_{Abs} are identical and finding the optimal solution for L_q is sufficient. We must then maximize

$$W_q(\boldsymbol{g}) = 2\mathrm{I} \quad E^{\pi(\boldsymbol{C}|by)} \sum_a \sum_b g_{ab} \left[c_{ab} - \frac{1}{2} \right] = 2 \sum_a \sum_b g_{ab} \left[Pr(c_{ab} = 1 \mid \boldsymbol{y}) - \frac{1}{2} \right].$$

The last expression shows that the value that maximizes $W_q(\boldsymbol{g})$ can be obtained by setting $g_{ab} = 1$ if and only if the correspondent coefficient is positive, i.e. when $Pr(c_{ab} = 1 \mid \boldsymbol{y}) > \frac{1}{2}$.

B) When L_{FMR} is used, it is trivial to note that FMR is minimized if the conservative behaviour of not declaring any match is adopted. Indeed, in this case, the posterior expected loss is always zero, regardless of the posterior distribution. The optimal solution is thus given by $\boldsymbol{g}^*_{a,b} = 0$, for all (a,b).

It must be emphasized that all optimal solutions derived through Theorem 1 are based on the marginal posterior probabilities that the various pairs (a,b) are matches. This is a consequence of the fact that the loss functions set out above are additive, and basically "sum" all losses due to individual mismatches.

Part B of Theorem 1 is also important. The Part states that from a decision-theoretic perspective, the FMR is not a valid measure of performance, because it controls only one type of error. Every reasonable loss function should also take into account a measurement of the number of undiscovered matches.

From this perspective, a possible loss function for record linkage is given by the *Global Error Rate*

$$L_{TOT}(\boldsymbol{C}, \boldsymbol{g}) = L_{FMR}(\boldsymbol{C}, \boldsymbol{g}) + \frac{\sum_a \sum_b (1 - g_{ab}) I_{c_{ab}=1}(c_{ab})}{\sum_a \sum_b (1 - g_{ab})}.$$

The loss L_{TOT} is actually capable of capturing errors due to missing true matches. However, the improvement is more theoretical than practical: indeed, in the previous formula, the second factor's denominator is so much greater than the denominator of L_{FMR} that the results obtained with L_{TOT} are not significantly different from those derived under loss L_{FMR}.

1.5 STATISTICAL INFERENCE WITH LINKED DATA: PROBLEMS AND SOLUTIONS

In this Report, we will describe and extend some recent advancements on a general Bayesian methodology for performing record linkages and for making inferences, using the resulting matched units. In particular, we will frame the record linkage process into

a formal statistical model that comprises both the matching variables and the other variables included at the inferential stage. Researchers will therefore be able to account for the uncertainty in the matching process, in the context of inferential procedures based on probabilistically linked data; at the same time, they will also be able to generate a feedback propagation of the information, between the working statistical model and the record linkage stage.

We argue that this feedback effect is both

- essential in eliminating potential biases that would otherwise characterize the resulting linked data inference, and

- capable of improving record linkage performances.

The procedure's practical implementation is based on the use of standard Bayesian computational techniques, such as Markov Chain Monte Carlo algorithms.

Although the methodology is rather general, for convenience of exposition we will focus on the popular and important case of linear multiple regression set-up.

1.5.1 INTRODUCTION

From a methodological perspective, the operation of merging two (or more) datasets can be important for two different and complementary reasons:

- *per se*, to obtain a larger reference dataset or frame, suitable for more accurate statistical analyses;

- to calibrate statistical models through the additional information that cannot be extracted from either one of the individual datasets.

If the merging can be accomplished without errors (perhaps because a clear identification key is available and can be used to match the units in the two datasets), there are no specific consequences for the statistical procedures undertaken in both situations. In practice, however, identification keys are often unavailable and linkages between statistical units are performed with a certain degree of uncertainty. This is when record linkage must be performed, and the possibility of making wrong matching decisions must be accounted for, especially when reporting the subsequent statistical analyses.

In this introductory section, we will first qualitatively define record linkage and identify when it is useful in statistical inference. Then, we will briefly sketch some historical notes on the development of record linkage theory. In the following sections, we will describe and extend the methods for regression analysis developed by Liseo and A. Tancredi (2011b). In the approach proposed, it is not necessary to impose constraints on the matching patterns, as suggested in Lahiri and Larsen's seminal paper (2005): this remarkable feature greatly broadens the method's potential applicability. In the next section, we will briefly recall the Bayesian approach to record linkage proposed by Tancredi and Liseo (2011) and introduce a simplified version of that model. In Section

5.5, we generalize the method to include a regression procedure step, and we compare the present approach to existing methodologies.

To briefly explain what record linkage is, let us suppose that there are two datasets F_1 and F_2, whose records relate, respectively, to statistical units (e.g. individuals, firms, etc.) of partially overlapping samples (or populations) S_1 and S_2. The records in the two datasets consist of several fields, or variables, either quantitative or categorical, that can be observed with a potential dose of noise. For example, in a file of individuals, the fields can be *surname, age, sex*, etc.

The goal of a record linkage procedure is to detect all the pairs of units (j, j'), with j in S_1 and j' in S_2, such that j and j' actually refer to the same unit, through the information provided by the records observed in the two datasets.

If the main goal of record linkage is that outlined above (case i), a new dataset is created by merging three different subsets of units: those present in both datasets, those belonging to S_1 only and those belonging to S_2 only. Naturally, information regarding the first group of individuals will be richer.

Appropriate statistical data analyses can be then performed on this expanded dataset: however, since the linkage step is executed with uncertainty, it may suffer the presence of duplicate units and a certain lack of efficiency, essentially due to erroneous matching in the merging process.

On the other hand, situation ii) – more important for the scope of this Report and discussed throughout this paper – is even more challenging, from both practical and methodological perspectives.

Let us assume that the observed variables in F_1 are denoted by $(Y, V_1, V_2, \ldots, V_h)$ while the observed variables in F_2 are $(X, V_1, V_2, \ldots, V_h)$. One might then be interested in performing a linear regression analysis (or any other more complex association model) between Y and X, restricted to those pairs of records that are declared matches pursuant to a record linkage analysis based on variables $(V_1, \ldots, V_h)$. The intrinsic difficulties of such a simple problem are well-documented in Neter *et al.* (1965) and discussed in Scheuren and W. E. Winkler (1993), Scheuren and W. E. Winkler (1997) and Lahiri and Larsen (2005). However, it is easy to note that in the regression case, the presence of false matches (matching record pairs that do not refer to the same statistical unit) reduces the level of association observed between Y and X; as a consequence, it introduces a bias effect towards zero in estimating the slope of the regression line. Similar biases arise in any statistical procedure, and often take a specific direction. As another example, when linkage procedures are used to estimate the size N of a population through a capture-recapture approach, the presence of false matches may severely reduce the final estimate of N.

At this point, it should also be noted that in practical uses of record linkage, it is quite usual for the linker (the researcher who matches the two files) and the analyst (who performs the statistical analysis) to be two different persons, working separately. However, as Scheuren and Winkler (1993) state, *"... it is important to conceptualize the linkage and analysis steps as part of a single statistical system and to devise appropriate strategies accordingly"*. Placing this statement in a broader perspective, let us suppose that we wish to observe variables $(Y_1, Y_2, \ldots, Y_k, V_1, V_2, .., V_h)$ on n_1 units in file F_1 and variables $(X_1, X_2, \ldots, X_p, V_1, V_2, \ldots, V_h)$ on n_2 units in file F_2. In this scenario, we consider the twofold objective of first, using the key variables $V_1, V_2, \ldots, V_h$ to draw inferences concerning the true matches between sources F_1 and F_2; and, at the same time, of using a statistical model M to perform an analysis based on variables Y's and X's (or even including V's), restricted to the records that have been recognized as matches. To perform this double task, we argue that a fully Bayesian analysis is the simplest method to obtain an integrated use of the information, that

- improves the execution of the linkage (through the use of the additional information contained in the Y's and X's. This occurs because pairs of records that do not adequately fit model M will be automatically down-weighted in the matching estimation;
- enables accounting for matching uncertainty in the estimation procedure related to model M, involving Y's and X's, in a natural manner; and
- improves the accuracy of the estimators of model M's parameters, at least in terms of bias.

An early attempt to frame the statistical problem of record linkage from a Bayesian perspective can be found in Fortini *et al*. (1990): the likelihood function provided by the set of multiple comparisons of different records in the two datasets – comparisons that could involve several different variables – was used to estimate the matching configuration, through use of a specific Markov Chain Monte Carlo technique. This approach, together with that formulated by Lahiri and Larsen (2005), can be interpreted as a Bayesian alternative to the classic record linkage approach formalized by Jaro (1989), which follows Fellegi and Sunter's seminal paper (1969). Recently Tancredi and Liseo (2011) have proposed a different Bayesian matching procedure that is particularly suited for categorical variables. They explicitly model the records fully observed through a particular measurement error model, inspired by the "hit-and-miss" strategy proposed by Copas and Hilton (1990). In the same paper, the problem of uncertainty in population size estimation, based on capture-recapture models with linkage uncertainty, is discussed in detail. Liseo and A. Tancredi (2011) have also introduced a record linkage model for continuous data, based on a multivariate normal model with measurement error.

In recent years, several authors have considered the problem of estimating the parameters of a regression model using linked data. Extending the pioneering works of Scheuren and W. E. Winkler (1993), (1997), and assuming that the two data sets represent a permutation of the same list of units, Lahiri and Larsen (2005) have proposed an estimator (LL) of the regression coefficients that is unbiased conditionally upon the matching probabilities provided by the record linkage process. Their approach was

extended by Hof and Zwinderman (2012) to more complex and realistic linkage scenarios and logistic regression problems. Generalizations of the LL estimator were also provided by Kim and Chambers (2012), who adopt an approach based on estimating equations. Goldstein *et al*. (2012) prposed a different approach: the authors consider the probabilities of being a match, provided by the record linkage algorithms, as an ingredient to be used within a multiple imputation scenario. Finally, Harvey *et al*. (2012) have proposed a Bayesian procedure that jointly models the record linkage and association between variables in two different dataset files.

In this paper, the authors consider the (computationally) simpler situation in which the number of records to be matched in the two datasets is low; this is obtained after an ample and informative blocking step. They show how the joint model improves both the matching procedure and the accuracy of the estimation of regression parameters, in a real data example concerning the "end-of life cost" data. Another limitation in Harvey *et al*. (2012) is that the authors assume a specific matching pattern; actually, for each individual block of comparisons, all cases in the smaller list also feature in the other list.

In the sections below, we sketch the model's mathematical framework. First, we discuss in detail the Bayesian model for record linkage. We then show how to link it with the inference model M.

1.5.2 A BRIEF REVIEW OF RECORD LINKAGE METHODOLOGY

Suppose there are two matrices of record, say V_1 and V_2, of different sizes n_1 and n_2 respectively. Here

$$V_1 = (v_{11}, \ldots v_{1n_1}) \qquad \text{and} \qquad V_2 = (V_{21}, \ldots, v_{2n_2})$$

and each individual v_{ij} can be represented as $v_{ij} = (v_{ij1}, \ldots, v_{ijh})$, i.e. it contains the observed values of a categorical random vector $v = (v_1, \ldots, v_h)$ whose support is $\mathsf{V} = \{v_{j_1 j_2, \ldots, j_h} = (j_1, \ldots, j_h) \quad j_1 = 1 \ldots, k_1; \ldots; j_h = 1, \ldots k_h\}$.

Also, consider the sets M and U of "true matches" and "true non-matches" respectively, introduced above. More precisely,

$$M = \{(j, j') : \text{record } j \in V_1 \text{ and } j' \in V_2 \text{ refer to the same unit}\},$$

and, of course, $U = M^c$.

The main goal of any record linkage technique is to identify which pair of records should be assigned to M.

The statistical model for a record linkage analysis is built upon the "comparison vectors" $q_{jj'} = (q_{jj'1}, \cdots, q_{jj'h})$, where

$$q_{jj'l} = \begin{cases} 1 & v_{1jl} = v_{2j'l} \\ 0 & v_{1jl} \neq v_{2j'l} \end{cases}, \qquad l = 1, \ldots, h.$$

Comparison vectors $q_{jj'}$ are assumed to be independent[4] and identically-distributed random vectors, with a distribution given by the mixture

$$p(q_{jj'} \mid m, u, w) = w \prod_{l=1}^{h} m_l^{q_{jj'l}} (1 - m_l)^{1 - q_{jj'l}} + (1 - w) \prod_{l=1}^{h} u_l^{q_{jj'l}} (1 - u_l)^{1 - q_{jj'l}}. \quad (1.7)$$

In the formula above, w represents the marginal probability that a random pair of records belongs to the same unit. In other words, w represents the percentage of overlapping between the two datasets. The quantities m_l and u_l, $l = 1, \ldots, h$ are the parameters of the two multinomial distributions related to the two sets of comparison M and U, i.e.

$$m_l = Pr(\text{ key variable } X_l \text{ agree} \mid \text{pair} \in M)$$

Also, individual key variables are assumed to be independent. To perform a record linkage procedure, one can consider the likelihood ratio

$$\lambda = \frac{P(q_{jj'} \mid (j, j') \in M)}{P(q_{jj'} \mid (j, j') \in U)} = \frac{\prod_{l=1}^{h} m_l^{q_{jj'l}} (1 - m_l)^{1 - q_{jj'l}}}{\prod_{l=1}^{h} u_l^{q_{jj'l}} (1 - u_l)^{1 - q_{jj'l}}}.$$

Otherwise, one may consider the posterior probability that an individual pair is a match $p((j, j') \in M \mid q_{jj'})$. Generally, pairs with a likelihood ratio λ – or a posterior probability – over an established threshold are declared matches. On the other hand, choosing the threshold can be problematic. In this case, optimization techniques may be helpful in ruling out the multiple matches issue, i.e. the possibility that an individual unit in dataset A is declared to match with two different units in dataset B.

Several extensions of this basic structure have been proposed. In particular, Larsen and Rubin (2001) introduced potential interactions between key variables; see also Winkler (1994) and references therein.

[4] Strictly speaking, this assumption is untenable: after comparing record A_1 with records B_1 and B_2, and then record A_2 with B_1, we are able to predict the result of comparison between A_2 and B_2.

1.5.3 BAYESIAN RECORD LINKAGE

We model the observed data matrices V_1 and V_2 of key variables, taking into account both potential measurement errors and relevant matching constraints. Let $\tilde{v}_{ijl}$ be the true unobserved value for the field l of record j on data set V_i. We assume that

$$v_{ijl} \mid \tilde{v}_{ijl} : \gamma_l \delta_{\tilde{v}_{ijl}}(v_{ijl}) + (1-\gamma_l) p(v_{ijl}) \quad \forall ijl.$$

Notice that v_{ijl} is a mixture of two components, the former degenerate on the true value while the latter is distributed over the other possible values of the variable. γ_l is the probability that the variable V_l is observed without noise; for each l, γ_l is assumed to be the same throughout the different datasets. This model, known as "hit-and-miss", was introduced by Copas and Hilton (1990) and recently adapted in the Bayesian framework by Tancredi and Liseo (2011) and Hall *et al*. (2013).

To build a model for true values $\tilde{v}_{ijl}s$, we introduce the matching matrix C. In particular, let C be a $n_1 \times n_2$ unknown matrix whose entries are either 0 or 1; $C_{jj'} = 1$ represents a match, and $C_{jj'} = 0$ denotes a non-match. We assume that each dataset does not contain replications of the same unit, so that $\sum_{j'} C_{jj'} \leq 1$, and $\sum_{j} C_{jj'} \leq 1$. Harvey *et al*. (1990) have used a similar matching matrix in a slightly different context, the alignment of unlabelled points for reconstructing molecular shapes. We assume that the true values distribution depends on the entries for C. More precisely, we assume that

$$p(\tilde{V}_1, \tilde{V}_2 \mid C) = \prod_{j: C_{jj'}=0 \forall j'} p(\tilde{v}_{1j}) \prod_{j': C_{jj'}=0 \forall j} p(\tilde{v}_{2j'}) \prod_{jj': C_{jj'}=1} p(\tilde{v}_{1j}, \tilde{v}_{2j'}). \quad (1.8)$$

Moreover, we take an independent multinomial sampling for the "non-match" true values, i.e.

$$p(\tilde{v}_{ij} = (j_1, \ldots, j_h)) = \prod_{l=1}^{h} \theta_{l j_l}$$

and a degenerate joint distribution for the true match values

$$p(\tilde{v}_{1j}, \tilde{v}_{2j'}) = \begin{cases} 0 & \text{if } \tilde{v}_{1j} \neq \tilde{v}_{2j'} \\ \prod_{l=1}^{h} \theta_{l j_l} & \text{otherwise} \end{cases}.$$

Note that the model above represents a simplified version of the model proposed in Tancredi and Liseo (2011), where an additional layer – when formulating a super-population model – was added at the apex of the hierarchy. This simplified model, also

used in Hall *et al*. (2013), can be obtained by integrating out with respect to the additional layer of hierarchy, under specific prior assumptions.

With respect to the prior distribution for C, this can be given in two steps. The first consists of a prior distribution $\pi(h)$, $h = 0,1,2,\ldots n_1 \wedge n_2$ on the random variable H: "number of matched pairs in the two data sets". In this step, researchers can easily collect information by studying previous experiences or the statistical characteristics of the datasets (e.g. if the two datasets refer to a census and a sample, respectively, a great number of matched pairs can be expected).

The second step consists of a conditional distribution of the configuration matrix C, given the number of matches. We take the natural noninformative choice of a uniform conditional prior on the set

$$C^{(h)} = \{C : \sum_{j}\sum_{j'} c_{j,j'} = h\},$$

that is $\pi(C \mid H = h) = \left(\text{card}(C^{(h)}\right)^{-1}$.

The model outlined cannot be analyzed in closed form; some form of simulation from the posterior distribution is required. In particular, we used a Metropolis-Hastings algorithm, which we will now describe. The C matrix is updated by adding, deleting or switching matches. For example, when proposing a move from $C_{jj'} = 0$ to $C_{jj'} = 1$, we accept the move with a probability given by

$$1 \quad \wedge \quad \frac{q(C \mid C')}{q(C' \mid C)} \frac{p(V_1, V_2 \mid C', \theta, \gamma)}{p(V_1, V_2 \mid C.\theta, \gamma)} \frac{p(C')}{p(C)}$$

where

$$\frac{p(V_1, V_2 \mid C', \theta, \gamma)}{p(V_1, V_2 \mid C.\theta, \gamma)} = \frac{p(v_j, v_{j'} \mid \theta, \gamma)}{p(v_j \mid \theta, \gamma) p(v_{j'} \mid \theta, \gamma)}, \tag{1.9}$$

$q(C \mid C')$ is the proposal density of the Metropolis-Hastings algorithm and

$$p(C) = \sum_{h=0}^{n1 \wedge n_2} \pi(C \mid h)\pi(h).$$

If the move is accepted, we propose new true values $(\tilde{v}_j, \tilde{v}_{j'})$ by sampling from their full conditional distributions, given the new status $C_{jj'} = 1$. It must be noted that the ratio (9) appearing in the above acceptance probability represents the Bayes factor, that

compares the hypothesis that the pair (j, j') is a match to the hypothesis that it is not a match: see e.g. Lindley (1977) and Liseo and Tancredi (2011) for similar expressions, when Gaussian distributions are adopted.

After a reasonably large sample is drawn from the posterior distribution, we propose estimating the matching configuration through the following – rather natural – point estimate of the matrix C, namely

$$\hat{C}_{ij} = \begin{cases} 1 & \text{if } p(C_{ij} = 1 \mid V_1, V_2) \geq \dfrac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Some observations must be made. First, the estimator $\hat{C}$ is to a certain extent suggested by simple decision-theoretic considerations (see Section 4.2). In this situation, the posterior mean cannot be used, since it would only provide useless real numbers between 0 and 1. Second, the estimated matrix $\hat{C}$ should only be used when the linkage procedure is the ultimate goal of the statistical analysis and a set of potential matches must be declared. If the merged data set is, on the other hand, the starting point of a new statistical analysis, the uncertainty concerning C provided by the posterior distribution of the matrix itself should be accounted for. We illustrate this below, in the particular case of the linear multiple regression model.

1.5.4 BAYESIAN REGRESSION WITH LINKED DATA

Suppose the first dataset is a $n_1 \times (h+1)$ matrix consisting of the variables (y, V_1), while the other dataset is a $n_2 \times (h+p)$ matrix, consisting of variables $(V_2, \boldsymbol{X})$ where $\boldsymbol{X} = (X_1, \ldots X_p)$. Let $\widetilde{\boldsymbol{X}}$ be the matrix containing the true (unobserved) covariate values for the corresponding entries of Y. $\widetilde{\boldsymbol{X}}$ has dimensions $n_1 \times p$. Conditionally on $\widetilde{\boldsymbol{X}}$ and on the true matching variables $\widetilde{V}_1$ and $\widetilde{V}_2$, we assume a Gaussian linear regression model for Y:

$$Y \mid \widetilde{\mathbf{X}}, \boldsymbol{\beta}, \sigma^2 \sim N(\widetilde{\mathbf{X}}\boldsymbol{\beta}, \sigma^2 I).$$

In addition, given the matrices of true values $\widetilde{V}_1$ and $\widetilde{V}_2$, we assume, for V_1 and V_2, the "hit-and-miss" model illustrated previously.

Conditionally on the matching matrix C, we also assume that the covariates for y_j are given by the vector $x_{j'}$ (the x-part of the j'-th row of data set B) only when $C_{jj'} = 1$; otherwise, we assume that they are unknown with a specific distribution $p(\tilde{x})$. The choice of $p(\cdot)$ is not crucial; generally, a multivariate Gaussian distribution will be used. More precisely, we have

$$p(\widetilde{\boldsymbol{X}} \mid C) = \prod_{jj':C_{jj'}=1} \delta_{x_{j'}}(\widetilde{x}_j) \prod_{j:C_{jj'}=0 \forall j'} p(\widetilde{x}_j)$$

For the matrices of true values $\widetilde{V}_1$ and $\widetilde{V}_2$, we will adopt the same model expressed by (8).

Note that the covariates vector for non-matches pairs are handled as missing variables.

The posterior simulation can be conducted easily through a standard Metropolis-Hastings algorithm. In this case, an "add-one-match" move will be accepted, with a probability depending on

$$\frac{p(y,V_1,V_2 \mid C',\theta,\gamma)}{p(y,V_1,V_2 \mid C,\theta,\gamma)} = \frac{\phi(y_j;x_{j'}^T\beta,\sigma^2)}{\int \phi(y_j;\widetilde{x}^T\beta,\sigma^2)p(\widetilde{x})d\widetilde{x}} \frac{p(v_j,v_{j'} \mid \theta,\gamma)}{p(v_j \mid \theta,\gamma)p(v_{j'} \mid \theta,\gamma)}. \quad (1.10)$$

These formulas prompt several important comments.

1. There is a feedback effect between the parameters of the model's linkage block and those of the regression part. This occurs because all of these appear in the acceptance probability. This implies, for example, that the posterior distribution of the $\boldsymbol{\beta}$ vector will not be independent of C. This must be interpreted as a bias-correction effect.
2. A closed-form expression for

   $$p(y_j;\boldsymbol{\beta},\sigma^2) = \int \phi(y_j;\widetilde{x}^T\boldsymbol{\beta},\sigma^2)p(\widetilde{x})d\widetilde{x}$$

   can be obtained, for example, by assuming a multivariate normal for $p(\widetilde{x})$.
3. When the "add-one-match" move is accepted, we update the true values $(\widetilde{x}_j,\widetilde{v}_j,\widetilde{v}_{j'})$ by drawing from their full conditional distributions conditionally, on the new status $C_{jj'}=1$.

1.5.5 CONNECTION WITH THE LL ESTIMATOR

Lahiri and Larsen (2005) consider a special case of regression with linked data, in which the two datasets consist of the same n units, that is $n = n_1 = n_2$, and, in our notation, $\sum_{j'} C_{jj'} = 1,$ for all j. Let z_j be the response value associated to covariates $\boldsymbol{x}_j$, after a record linkage procedure has been performed. Essentially, Lahiri and Larsen (2005) assume that

$$z_j \sim \sum_{k=1}^{n} p_{jk}\phi(\mathbf{x}_k^T\boldsymbol{\beta},\sigma^2) \quad i = 1,\ldots,n$$

where $p_j = (p_{j1}, \ldots, p_{jn})'$ is a vector of known matching weights – from the linkage procedure – such that $\sum_k p_{jk} = 1$. They notice that the expected value of z_j is given by

2

Application

2.1 DATA DESCRIPTION

In this Chapter, two main data sources will be considered. The first dataset is the 2007 Mozambique Census of Population and Housing (MCPH), limited to the Guija, Mandlacaze and Mabalane districts of Gaza Province. From this dataset, we extracted all households engaging in agricultural activities that satisfed the criteria for inclusion in the subsequent 2009-2011 Census of Agriculture and Livestock, as established in the relevant Statistics Mozambique methodology report (2012). We thus obtained a file with 54,007 records, from which we considered the following list of observed variables:

S_1:	District	V_1:	Number of cashew trees	V_5:	Number of sheeps
S_2:	Administrative post	V_2:	Number of coconut trees	V_6:	Number of pigs
S_3:	Locality	V_3:	Number of cows	V_7:	Number of chickens
S_4:	Enumeration area	V_4:	Number of goats	V_8:	Number of ducks

In Mozambique, each province is divided into districts, admistrative posts and localities. For the census operations, each locality was also divided into enumeration areas. The Gaza province households engaging in agricultural or livestock operations for the 2007 MCPH dataset are displaced across 730 different enumeration areas. The top panel of Figure 2.1 below shows the number of these households for each enumeration area. The average number of households per enumeration area is 74 (s.d. 33), with a minimum of 10 and a maximum of 291.

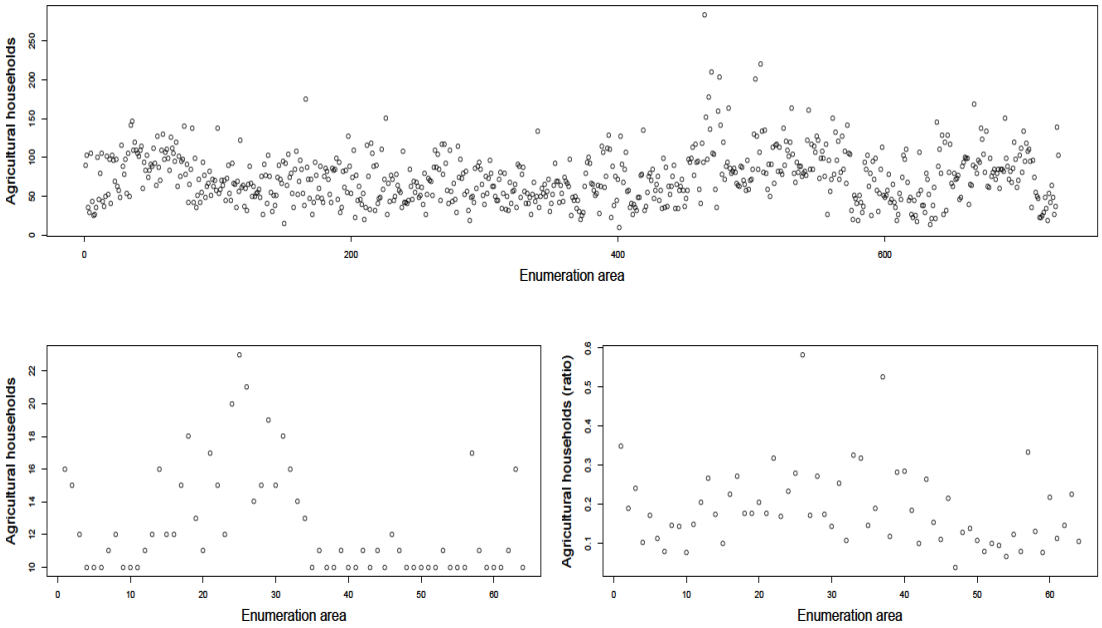

Figure 2.1: Number of households in each enumeration area. Top panel: MCPH dataset. Bottom-left panel: CAPII dataset. Bottom-right panel: ratios between the number of households in each enumeration area common to the two datasets.

The other data source is the 2009 Mozambique Census of Agriculture and Livestock (CAP II). This is a stratified two-stage sampling survey, with the first sampling level given by the enumeration areas of the 2007 MCPH dataset, and the second level given by the agricultural households. For this dataset too, we consider only the Guija, Mandlacaze and Mabalane districts of the Gaza province. Of the 730 enumeration areas of the 2007 MCPH dataset, 64 were observed in this survey too, corresponding to a percentage of 8.7%. In each enumeration area, all agricultural households that could be assimilated to medium farmholders were included in the sample, in addition to a sample of about 10 family farms. The final file comprises 801 records, and reports several variables, including the list of observed variables of the MCPH dataset outlined above. The bottom-left panel of Figure 2.1 above shows the number of records for each enumeration area observed in the CAP II dataset. The average number of records per enumeration area is 12.5 (s.d=3.2), with a minimum of 10 and a maximum of 23. In principle, each sampled unit in the CAP II dataset should also be present in the corresponding enumeration area of the MCPH dataset. If this were the case, the ratio between the number of CAPII records and the number of MCPH records could be interpreted as a nominal coverage level of the CAPII survey. The bottom-right panel of Figure 2.1 shows these ratios for each enumeration area. It must be noted that the average coverage per enumeration area is of 16%, with a minimum of 3.8% and a maximum of 53.8%.

Since the two datasets do not share a common identifier, it is not possible to

deterministically link the households of the two surveys. However, it could be practicable to perform a probabilistic record linkage, considering the enumeration area as a blocking variable and the variables measuring the agricultural and livestock operations as key variables. In record linkage literature, blocking strategies are very popular; basically, these consist of partitioning, into homogeneous groups, all possible comparisons among records, to reduce the computational burden (see for example Newcombe (1967) or Winkler (2004)). To illustrate the blocking operation, we consider all records of a specific enumeration area in the district of Mandlacaze, observed in the MPCH dataset (see Table 2.1 below) and in the CAPII dataset (Table 2.2 below). Since the ten records in Table 2.2 may be co-referential only with ten of the nineteen units observed in Table 2.1, it is natural to limit record comparison only to the $190 = 10 \times 19$ pairs obtained by matching the units of Table 2.2 only with the units of Table 2.1, without extending the comparison to the MPCH units recorded in the other enumeration areas. This, essentially, is the blocking concept.

	CASHEWS	COCONUTS	COWS	GOATS	SHEEPS	PIGS	CHICKENS	DUCKS
13135	10	0	0	1	0	0	5	1
13136	15	0	0	2	0	1	4	0
13137	12	0	0	0	0	1	3	0
13138	15	0	0	0	0	1	5	0
13139	20	1	0	4	0	7	5	0
13140	0	0	0	0	0	0	0	0
13141	20	2	0	1	0	0	5	0
13142	20	3	0	1	0	1	5	2
13143	30	0	0	3	0	1	3	0
13144	20	0	0	1	0	10	10	0
13145	15	0	0	5	0	3	8	5
13146	7	0	0	0	0	0	0	0
13147	1	1	0	2	0	0	5	0
13148	0	1	0	0	0	0	7	0
13149	4	0	0	0	0	0	4	0
45391	30	2	2	7	0	0	10	2
47072	10	0	3	0	0	1	0	0
48322	11	0	4	10	0	0	6	0
50154	40	0	6	0	0	1	10	0

TABLE 2.1. MCPH dataset: values of the key variables observed in Enumeration Area 12, Locality 3, Administrative Post 1 of Mandlacaze District

	CASHEWS	COCONUTS	COWS	GOATS	SHEEPS	PIGS	CHICKENS	DUCKS
500	1	1	0	0	0	1	1	0
501	15	2	0	0	0	2	16	1
502	7	0	0	3	0	0	7	0
503	8	1	0	0	0	0	0	0
504	20	0	3	2	0	0	2	0
505	10	0	0	5	0	0	25	0
506	24	1	2	0	0	4	12	0
507	4	0	0	0	0	0	6	0
508	17	2	0	0	0	0	6	0
509	0	0	0	0	0	0	0	0

TABLE 2.2. CAPII dataset: values of the key variables observed in Enumeration Area 12, Locality 3, Administrative Post 1 of Mandlacaze District

It must also be noted that the records observed for the same family farm during the two different sampling occasions may differ, both due to an actual change in certain variables observed and to measurement errors. Real changes may occur for various reasons. Indeed, the second dataset was recorded two years after the first; in the interval between recordings, cattle may have been sold, new animals may have been bought or were born, or new trees may have been planted. As for measurement errors, the records' actual values may have been modified by transcription mistakes or false declarations. However, although changes may have affected the two sets of observations relating to the same family farm, we expect to find more similar profiles between these records, than for the records referring to two different units. A very approximate way to identify the pairs with similar profiles, with respect to agricultural and livestock operations, could be to measure the distance between each pair of units, for example by the following standardized version of the Euclidean distance:

$$dist(a,b) = \sum_{j=1}^{d} \sqrt{\frac{(x_{a,j} - x_{b,j})^2}{Var(X_j)}} \qquad (2.1)$$

where $x_{a,j}$ is the value of the $j-th$ key variable observed on unit a of the first file, x_{bj} is the value of the same key variable observed on unit b of the other file, and $Var(X_j)$ is the variance of the $j-th$ variable estimated, for example, considering all the units of the greater dataset. In Table 2.3 below, we report the distance (2.1) for each pair of records reported in Tables 2.1 and 2.2.

	1	2	3	4	5	6	7	8	9	10
1	1.54	1.80	0.95	0.90	1.33	1.90	3.59	0.81	0.88	1.03
2	0.75	1.85	0.97	1.18	1.07	1.91	2.68	1.16	1.04	1.31
3	0.32	1.62	1.28	0.75	1.43	2.23	2.44	0.80	0.80	0.88
4	0.45	1.48	1.26	0.88	1.44	2.21	2.31	0.79	0.67	1.01
5	4.38	4.34	4.25	4.81	4.56	4.86	2.85	4.77	4.51	4.99
6	0.62	2.50	0.90	0.19	1.17	1.96	3.32	0.30	0.62	0.00
7	1.25	2.22	0.72	0.62	0.70	1.67	2.91	0.58	0.26	0.80
8	1.71	1.72	2.25	2.15	2.23	3.19	3.37	2.10	1.79	2.33
9	1.17	2.35	1.13	1.60	1.30	2.08	2.82	1.65	1.46	1.73
10	5.69	5.28	6.01	6.12	6.15	6.74	3.78	5.94	5.75	6.25
11	4.88	3.65	4.54	5.31	5.20	4.74	4.38	5.13	5.02	5.44
12	0.72	2.36	0.76	0.05	1.03	1.83	3.18	0.28	0.48	0.14
13	1.01	2.59	0.39	0.66	0.88	1.45	3.42	0.46	0.71	0.57
14	0.78	2.21	0.67	0.42	1.31	1.73	3.03	0.14	0.40	0.29
15	0.73	2.27	0.67	0.26	1.09	1.74	3.09	0.07	0.39	0.23
16	3.84	3.48	2.50	3.21	2.48	2.56	4.48	3.10	2.78	3.40
17	0.61	2.14	1.72	0.97	1.14	2.67	2.47	1.24	1.32	1.10
18	3.10	4.21	1.77	2.47	1.78	2.05	4.54	2.29	2.33	2.60
19	1.87	2.52	2.53	2.30	1.93	3.25	2.50	2.13	1.93	2.43

TABLE 2.3. Standardized Euclidean distances for all record pairs of Enumeration Area 12, Locality 3, Administrative Post 1 of Mandlacaze District

It is possible to link each unit of the CAP II survey with the nearest unit of the MCPH dataset, in accordance with the distance function (Formula 2.1), for example by solving the following optimization problem:

$$\min_{z} \sum_{a=1}^{n^A} \sum_{b=1}^{n^B} z_{ab} dist(a,b), \qquad (2.2)$$

subject to constraints $\sum_{a=1}^{v_A} z_{ab} \leq 1 \quad \forall b \quad \sum_{b=1}^{v_B} z_{ab} = 1 \quad \forall a$ and $z_{ab} \in \{0,1\} \quad \forall (a,b)$. Indeed, by solving the above Problem 2.3, we obtain the matching configuration with the minimum distances as expressed in Formula 2.1 satisfying the constraint that each unit of the second file must be matched with one unit of the first file. It must also be noted that this optimization problem is easily solved with the R package lpSolve. For the single block under analysis in this Section, we obtained the following list of pairs:

$$\{(3,1),(4,2),(13,3)(12,4),(7,5),(1,6)(17,7)(15,8)(14,9)(6,10)\}.$$

Some of these pairs show very similar profiles: for example, Pair (12,4), in which there is no livestock activity and the number of cashew and coconuts tree differ by only two units. Notice also Pair (6,10), in which all variables reported are equal to 0. This pair could be a true match if it corresponds to an household engaging in agricultural or livestock activities different from those reported by the key variables, but it could also be a false match, due to non-responses recorded as 0 values.

	CASHEW	COCONUTS	COWS	GOATS	SHEEPS	PIGS	CHICKENS	DUCKS
Min.	0	0	0	0	0	0	0	0
1st Qu.	1	0	0	0	0	0	1	0
Median	7	0	0	0	0	0	4	0
Mean	20	8.8	1.6	2.2	0.21	0.78	5.5	0.67
3st Qu.	20	5	0	3	0	1	7	0
	3500	2437	958	743	98	74	6000	100

Table 2.4: Summary statistics for the key variables in the MCPH 2007 dataset

	CASHEW	COCONUTS	COWS	GOATS	SHEEPS	PIGS	CHICKENS	DUCKS
Min.	0	0	0	0	0	0	0	0
1st Qu.	0	0	0	0	0	0	0	0
Median	2	0	0	0	0	0	4	0
Mean	16	5.6	5.3	3.7	0.52	1.1	6.7	0.44
3rd Qu.	10	2	6	5	0	1	10	0
Max.	580	310	96	106	38	29	60	32

TABLE 2.5. Summary statistics for the key variables in the CAP II dataset

We conclude this Section by comparing the distribution of the key variables in the two datasets. Tables 2.4 and 2.5 above report the summary information. The key variable distributions present some changes between the two datasets. In particular, in relation to the MPCH dataset, the location summaries of the CAPII survey are lower for the cashew and coconut trees, and higher for the livestock matching variables.

2.2 AIMS OF THE LINKAGE PROCEDURE

As we have already seen, the operation of merging two or more datasets can be important, to obtain statistical information that cannot be extracted from either one of the two individual datasets. Suppose, for example, that we wish to compare the average agricultural and livestock activity of Mozambique family farms in 2007 and 2009, i.e. at the times of the MCPH and of the CAP II surveys, to ascertain whether some of these activities have changed over time. From a statistical point of view, it would be best to obtain a panel of family farms, and to observe the changes occurring in these farms during the observation period. In the data at hand, such a panel is not immediately available; however, after the record linkage procedure, the matched units can be used across the two files for this very purpose. Consider, again, the data from the limited enumeration area presented in the previous Section, and the corresponding list of matched pairs, with regard to the distance function (2.1) and the minimization problem (2.3). Table 2.6 below reports the list of units and the number of cashew trees observed on the two occasions. With these data, it is possible to perform, for example, a test for paired data, to ascertain whether there has been a change in the average number of cashew trees in this enumeration area. For example, the t test statistics for paired data is equal to -1.0728, and the corresponding p-value is equal to 0.3113; this suggests that there have been no changes in this enumeration area.

Pairs	(3,1)	(4 ,2)	(13,3)	(12,4)	(7,5)	(1,6)	(17,7)	(15,8)	(14,9)	(6,10)
Cashews (MCPH)	12	15	1	7	20	10	10	4	0	0
Cashews (CAP II)	1	15	7	8	20	10	24	4	17	0

TABLE 2.6. Cashew trees for the 10 matched pairs of Enumeration Area 12, Locality 3, Administrative Post 1 of Mandlacaze District

Another type of analysis that can be performed with the linked data is a simple regression analysis, to predict the number of cashew trees in 2009 using the same variable observed in 2007 as covariate. Figure 2.2 below shows the scatter-plot for these two variables, considering the matched pairs and the estimated regression line thus ensuing.

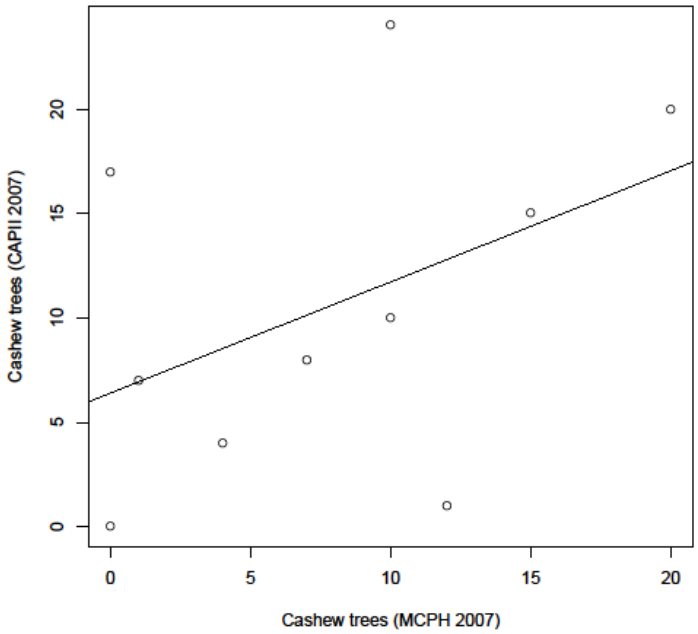

FIGURE 2.2. Scatter-plot for the number of cashew trees in 2007 and 2009 for the 10 matched pairs of Enumeration Area 12, Locality 3, Administrative Post 1 of Mandlacaze District

2.3 PRELIMINARY STEPS FOR THE RECORD LINKAGE PROCEDURE

To perform both the classical and the full Bayesian approach described in the previous Chapter, we must first categorize the matching variables. This is necessary to apply the Bayesian model described in Section 1.5.4 above, but it could also be valuable for the implementation of the comparison vector models, from both classical and Bayesian points of view. Indeed, without a categorization of the matching variables, minimal differences in a key variable would provide the same evidence as large differences, contrary to the scenario where a pair of records is co-referent; this is because the comparison value is equal to 1 only in the case of identical key variables. Therefore, we adopted the following rule: let z_α be the α% quantile of the distribution of the key variable V_j conditional on $V_j > 0$. The quantile z_α is to be estimated in light of the entire MPCH dataset. We then categorized the observed values for V_j into the following five classes: $V_j = 0$, $0 < V_j \leq z_{0.25}$, $z_{0.25} < V_j \leq z_{0.5}$, $z_{0.5} < V_j \leq z_{0.75}$ and $z_{0.75} < V_j \leq \max V_j$. In this way, two records that both had positive values for key variables, but that differed for small units, would receive greater evidence of being a match than two records that differ by the same amount, but one of which shows that the agricultural or livestock activity represented by the key variable was absent. The values

for the matching variables, categorized for the enumeration area described above, are listed in Table 2.6 below.

V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8
2	1	1	1	1	1	2	1
3	1	1	1	1	1	2	1
3	1	1	1	1	1	1	1
3	1	1	1	1	1	2	1
3	1	1	3	1	4	2	1
1	1	1	1	1	1	1	1
3	1	1	1	1	1	2	1
3	2	1	1	1	1	2	1
4	1	1	2	1	1	1	1
3	1	1	1	1	4	4	1
3	1	1	3	1	3	3	4
2	1	1	1	1	1	1	1
1	1	1	1	1	1	2	1
1	1	1	1	1	1	3	1
1	1	1	1	1	1	2	1
4	1	1	4	1	1	4	1
2	1	2	1	1	1	1	1
3	1	2	4	1	1	3	1
4	1	3	1	1	1	4	1

V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8
1	1	1	1	1	1	1	1
3	1	1	1	1	2	4	1
2	1	1	2	1	1	3	1
2	1	1	1	1	1	1	1
3	1	2	1	1	1	1	1
2	1	1	3	1	1	4	1
3	1	1	1	1	4	4	1
1	1	1	1	1	1	3	1
3	1	1	1	1	1	3	1
1	1	1	1	1	1	1	1

TABLE 2.7. Values of the categorized key variables observed in Enumeration Area 12, Locality 3, Administrative Post 1 of Mandlacaze District

2.4 RESULTS

In this Section, we illustrate the results obtained by matching the two Mozambique datasets with the various record linkage techniques presented in the Part 3, Chapter 1, and the transformation of the key variables into categorical variables, as outlined in Section 2.3 above. We begin by illustrating the results of the classical approach based on the comparison vectors, as defined in Section 2.3, in the single enumeration area introduced in Section 1. Table 2.7 below shows the frequency distribution for the comparison vector $q_{jj'} = (q_{jj'1}, \cdots, q_{jj'h})$, where

$$q_{jj'l} = \begin{cases} 1 & v_{1jl} = v_{2j'l} \\ 0 & v_{1jl} \neq v_{2j'l} \end{cases}, \qquad l = 1, \ldots, h.$$

Here, the frequency is greater than zero. Note that the 5 pairs agree with respect to all key variables, i.e. pairs $\{(6,1),(6,10),(10,7),(12,4),(14,8)\}$.

V1	V2	V3	V4	V5	V6	V7	V8	Freq	V1	V2	V3	V4	V5	V6	V7	V8	Freq
1	1	0	0	1	0	0	0	1	1	1	0	0	1	1	1	1	1
0	1	1	0	1	0	0	0	3	1	1	1	0	1	1	0	1	6
1	1	1	0	1	0	0	0	2	1	0	0	1	1	1	0	1	1
0	1	1	1	1	0	0	0	1	0	1	0	1	1	1	0	1	12
0	1	1	0	1	0	1	0	2	1	1	0	1	1	1	0	1	3
1	1	1	0	1	0	1	0	1	0	0	1	1	1	1	0	1	4
1	1	0	0	1	0	0	1	3	1	0	1	1	1	1	0	1	1
0	1	1	0	1	0	0	1	8	0	1	1	1	1	1	0	1	26
1	1	1	0	1	0	0	1	2	1	1	1	1	1	1	0	1	14
0	1	0	1	1	0	0	1	2	0	1	1	0	1	0	1	1	3
1	1	0	1	1	0	0	1	1	0	1	0	1	1	0	1	1	2
1	0	1	1	1	0	0	1	2	1	1	1	1	1	0	1	1	1
0	1	1	1	1	0	0	1	17	0	1	0	0	1	1	1	1	4
1	1	1	1	1	0	0	1	9	0	1	1	1	1	1	1	1	8
0	1	0	0	1	1	0	1	6	0	1	1	0	1	1	1	1	5
1	1	0	0	1	1	0	1	2	0	1	0	1	1	1	1	1	4
0	0	1	0	1	1	0	1	2	1	1	0	1	1	1	1	1	2
0	1	1	0	1	1	0	1	24	1	1	1	1	1	1	1	1	5

TABLE 2.8. Frequency distribution for the observed comparison vector in Enumeration Area 12, Locality 3, Administrative Post 1 of Mandlacaze District

With these comparison vector data, we estimated the mixture model

$$p(q_{jj'} \mid m,u,w) = w\prod_{l=1}^{h} m_l^{q_{jj'l}}(1-m_l)^{1-q_{jj'l}} + (1-w)\prod_{l=1}^{h} u_l^{q_{jj'l}}(1-u_l)^{1-q_{jj'l}}.$$

We thus obtained the following maximum likelihood estimates for parameters $m = (0.326,0.952,\ \ 0.801,0.699,1,0.761,0.227,0.991)$, $u = (0.184,0.928,\ 0.625,0.190,1,0.342,0.08,0.754)$ and $p = 0.816$.

As was expected, the estimates of parameter m are greater than the estimates of parameter u, since the m parameters represent the probability of a field agreement for the matches, while the u parameters represent the same probability for the non-matches. Once we have estimated parameters m and u, we can progress in the matching process by evaluating the likelihood ratios

$$\lambda = \frac{P(q_{jj'} \mid (j,j') \in M)}{P(q_{jj'} \mid (j,j') \in U)} = \frac{\prod_{l=1}^{h} m_l^{q_{jj'l}}(1-m_l)^{1-q_{jj'l}}}{\prod_{l=1}^{h} u_l^{q_{jj'l}}(1-u_l)^{1-q_{jj'l}}}.$$

V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	Freq	$\log \lambda$
0	0	0	0	1	0	0	0	0	-6.70
1	0	0	0	1	0	0	0	0	-5.90
0	1	0	0	1	0	0	0	0	-6.30
1	1	0	0	1	0	0	0	1	-5.50
0	0	1	0	1	0	0	0	0	-5.80
1	0	1	0	1	0	0	0	0	-5.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
0	1	0	1	1	1	1	1	4	2.60
1	1	0	1	1	1	1	1	2	3.40
0	0	1	1	1	1	1	1	0	3.10
1	0	1	1	1	1	1	1	0	3.80
0	1	1	1	1	1	1	1	8	3.50
1	1	1	1	1	1	1	1	5	4.30

Table 2.9: Logarithm of the likelihood ratio, for some comparison vectors

Table 2.9 above reports the value of the logarithm of λ for certain comparison vectors. As expected, the likelihood ratio λ increases with the number of agreements between the key variables. Note also the left panel of Figure 2.3, which shows the logarithm of the likelihood ratio λ with respect to the Euclidean distance (Formula 2.1) for each possible pair. Aside from a few observations, there is substantial agreement between the two ways of measuring the matching evidence, with the log-likelihood ratio decreasing in relation to the Euclidean distance (Formula 2.1). In the right panel of Figure 2.3, for each pair, we graphically represented the posterior probability of being a match

$$P((j,j') \in M \mid q_{j,j'}) = \frac{w\prod_{l=1}^{h} m_l^{q_{jj'l}} (1-m_l)^{1-q_{jj'l}}}{w\prod_{l=1}^{h} m_l^{q_{jj'l}} (1-m_l)^{1-q_{jj'l}} + (1-w)\prod_{l=1}^{h} u_l^{q_{jj'l}} (1-u_l)^{1-q_{jj'l}}}.$$

A high number of potential matches can be found with this approach. This can be explained by the fact that, due to the assumption that comparison vectors are independent, and to the absence of constraints on the matching process, the marginal matching probabilities depicted depend only on the information provided by the comparison, and not on the information provided by the entire dataset.

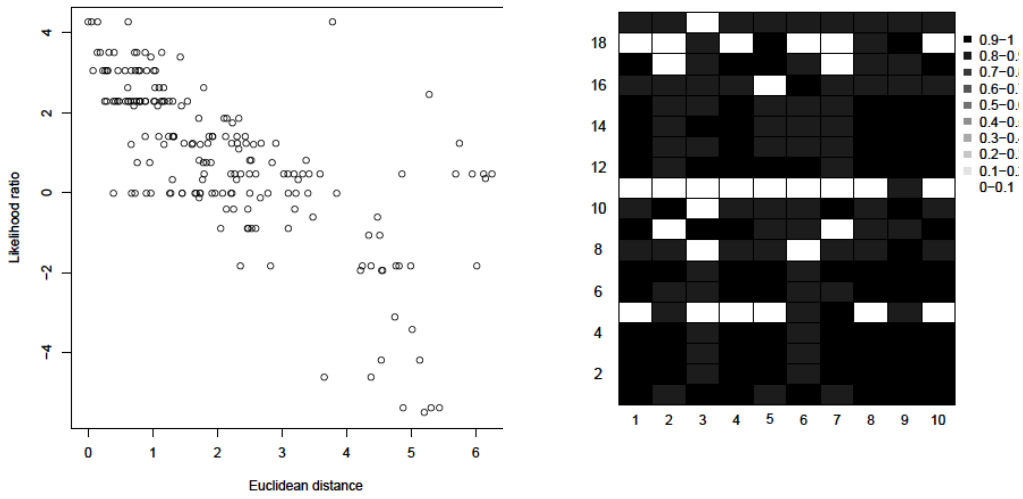

FIGURE 2.3. Left panel: scatter-plot for the log likelihood ratios against the Euclidean distance. Right panel: Marginal posterior probability of being a match for the classical approach

Finally, a 1-to-1 assignment of the CAP II sample units can be obtained by solving the problem

$$\max_{z} \sum_{j=1}^{n_1} \sum_{j'=1}^{n_2} z_{jj'} \log \lambda_{jj'} \ ,$$

subject to the constraints $\sum_{j=1}^{n_1} z_{jj'} \leq \ \forall j'$, $\sum_{j=1}^{n_2} z_{jj'} = 1 \ \ \forall j$ and $z_{jj'} \in \{0,1\} \ \ \forall (j,j')$. In this case, we obtained the following list of pairs:

$$\{(3,1),(4,2),(9,3)(12,4),(17,5),(16,6)(10,7)(14,8)(2,9)(6,10)\} \ ,$$

with 4 pairs common to the distance function approach described in Section 2.3.

2.4.1 SINGLE-BLOCK BAYESIAN ANALYSIS

We now show the results obtained for the enumeration area with the two different Bayesian models described in Part 3, Chapter 1 above. The first model is described in Section 3.4, i.e. the Bayesian version of the classical approach based on the comparison vectors. This model uses the same data used in the classical approach, i.e. the comparison vectors under the independence assumption, but automatically rules out multiple matches, by the means of the constraints on the matching matrix C. The other model, which we name the full Bayesian model, is the matching model described in Section 3.5, based on the categorical variables. For both models, we slightly changed the priority for the matching matrix to impose matching all CAP II sample units. This was obtained

simply by fixing the total number of matches and proposing, in the MCMC algorithms, only moves switching the matches of all CAP II units.

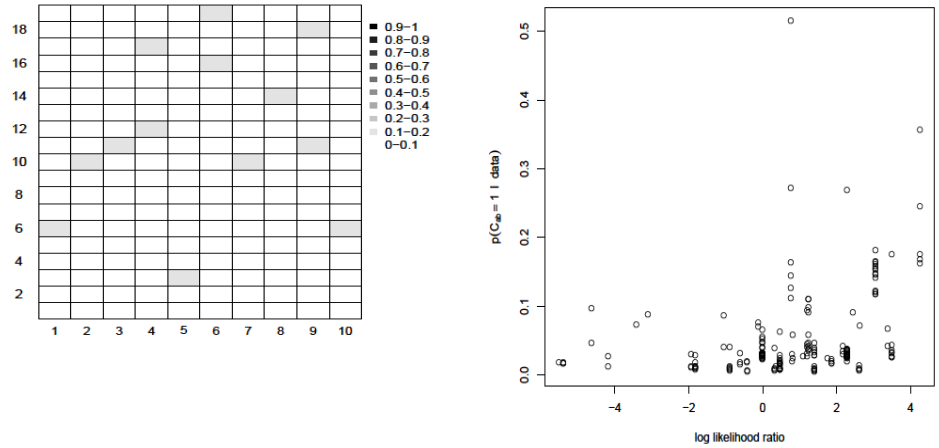

FIGURE 2.4. Left panel: $P(C_{jj'}=1\,|\,data)$ **using the Bayesian version of the comparison vector model. Right panel: Scatter-plot with** $P(C_{jj'}=1\,|\,data)$**, vs the the log-likelihood**

The left panel of Figure 2.4 above shows the marginal posterior probabilities $p(C_{jj'}=1\,|\,data)$ using the comparison vector model. Note that the maximum value for these probabilities is equal to 0.17, which is rather low. Therefore, at least for this example, when all sources of uncertainty are correctly accounted for, the information provided by the comparison vectors alone does not enable distinguishing between matches and non-matches. In this connection, it must be noted that the analogue posterior probability $P((j,j')\in M\,|\,q_{jj'})$ estimated with the classical approach does not take into account the uncertainty of parameters m and u, unlike the Bayesian marginal posterior probabilities $p(C_{jj'}=1\,|\,data)$. This partly explains the great difference between the right panel of Figure 2.3 and the left panel of Figure 2.4. However, when comparing the probabilities $p(C_{jj'}=1\,|\,data)$ with the logarithm of the likelihood ratios λ (see the right panel of Figure 2.4), there is substantial concordance between the two ways to measure the matching evidence, as the pairs with higher probabilities also have higher values for λ.

We conclude this Section by briefly illustrating the results of the full Bayesian approach. The left panel of Figure 2.4 shows the marginal posterior probabilities $p(C_{jj'}=1\,|\,data)$. The maximum value for these probabilities is now equal to 0.51; we have other pairs which present stronger evidence of being a match in relation to the Bayesian analysis of the comparison vector model. In our opinion, this means that by modeling all the observed values of the categorical, and not only the comparison, vector, we provide further evidence for distinguishing between matches and non-matches. Moreover, when comparing the probabilities $p(C_{jj'}=1\,|\,data)$ provided by the full Bayesian model with

the logarithm of the likelihood ratios $\lambda_{j,j'}$ (see the right panel of Figure 2.5 below), we have a lower correlation than in a case with , $p(C_{jj'}=1\,|\,data)$ estimated by the Bayesian version of the comparison vector model. Indeed, the likelihood ratio $\lambda_{j,j'}$ was estimated by the comparison vectors, even if by means of a classical framework; this fact should explain the higher correlation in the latter case.

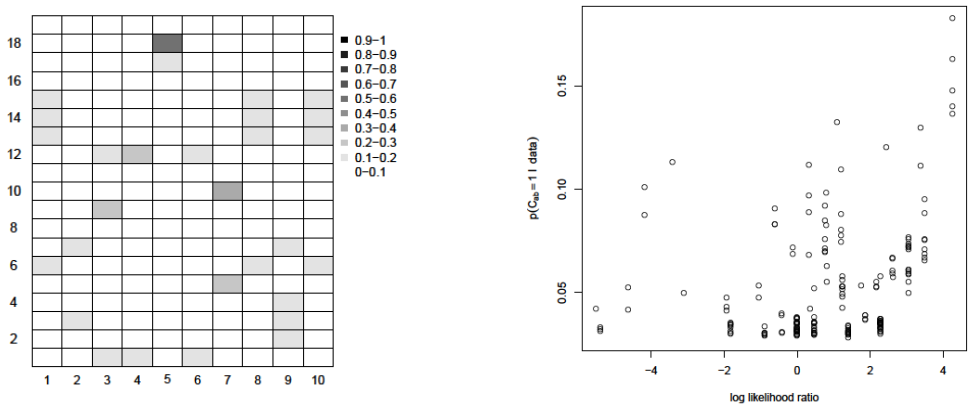

FIGURE 2.5. Left panel: $P(C_{jj'}=1\,|\,data)$ **using the full Bayesian model. Right panel: Scatter-plot with** $P(C_{jj'}=1\,|\,data)$ **vs the log-likelihood ratios of the classical approach**

2.4.2 EXAMPLES OF STATISTICAL INFERENCE WITH LINKED DATA

We conclude this Chapter by illustrating the results of the record linkage exercise performed on all 64 enumeration areas common to the two agricultural surveys. In particular, by applying the full Bayesian model and declaring all the pairs with $P(C_{j,j'}=1\,|\,data)>0.5$ as matches, we found 130 pairs.

Considering the number of cashew trees observed on these 130 units during the two sampling occasions, we performed a t-test for paired data, to evaluate the difference in the average number of cashew trees between the two sampling occasions; we obtained a non-significant result. The right panel of Figure 2.6 below shows the histogram of the distribution of the differences between the cashew tree numbers for these matched pairs. The outliers in both sides of the histogram may be false matches, introduced erroneously into the set of panel observations. These outliers could have a serious adverse impact on the resulting inference; their effect could be reduced by developing more appropriate models that are also able to both account for matching uncertainty, and to estimate the interesting parameters of the linked-data model. A similar argument was developed in Part 3, Chapter 1, when discussing the inference for regression models based on linked data. In this regard, the right panel of Figure 2.6 shows the scatter-plot with the number of cashew trees in 2009 ($y-$axis) and the same variable recorded in 2007 (x -axis), for all 130 units considered to be true matches. We also report the estimated regression line,

considering these 130 units as true matches. Note the outliers that pull the slope of the regression line towards zero. As discussed in Part 3, Chapter 1, this is a common situation for regression analysis based on linked data, and is due mainly to the presence of false matches. Indeed, for these pairs, the value of the response variable is independent of the corresponding values of the covariates.

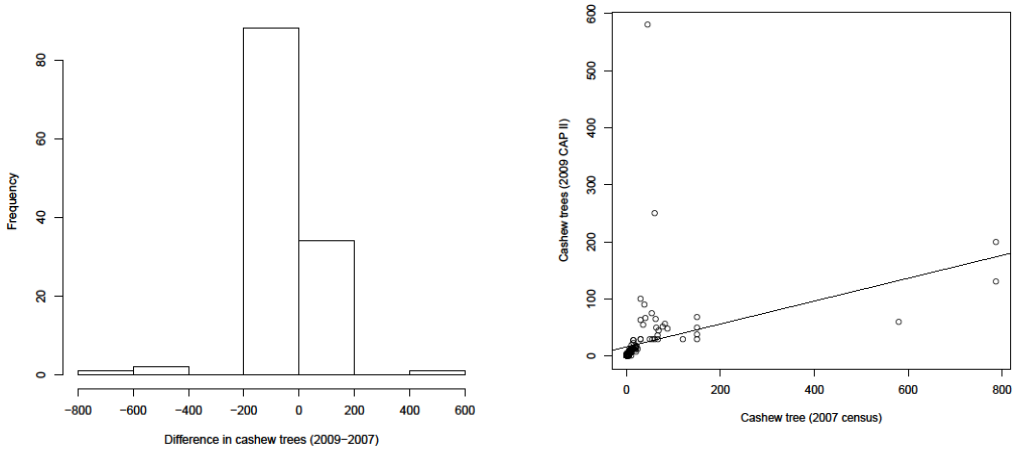

FIGURE 2.6. Left panel: Distribution of the differences between the cashew tree numbers for the matched pairs. Right panel: scatter-plot with the number of cashew trees in 2009 ($y-$axis) and the same variable recorded in 2007 (x-axis) for the matched pairs.

PART 4

Sampling

In this Part, we propose a sampling strategy for a multipurpose survey aimed at achieving an integrated observation of individuals, households, farms and land parcels (Chapter 1). The strategy adopted is the following: beginning with the observation of one population, the units of the other populations are surveyed by referring to the links connecting them to the units of the first population. Thus, like in an indirect sampling approach, the other populations can be considered sampled from an imperfect frame, i.e. the frame referring to the first population. We also consider the hypothesis of frame imperfections in observing the first population. Finally, a Multiple Frame approach is proposed. The simulation study (presented in Chapter 3) outlines some of the main strengths of the proposed integrated sampling strategy: (*i*) the estimator is design-unbiased; (*ii*) the loss in terms of efficiency when estimating a population total (compared to a strategy based on independent observations) is often negligible; and (*iii*) the proposed observational strategy is more accurate and efficient in estimating an integrated parameter referred to two distinct populations.

We also propose some solutions for optimal sampling design (Chapter 2). The observational strategy adopted for direct and indirect sampling design required some changes to be made to the standard techniques for determining the optimal inclusion probability. The applications to artificial data based on the real agricultural survey data of a developing country (Chapter 3) indicate that when the inclusion probabilities (the sample allocation or the sample size determination) are defined, taking into account the estimation targets of all integrated populations globally, the estimates' overall reliability improves. On the other hand, if the estimation targets are taken into account only partially, the risk of highly unreliable estimates is not entirely negligible.

Part 4A: Theory

1

Direct and Integrated Observation of Different Populations

1.1 INTRODUCTION

The basic concept proposed in this Chapter is the development of a framework for the joint observation of different populations, to estimate target parameters referring to units common to both populations considered. Thus, when information on land parcels, households and farms are interlinked, the methods developed in the following paragraphs enable consistent statistics on the environmental, social and economic dimensions of agriculture to be provided.

In any conventional survey, random selection of the sample requires an updated list, which records all individuals eligible for the survey (and only these), each identified by a label. This perfect list, i.e. the sampling frame, is used to identify the elements of the target population. When the sampling frame is available, one crucial statistical issue is the assessment of this list's actual coverage of the target population. A sampling frame is perfect when there is a one-to-one mapping of frame elements to target population elements. However, in statistical practice, perfect frames seldom exist, and problems always arise to disrupt the ideal one-to-one mapping. For example, the sampling frame can suffer from either or both under-coverage and over-coverage. There is under-coverage when the available frame is incomplete, because it includes only part of the target population, the missing elements cannot appear in any sample drawn for the survey. On the contrary, there is over-coverage when the sampling frame contains duplications of the same units or units that are not included in the target population. However, in statistical practice, there may also be frame imperfections of yet other types: for example, in certain circumstances, one may not possess the collection units desired, but rather another frame of units linked to the list of collection units. Also, although a frame may be available, in a dynamic environment it quickly becomes outdated, thus representing a situation that is rather different from the reality.

This Chapter considers all these issues in developing a unified sampling strategy for integration in agricultural and rural statistics, for developing countries.

The following strategy will be adopted: starting with the observation of one population, the units of the other populations are surveyed by reference to their links with the units of the first population. Thus, as would occur with an indirect sampling approach, the other populations can be considered sampled from an imperfect frame, i.e. the frame referring to the first population. We will also consider frame imperfections in the observation of the first population. Finally, a Multiple Frame approach is proposed.

Briefly, the methodological approach proposed herein extends the use of Indirect Sampling (Lavallée 1995, 2002) to the production of integrated estimates on more than one target population, in the context of Multiple Frame surveys (Hartley, 1974; Mecatti *et al*. 2007, 2011).

The techniques proposed are relatively flexible and can be tailored to the diverse information contexts that characterize the production of agricultural statistics in developing countries. Furthermore, under rather general conditions, they enable the production of unbiased statistics, thus overcoming most of the problems caused by imperfect sampling frames.

These two approaches can be combined through the concept of *Multiplicity*, first introduced by Birnbaum and Sirken (1965) in their presentation of Network Sampling as a strategy for surveying rare or elusive populations. Also known as Multiplicity Sampling or Snowball Sampling, this is a link-tracing sampling procedure in which a sample is obtained by following existing links from one respondent to another. This sampling methodology applies, for example, in estimating the country-prevalence of a rare disease, when a frame that fully represents the target population is not available. Selection units and target units may either coincide, be related or be unrelated, according to a one-to-many linkage rule. Thus, for each target unit, multiplicity is defined as the number of selection units to which it is linked, and a multiplicity-adjusted estimator is suggested. In Indirect Sampling, the notion of multiplicity is essentially the same, except that a many-to-many linkage pattern must be considered. To adjust for possible data duplication at the estimation stage, it is suggested to use the Generalized Weight Share Method (GWSM) to provide an estimation weight for each target unit in the selected sample; in fact, this is a multiplicity adjustment. On the other hand, in the context of Multiple Frames surveys, multiplicity is defined as the number of frames from which a unit can be selected.

The present Chapter is organized as follows. In Section 3, an integrated observational strategy based on the concept of Multiplicity is presented. The two alternative observational strategies are described in further detail in Sections 3 and 4. In Section 5, the problems of using an existing survey as a frame are considered. Section 6 presents a calibration strategy for estimation, as an instrument to enable the inclusion of the auxiliary information available from the estimates.

1.2 OBSERVATIONAL STRATEGY

Essentially, the observation is based on a two-step scheme.

- **Step 1** consists of selecting one pair from populations *I-H* or *L-F*. Then, data on the units of the populations of the chosen pair are collected.
    - ***Couple I-H.*** The observation can begin from either *I* (individuals) or from *H* (households). Starting from *I,* if individual *i* is ***observed***, then all individuals of her/his household *h* are observed too. Likewise, starting from *H,* if household *h* is ***observed***, then all individuals *i* belonging to it are also observed.
    - ***Couple L-F***. If the observation begins with *L*, the farm *f* to which the ***observed*** land parcel *l* belongs is also observed. Likewise, if the observation starts from *F*, information on the land is collected on all land parcels *l* belonging to the ***observed*** farm *f*.
- In **Step 2**, the units of the populations not observed in the first step are reached through indirect sampling, using the links with the units in the direct sample.
    - If populations *I-H* are observed first, then populations *L-F* are observed in Step 2. In particular, in this second step, data are collected on all the farms in which the individuals observed in Step 1 work. The information related to L are also collected for these farms.
    - If *L-F* are observed first, then populations *I-H* are observed in Step 2. In particular, in this second step, data are collected on all the households of the farm workers observed in Step 1. For these households, information related to *I* is also collected.

An example of this process is given in Figure 1.1 below. The oval shapes represent the households or the farms, and the circles represent the individuals.

- In **Step 1**, a sample of two households is selected. All individuals (A, B, C, D, E) living in the two households are also observed.
- In **Step 2**, Farm 1 (in which individuals A and D work) is observed, along with Farm 2 (where individual C works) and Farm 3 (in which individual E is the *farm-holder*). All the land of the three farms is also observed.

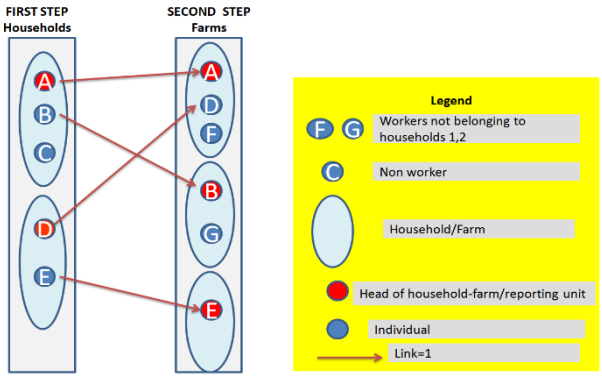

FIGURE 1.1. Example of the two steps of the observational strategy

In other words, the observation can be organized into these two alternative chains:

	Name of the chain	First step		Second step
Chain 1	From *households* to *farms*	*I-H*	→	*L-F*
Chain 2	From *farms* to *households*	*L-F*	→	*I-H*

Conceptually, the selection can begin from both land *I-H* and land *L-F*. Each choice entails pros and cons, which must be examined thoroughly, and in light of the country's specific information context.

However, the two chains are not entirely equivalent. Indeed, the chain from farms to households does not include urban households, i.e. those without any members working on a farm. In this case, if the survey seeks to compare rural and urban households, an additional survey specifically targeting the population of urban households must be conducted.

Highlights

Choosing the chain is a complex process, in which several different aspects must be considered. Some of these are:

- ✓ **the survey objective.** For example, if it is sought to conduct a survey to compare rural and urban households, the first chain is preferable, as it does not require additional samples.
- ✓ **existing registers.**
  - If a good quality register on farms is available, the second chain becomes more attractive.
  - Chain 2 is also more attractive if high-quality geographic information is available.
  - Chain 1 can be more convenient if a dataset with rich information for each Enumeration Area (e.g. data from the last Population Census) is available.
- ✓ **existing surveys conducted on a regular basis.**
  - Chain 1 may be more appropriate if a broad survey on population is conducted on a yearly basis.
  - Chain 2 may be more appropriate if a broad survey on farms is conducted on a yearly or biennial basis.

These two alternative chains will be examined in depth in the next two sections.

1.3 FROM HOUSEHOLDS TO FARMS

1.3.1 SAMPLING

Step 1

Let us suppose that the observation begins from populations I-H; let us assume, furthermore, that these populations can be reached through a collection of Q frames.

Let A_I^q be the generic frame, and denote with the same symbols (A_I^q) the number of elementary units recorded within it. Let v be the generic unit in the frame ($v = 1, \ldots, A_I^q$).

Let us suppose that the generic unit in frame v can be linked with one or more individuals I; it is then possible to define an $A_I^q \times I$ matrix, Λ^{qI}, with a generic element given by

$$\lambda_{vi}^q = \begin{cases} 1, & \text{if the } v-\text{th unit of } A_I^q \text{ is linked to the individual } i \\ 0, & \text{otherwise} \end{cases}.$$

The sample is selected according to the following steps:

1. **Sampling from the frames.** A sample Ω_I^q, is selected from each frame A_I^q (q=1,…,Q), with a generic sample design, where unit $v \in A_I^q$ is selected with an inclusion probability equal to π_v^q. At the end of these operations, we have Q different and partially overlapping samples from the frames.

2. **From frames to individuals.** Each sample Ω_I^q (q=1,…,Q) drives the definition of the sample $S_I^q = \{i \in u_h\colon (\lambda_{vi}^q = 1) \vee (\lambda_{vi'} = 1 \wedge i' \in u_h), v \in \Omega_I^q\}$ of individuals, which encompasses two main subsets:

- The first subset includes all individuals having a non-null link with the units v included in Ω_I^q.

- The second includes all individuals living in the same households of the individuals belonging to the first subset $S_I^q, q = 1, \dots, Q$.

3. **From individuals to households.** All households u_h of the individuals belonging to S_I^q are included in the sample of households; in symbols, $S_H^q = \{u_h \ni i\colon i \in S_I^q\}$. At the end of this operation, we have Q different and partially overlapping samples of households.

The overall process can be represented as Q chains of samples $\Omega_I^q \rightarrow S_I^q \rightarrow S_H^q$ (q=1,…,Q). Operations 2 and 3 can be reversed for some Q chains. Indeed, if the frame Ω_I^q can be linked only to households, for that frame the sample chain is $\Omega_I^q \rightarrow S_H^q \rightarrow S_I^q$.

At the end of this process, we obtain Q different and partially overlapping samples of individuals S_I^q, and of households, S_H^q.

Examples

Example 1. A_I^q can be a list of Enumeration Areas (EA) in which each v corresponds to a given EA. In this case:

- ✓ λ_{vi}^q is equal to 1 for all individuals living in the same EA.
- ✓ All households and individuals in the same EA have equal inclusion probabilities.
- ✓ Ω_I^q is a sample of EAs.
- ✓ The sample of individuals S_I^q and that of households S_H^q encompass all individuals and households living in the EAs selected in sample Ω_I^q .

Example 2. A_I^q can be a list of households collected during the previous Census. In this case:

- ✓ The elementary records v in the frame are the households collected in the previous Census. By the time of the current survey, the situation of a household may have changed for demographic or social reasons: e.g. births, deaths, weddings, immigration or emigration. Accordingly, A_I^q may represent an *outdated* register of *households* and sample Ω_I^q might contain an *outdated* list of households.
- ✓ λ_{vi}^q is equal to 1 for all individuals living in the same household during the previous Census' reference period.
- ✓ All individuals living in the same household during the previous Census' reference period are observed, with an equal inclusion probability.
- ✓ Sample S_I^q encompasses the individuals of Ω_I^q still alive, as well as the *individuals* living in their new household, even if these were not included in Ω_I^q.
- ✓ Sample S_H^q encompasses the new households of the Ω_I^q individuals still alive.

To ensure that the sampling strategy proposed is unbiased, the following condition must be fulfilled.

Condition 1.1. Each household u_h of H must have at least 1 non–null link with one of the Q frames, formally:

$$\sum_{q=1}^{Q}\sum_{v=1}^{A_I^q} \lambda_{vh}^q > 0 \text{ for } h, \dots, H \qquad (1.1)$$

being

$$\lambda_{vh}^q = \sum_{i=1}^{n_h} \lambda_{vi}^q \qquad (1.2)$$

i.e. the number of individuals of I represented by frame A_I^Q.

Step 2

In the second step, the sample is selected with the following operations:

1. **From individuals to farms.** Q different and partially overlapping samples of farms S_F^q $(q = 1, \cdots, Q)$ are selected through an Indirect Sampling mechanism. The sample S_F^q includes all farms in which individuals S_I^q work: $S_F^q = \{u_f \ni i: (\lambda_{if} = 1) \wedge (i \in S_I^q)\}$.

2. **From farms to land**. In the last sample of the chain, all the land parcels of the farms in S_F^q $(q = 1, \cdots, Q)$ are surveyed. Thus, Q indirect samples S_L^q from the population of land parcels are observed**.**

At the end of this process, we have Q different and partially overlapping samples of farms S_F^q, and Q different and partially overlapping samples of land parcels S_L^q.

To ensure that the sampling strategy proposed for Step 2 is unbiased, **Conditions 1.1** above, and **1.2** below, must be fulfilled.

Condition 1.2. Each farm u_f must have at least 1 non–null link with one of the I individuals in the H households, formally

$$\sum_{h=1}^{H} \sum_{i=1}^{n_h} \lambda_{if} > 0 \qquad (1.3)$$

In other words, the entire sequence of Q samples can be represented as follows:

$$\Omega_I^q \rightarrow S_I^q \rightarrow S_H^q \rightarrow S_F^q \rightarrow S_L^q \ \ (q = 1, \cdots, Q).$$

Highlights

It is absolutely crucial to ensure that:

- the population of households H is entirely covered by the union of the various frames. It is not necessary to ensure a perfect match between the records in the frame and the households; rather, the only mandatory condition is that each household is linked to at least one unit in the frames.
- the population of farms F can be obtained from the links with the population of households H. For example, the definition of H that excludes the households living in urban areas may cause distortions, if there are farms operated by individuals living there.

1.3.2 MULTIPLICITY IN THE OBSERVATIONAL STRATEGY DESCRIBED

Step 1

In the first step of the sampling process, the multiplicity consists in the following factors:

- A unit v in the q-th frame may be related to more than one individual of population I (see Table 2.1)
- A given individual may be included in more than one frame A_I^q (q=1,…,Q) (see Table 2.2)
- A household may include more than one individual who can be reached by a sampling mechanism

Thus, the multiplicity of an individual i is given by:

$$m_i = \sum_{q=1}^{Q} m_i^q = \sum_{q=1}^{Q} \sum_{v=1}^{A_I^q} \lambda_{vh}^q \quad \text{for} \quad i \in u_h \,, \tag{1.4}$$

being $\lambda_{vh}^q = \sum_{i=1}^{n_h} \lambda_{vi}^q$.

The unit multiplicity for a given household u_h is

$$m_h = \sum_{q=1}^{Q} m_h^q = \sum_{q=1}^{Q} \sum_{v=1}^{A_I^q} \lambda_{vh}^q \tag{1.5}$$

Thus, $m_i^q = m_h^q$ and $m_i = m_h$ for $i \in u_h$, i.e. the individuals belonging to the same household have the same multiplicity, and this multiplicity is equal to that of the household.

Step 2

In the sampling strategy illustrated above, the multiplicity factor of a farm u_f is given by the number of workers in the farm.

$$m_f = \sum_{h=1}^{H} \sum_{i \in u_h} \lambda_{if} = \sum_{h=1}^{H} n_{hf} = \sum_{h=1}^{H} \tilde{\lambda}_{hf} n_h \tag{1.6}$$

Note that m_f can be lower than n_f. This may occur if farm *f* includes individuals who do not belong to population *I*, e.g. if among the workers of f there are immigrants, while population I includes only local individuals.

Useful reformulations of the parameter of interest, taking into account the units' multiplicity

Let us consider, as target parameters, the population totals $\theta_I = \sum_{i=1}^{I} Y_i$, $\theta_L = \sum_{l=1}^{L} Y_l$, $\theta_H = \sum_{h=1}^{H} Y_h$, $\theta_F = \sum_{f=1}^{F} Y_f$ of populations I, L, H and F.

It must be noted that for the moment, it is sufficient to focus on parameters θ_H and θ_L, since parameters θ_I and θ_L represent specific subsets of the parameters θ_H and θ_F respectively. Indeed,

$$\theta_I = \sum_{h=1}^{H} \sum_{i \in u_h} Y_i = \sum_{h=1}^{H} Y_h = \theta_H \,, \quad \theta_L = \sum_{f=1}^{F} \sum_{l \in u_f} Y_l = \sum_{f=1}^{F} Y_f = \theta_F \,.$$

Parameter $\boldsymbol{\theta_H}$

Using the multiplicity factors defined above, a reformulation of parameter θ_H is:

$$\theta_H = \sum_{q=1}^{Q} \sum_{h=1}^{H} \alpha_h^q \, Y_h \tag{1.7}$$

where

$$\alpha_h^q = \frac{m_h^q}{m_h} \tag{1.8}$$

are fractional factors, adjusting for the multiplicity.

Theorem 1.1 If **Condition 1.1** holds true, the sum over *q* of α_h^q equals 1.

Proof:

$$\sum_{q=1}^{Q} \alpha_h^q = \sum_{q=1}^{Q} \frac{m_h^q}{m_h} = \frac{1}{m_h} \sum_{q=1}^{Q} \sum_{v=1}^{A_I^q} \lambda_{vh}^q = \frac{m_h}{m_h} = 1. \tag{1.9}$$

Conversely, if for a subset $\tilde{H}$ of *H*, it is

$$\sum_{v=1}^{A_I^q} \lambda_{vh}^q = 0 \text{ then } \sum_{q=1}^{Q} \alpha_h^q = 0 \text{ for } u_h \in \tilde{H}.$$

Under **Condition 1.1**, parameter θ_H may be reformulated as the result of Q sum totals, θ_H^q, (q=1,…,Q) referred to the frames A_H^q:

$$\theta_H = \sum_{q=1}^{Q}\sum_{h=1}^{H} \alpha_h^q Y_h = \sum_{q=1}^{Q}\sum_{h=1}^{H}\sum_{v=1}^{A_I^q} \frac{\lambda_{vh}^q}{m_h} Y_h = \sum_{q=1}^{Q}\sum_{v=1}^{A_I^q} Z_{vH}^q = \sum_{q=1}^{Q} \theta_H^q \quad (1.10)$$

being

$$Z_{vH}^q = \sum_{h=1}^{H} \frac{\lambda_{vh}^q}{m_h} Y_h \text{ and } \theta_H^q = \sum_{v=1}^{A_I^q} Z_{vH}^q \quad (1.11)$$

Parameter θ_F

The parameter of interest θ_F can be reformulated as

$$\theta_F = \sum_{q=1}^{Q}\sum_{f=1}^{F} \alpha_f^q Y_f \quad (1.12)$$

$$\text{where } \alpha_f^q = \sum_{h=1}^{H} \frac{n_{hf}}{m_f} \alpha_h^q . \quad (1.13)$$

Theorem 1.2 If Conditions 1.1 and 1.2 hold true, the sum over q of α_f^q equals 1.

Proof:

$$\sum_{q=1}^{Q} \alpha_f^q = \sum_{h=1}^{H} \frac{n_{hf}}{m_f} \sum_{q=1}^{Q} \alpha_h^q = 1$$

$$\text{since } \sum_{q=1}^{Q} \alpha_h^q = 1 \text{ under condition 1.1. and } \sum_{h=1}^{H} \frac{n_{hf}}{m_f} = \frac{1}{m_f} \sum_{h=1}^{H}\sum_{i=1}^{n_h} \lambda_{if} = \frac{m_f}{m_f}$$

under Condition 1.2.Conversely, if neither Condition 1.1 nor Condition 1.2 hold true, then

$$\sum_{q=1}^{Q} \alpha_f^q = 0.$$

Under Conditions 1.1 and 1.2, parameter θ_F may be reformulated as the result of Q sum totals, θ_F^q, (q=1,…,Q):

$$\theta_F = \sum_{q=1}^{Q}\sum_{f=1}^{F} \alpha_f^q \, Y_F = \sum_{q=1}^{Q}\sum_{f=1}^{F}\sum_{h=1}^{H} \frac{n_{hf}}{m_f}\alpha_h^q \, Y_F = \sum_{q=1}^{Q}\sum_{f=1}^{F}\sum_{h=1}^{H} \frac{n_{hf}}{m_f}\sum_{v=1}^{A_I^q} \frac{\lambda_{vh}^q}{m_h} \; Y_F$$

$$= \sum_{q=1}^{Q}\sum_{v=1}^{A_I^q} Z_{vF}^q = \sum_{q=1}^{Q} \theta_F^q \qquad (1.14)$$

$$\text{being } Z_{vF}^q = \sum_{h=1}^{H}\sum_{f=1}^{F} \frac{n_{hf}}{m_f}\frac{\lambda_{vh}^q}{m_h} Y_F \text{ and } \theta_F^q = \sum_{v=1}^{A_I^q} Z_{vF}^q \qquad (1.15)$$

1.3.3 ESTIMATION

Unified *Estimator* for Multiplicity

The Unified Multiplicity Estimator (UME) proposed in this section may be considered a direct generalization of either the Generalized Weight Share Method Estimator (GWSM; Lavallée, 2007) or the Unit Multiplicity Estimator (Mecatti, 2007). Indeed, the UME weights adjust for the multiplicity arising from both the Indirect Sampling process and the use of a set of frames for the first population surveyed.

The UME estimator assigns, to each unit of a selected cluster j ($j = h$ for an household or $j = f$ for a farm), a weight w_j^q that is equal for all individuals of the cluster. Therefore, the UME estimator for θ_H is

$$\hat{\theta}_H = \sum_{q=1}^{Q}\sum_{h=1}^{S_H^q} w_h^q Y_h \qquad (1.16)$$

in which

$$w_h^q = \frac{1}{m_h}\sum_{v=1}^{\Omega_I^q} \frac{\lambda_{vh}^q}{\pi_v^q} \qquad (1.17)$$

Introducing the sample indicator variables δ_v^q, which assume a value of 1 if the unit ν of frame A_I^q is included in the sample Ω_I^q, and equals 0 otherwise, the weights can be reformulated as follows:

$$w_h^q = \frac{1}{m_h} \sum_{v=1}^{A_I^q} \frac{\delta_v^q \ \lambda_{vh}^q}{\pi_v^q} \quad \text{and thus } \hat{\theta}_H = \sum_{q=1}^{Q} \sum_{h=1}^{H} w_h^q \ Y_h \ .$$

Highlights

For the computation of sampling weights w_h^q, the following operations must be performed.

1. **Sampling**. The *sample designer* must select *Q* samples Ω_I^q (*q* =1,...,*Q*), one from each frame. The inclusion probability π_v^q of unit *v* in sample Ω_I^q must be recorded and stored, in accordance with the direct sampling design adopted.
2. **Tracing**. This operation is conducted by the field *enumerator* and aims to identify the individual *i* or the household u_h that can be linked to the frame's record *v*. For example:
   - If the frame is a list of Enumeration Areas (EAs), and *v* is a specific EA, then in this operation all the households of the *v*-th EA are identified.
   - If the frame is a list of workers from the latest census and v is a specific worker, then this operation enables the worker and the household in which he/she lives to be reached.
3. **Collecting data for sample links**. In this step, the *field enumerator* must collect the data for each individual *i* of household u_h (traced in Step 2), for the computation of λ_{vh}^q (Formula 1.2). This variable indicates how many individuals in the household are linked to the sample unit *v*. Its value depends mainly on the frame's characteristics, e.g.:
   - If the frame is a list of EAs, and *v* is a specific EA, then all individuals of the households are linked to *v*; thus, $\lambda_{vh}^q = n_h$.
   - If the frame is a list of workers recorded during the reference period of the latest census, and v is a specific worker living in the household u_h, then $\lambda_{vh}^q = 1$.
4. **Collecting data for potential links**. The *field enumerator* must also collect data for the potential links of the households with each *Q* frame. For the household u_h, *Q* variables must be collected: m_h^q (*q*= 1,..., *Q*). Let us consider the above example in which *Q*=2 and there are two different frames, A_I^1 and A_I^2, A_I^1 being a list of enumeration areas and A_I^2 a list of agricultural workers recorded in the latest census. In this case,
   - $m_h^1 = n_h$.
   - m_h^2 is equal to the number of individuals in the household who were agricultural workers during the reference period of the most recent Census.
5. **Computing the weights**. The weights are computed according to Formula 1.17. The field operator collects the values of variables λ_{vh}^q and m_h^q (*q*=1,...,*Q*). The variable m_h is obtained by summing up the *Q* values m_h^q.

The UME for θ_F is

$$\hat{\theta}_F = \sum_{q=1}^{Q} \sum_{h=1}^{S_f^q} w_f^q \, Y_f \qquad (1.18)$$

in which

$$w_f^q = \sum_{h=1}^{S_H^q} \frac{n_{hf}}{m_f} \, w_h^q \qquad (1.19)$$

The following aspects must be noted:

- Household u_h may have been selected in more than one sample S_H^q. In this case, the final weight assigned to the household is equal to the sum of the weights assigned to each of the samples containing u_h.

- A farm u_f may have been selected in more than one sample S_F^q. In this case, the final weight assigned to u_f is equal to the sum of the weights corresponding to each sample including u_f.

Highlights

To compute sampling weights w_f^q, the following operations must be performed.

1. **Sampling and tracing.** The *sample enumerator* must select Q samples of farms S_F^q (q=1,...,Q) by means of an indirect sampling mechanism. All farms linked to households u_h belonging to sample S_H^q are included in S_F^q.
2. **Collecting the data for the sample links**. In this step, the *field enumerator* must collect the ratios n_{hf}/m_f for farms u_f. The value of each ratio is equal to the number of the households of the sample S_I^q to which the farm u_f is linked. Therefore, if the farm is linked to 3 households, then this number is equal to 3.
3. **Computing the weights**. The weights are computed according to Formula 1.17. Note that the variables w_h^q (h=1,..., u_{hf}) are defined in the *first step*. The field operator must collect the values of ratios $n_{hf}/\, m_f$.

The operators involved in computing the sampling weights are a sample designer, a sample enumerator, a questionnaire designer and a software specialist capable of linking data from different surveys.

Survey operations must be accurately defined, to ensure an adequate collection of the values of λ_{vh}^q, m_h^q and n_{hf}/m_f.

Bias

Theorem 1.3 If **Condition 1.1** holds, the estimator $\hat{\theta}_H$ is design-unbiased.

Proof. The expectation under the sample design, $E_P(\cdot)$, of w_h^q is equal to α_h^q:

$$E_P\left(w_h^q\right) = \frac{1}{m_h}\sum_{v=1}^{A_I^q} \frac{\lambda_{vh}^q \, E_P\left(\delta_v^q\right)}{\pi_v^q} = \frac{1}{m_h}\sum_{v=1}^{A_I^q} \lambda_{vh}^q = \frac{m_h^q}{m_h} = \alpha_h^q,$$

with $E_P\left(\delta_v^q\right) = \pi_v^q$.

From **Theorem 1.1**, it follows that

$$E_P\left(\hat{\theta}_H\right) = \sum_{q=1}^{Q}\sum_{h=1}^{H} E_P(w_h^q)\, Y_h \;\; = \sum_{q=1}^{Q}\sum_{h=1}^{H} \alpha_h^q \, Y_h \;\; = \theta_H.$$

Highlights

The only mandatory condition to ensure the unbiasedness of estimates $\hat{\theta}_H$ is that **Condition 1.1** must be fulfilled; each household u_h of *H* must have at least 1 non–null link with one of the *Q* frames.

The unbiasedness of the strategy is not based on the fact that the links are correct, but only on the fact that tracing and survey operations enable reaching all the households of the target population, starting from the existing frames.

Theorem 1.3 If **Conditions 1.1** and **1.2** hold, the estimator $\hat{\theta}_F$ is design-unbiased.

Proof. The expectation under the sample design of w_f^q is equal to

$$E_p(w_f^q) = \sum_{h=1}^{H} \frac{m_{hf}}{m_f} \, E_p(w_h^q) = \sum_{h=1}^{H} \frac{m_{hf}}{m_f} \, \alpha_h^q = \alpha_f^q.$$

Therefore, from **Theorem 1.2** it follows that

$$E_P\left(\hat{\theta}_F\right) = \sum_{q=1}^{Q}\sum_{f=1}^{F} E_P(w_f^q)\, Y_f \;\; = \sum_{q=1}^{Q}\sum_{f=1}^{F} \alpha_f^q \, Y_f \;\; = \theta_F.$$

Highlights

The only mandatory condition for ensuring the unbiasedness of the estimates $\hat{\theta}_F$ is that **Conditions 1.1**. and **1.2** are fulfilled, namely that each household u_h of *H* must have at least 1 non–null link with one of the *Q* frames, and that each farm u_f must have at least 1 non–null link with one of the *I* individuals in the *H* households.

Exact knowledge of the links between units of different populations is not necessary. The strategy's unbiasedness is not based on the fact that the links are correct, but only on the fact that tracing and survey operations enable reaching all the households and farms of the target population, starting from the existing frames.

It must also be noted that exact knowledge of the links between units of different populations is however necessary, when the focus is on estimating parameters related to the integrated parameters defined in **Chapter 4** of **Part 1**.

Variance

Estimator $\hat{\theta}_H$

The UME $\hat{\theta}_H$ can be expressed as

$$\hat{\theta}_H = \sum_{q=1}^{Q} \sum_{h=1}^{S_H^q} \frac{1}{m_h} \sum_{v=1}^{\Omega_I^q} \frac{\lambda_{vh}^q}{\pi_v^q} Y_h = \sum_{q=1}^{Q} \sum_{v=1}^{\Omega_I^q} \frac{Z_{vH}^q}{\pi_v^q} = \sum_{q=1}^{Q} \hat{\theta}_{H,ht}^q \quad (1.20)$$

in which the variables Z_{vH}^q are expressed in Equation 1.11, and $\hat{\theta}_{H,ht}^q$ is the *Horvitz-Thompson* estimator of the totals θ_H^q (see Formula 1.11).

If, as in most cases, samples Ω_I^q (q=1,…,Q) are selected independently, then the sampling variance of $\hat{\theta}_H$ is given by the sum of the variances of $\hat{\theta}_{H,ht}^q$:

$$V_P(\hat{\theta}_H) = \sum_{q=1}^{Q} V_P(\hat{\theta}_{H,ht}^q) \quad (1.21)$$

in which

$$V_P(\hat{\theta}_{H,ht}^q) = \sum_{v \in A_I^q} \sum_{v' \in A_I^q} (\pi_{vv'}^q - \pi_v^q \pi_{v'}^q) \frac{Z_{vH}^q}{\pi_v^q} \frac{Z_{v'H}^q}{\pi_{v'}^q} \quad (1.22)$$

with $\pi_{vv'}^q = E_P(\delta_v^q \, \delta_{v'}^q)$ as the joint inclusion probabilities of units v and v' of the frame A_I^q.

The sample estimator of the sampling variance is

$$\hat{V}_P(\hat{\theta}_H) = \sum_{q=1}^{Q} \hat{V}_P(\hat{\theta}^q_{H,ht}) \qquad (1.23)$$

with

$$\hat{V}_P(\hat{\theta}^q_{H,ht}) = \sum_{v\in\Omega^q_I} \sum_{v\prime\in\Omega^q_I} \frac{(\pi^q_{vv'} - \pi^q_v \pi^q_{v'})}{\pi^q_{vv'}} \frac{Z^q_{vH}}{\pi^q_v} \frac{Z^q_{v'H}}{\pi^q_{v'}} . \qquad (1.24)$$

Estimator $\widehat{\boldsymbol{\theta}}_{\boldsymbol{F}}$

With computations similar to those adopted for $\hat{\theta}_H$, it is clear that the UME $\hat{\theta}_F$ can be expressed as

$$\hat{\theta}_F = \sum_{q=1}^{Q} \sum_{v=1}^{\Omega^q_F} \frac{Z^q_{vF}}{\pi^q_v} = \sum_{q=1}^{Q} \hat{\theta}^q_{F,ht} \qquad (1.25)$$

in which variables Z^q_{vF} are expressed in Formula 1.15, and $\hat{\theta}^q_{F,ht}$ is the *Horvitz-Thompson* estimator of the total θ^q_F (see Formula 1.25).

Therefore, the variance of $V_P(\hat{\theta}_F)$ and its estimate $\hat{V}_P(\hat{\theta}_F)$ can be obtained with expressions similar to 1.22 and 1.13 above, by substituting variables Z^q_{vH} with variables Z^q_{vF}.

1.4. FROM FARMS TO HOUSEHOLDS

1.4.1 SAMPLING

First step

Let us reverse the observational chain and start the sample selection from the population of farms *F*, and assume that this population can be reached through a collection of Q frames.

Let A^q_F be the generic frame; we denote, with the same symbol (A^q_F), the number of elementary units recorded within it. Furthermore, suppose that the generic unit in A^q_F, denoted with v, can be linked with one or more farms in F.

Therefore, it is possible to define an $A^q_F \times F$ matrix, Λ^{qF}, where the generic element is given by

$$\lambda^q_{vf} \begin{cases} > 0 & \text{if the } v-\text{th unit of } A^q_F \text{ is linked to farm } f \\ = 0 & \text{otherwise} \end{cases} .$$

The sample selection is achieved by means of the following operations:

1. **Sampling from the frames.** A sample, Ω_F^q, is selected from each frame A_F^q (q=1,…,Q) with a generic sample design; the unit $v \in A_F^q$ is selected with an inclusion probability equal to π_v^q. At the end of this operation, we have Q different and partially overlapped samples from the frames.

2. **From frames to farms.** Each sample Ω_F^q (q=1,…,Q) founds the definition of the sample of farms S_F^q, which includes all farms linked to at least one unit v in Ω_F^q. Formally, $S_F^q = \left\{u_f:\ \lambda_{vf}^q > 0, v \in \Omega_F^q\right\}$.

3. **From farms to land.** All land parcels l of the farms belonging to S_F^q are included in the sample of land parcels; with symbols, $S_L^q = \left\{l: l \in u_f \wedge u_f \in S_F^q\right\}$. At the end of this operation, we have Q different and partially overlapping samples of land parcels.

The overall process may be represented as the union of Q chains of samples:

$$\Omega_F^q \rightarrow S_F^q \rightarrow S_L^q \quad (q=1,\ldots,Q).$$

In other words, at the end of the entire process, we have Q different and partially overlapping samples of farm land parcels, S_F^q and S_L^q.

Examples

Example 1. A_F^q can be a list of *points* in a geographic grid, with latitude and longitude coordinates and related information derived by satellite imagery on land use.
In this case:

- ✓ v is a given point in the grid
- ✓ λ_{vf}^q is equal to 1 only for one u_f
- ✓ Ω_F^q is a sample of points of a geographic grid
- ✓ The sample of farms S_F^q includes all farms of which a part of land includes one of the points selected in sample Ω_F^q
- ✓ The sample of land parcels S_L^q covers all land parcels of the farms included in S_F^q

Example 2. A_F^q can be a list of farms collected in a farm register. These registers usually include the largest farms, often described as *economic farms*. In this case:

- ✓ The elementary record *v* of the frame is one of the farms that belong to the farm register. When the survey is executed, a farm's actual situation may be different from that recorded in the available frame, for several reasons (splitting, merging or acquisition). Thus, A_F^q often represents an outdated register of farms and Ω_F^q is a sample containing an outdated list of farms. Furthermore, this register usually does not cover all populations of farms *F*.
- ✓ The linking variables can be defined in different ways, depending upon the organization of the information in the register:
  - λ_{vf}^q may be equal to the proportion of the area of the farm u_f shared with farm *v* in the frame. In this case, to reduce the risk of measurement error, the information on the register concerning the surface of farm *v* should be geo-referenced, and the surface of farm *f* should be measured with a GPS device. Alternatively, at a lower cost but with a greater risk of measurement error, this information could be collected by the farmholder of the actual farm u_f.
  - λ_{vf}^q may be equal to the proportion of the workers of farm u_f shared with farm *v*. In this case, the farm register should possess the names of the workers of farm *v*. Alternatively, at a lower cost but with a greater risk of measurement error, this information could be collected by the farmholder of farm u_f.
- ✓ The sample S_F^q encompasses the farms having a non-null link λ_{vf}^q with the farms belonging to the sample Ω_F^q, selected from the register.
- ✓ The sample of land parcels S_L^q covers all land parcels of the farms included in S_F^q.

To ensure the unbiasedness of the sampling strategy proposed, the following condition must be fulfilled.

Condition 1.3 Each farm u_f of *F* must have at least 1 non–null link with one of the Q frames; formally,

$$\sum_{q=1}^{Q}\sum_{v=1}^{A_F^q} \lambda_{vf}^q > 0 \text{ for } f = 1, \dots, F. \qquad (1.26)$$

Step 2

In the second step, the sample is selected through the following steps:

1. **From farms to households.** Q different and partially overlapping samples of households, S_H^q $(q = 1, \cdots, Q)$, are selected with an indirect sampling mechanism. The sample S_H^q includes all the households of the workers of the farms selected in the first step S_F^q: $S_H^q = \{u_h : \lambda_{hf} > 0 \wedge u_f \in S_F^q\}$.

2. **From households to individuals**. In the last sample of the chain, all individuals i in S_H^q $(q = 1, \cdots, Q)$ are surveyed. Thus, Q indirect samples S_I^q from I are observed**.** Formally, the sample of individuals is $S_I^q = \{i \in u_h : u_h \in S_F^q\}$.

Remark 1.1 Sub-sampling the households of large farms. In the case of economic farms, the number of households to be interviewed is often too great. This may cause an unexpected increase of survey costs. In these cases, a simple solution is to sub-sample the farms' households. This sub-sampling process is described in Section 1.4.4 below.

At the end of this process, we have Q different and partially overlapping samples of households S_H^q and individuals S_I^q.

To ensure the unbiasedness of the sampling strategy proposed for Step 2, **Condition 1.3** must be satisfied, along with the following condition:

Condition 1.4 Each household u_h must have at least 1 non–null link with one of the farms u_f in F; formally,

$$\sum_{f=1}^{F} \lambda_{hf} > 0 \qquad (1.27)$$

In other words, the entire sequence of Q samples may be represented as follows:

$$\Omega_F^q \rightarrow S_F^q \begin{cases} \rightarrow S_H^q \rightarrow S_I^q \\ \rightarrow S_L^q \end{cases}, (q=1,\ldots,Q).$$

Highlights

It must be ensured that:

- The population of farms is entirely covered by the union of the available frames. It is not necessary to ensure a perfect match between the records in the frame and the farms. The binding condition is that each farm is linked to at least one unit in the frames.
- The population of households *H* can be identified using the links with the population of farms *F*. This condition automatically excludes all households without any members working in farms F. Therefore, this observational strategy does not enable the production of unbiased statistics on the differences between rural and urban households.

1.4.2 MULTIPLICITY IN THE OBSERVATIONAL STRATEGY

Step 1

In the sampling process' first step, the multiplicity is related to the facts that:

- A unit v in the q-th frame may be related to more than one farm (see Example 2 above).
- A given farm may be included in more than one frame A_F^q (q=1,…,Q).
- A farm may include more than one land parcel, which can be reached by means of a sampling mechanism.

Thus, the multiplicity of a given farm u_f is:

$$m_f = \sum_{q=1}^{Q} m_f^q = \sum_{q=1}^{Q} \sum_{v=1}^{A_I^q} \lambda_{vf}^q \quad \text{for} \quad f = 1, \dots, F \qquad (1.28)$$

Step 2

In the sampling strategy illustrated above, the multiplicity factor of a household u_h is given by the number of farms in which its members work. Formally,

$$m_h = \sum_{f=1}^{F} \lambda_{hf}\,. \qquad (1.29)$$

Useful reformulations of the parameter of interest, taking into account the units' multiplicity

Parameter θ_F

Using the multiplicity factors defined above, the parameter of interest θ_H is reformulated below:

$$\theta_F = \sum_{q=1}^{Q} \sum_{f=1}^{F} \alpha_f^q \, Y_f \tag{1.30}$$

where

$$\alpha_f^q = \frac{m_f^q}{m_f} \tag{1.31}$$

are fractional factors, adjusting for the multiplicity.

Theorem 1.4 If **Condition 1.3** holds true, the sum over q of α_f^q equals 1.

Proof:

$$\sum_{q=1}^{Q} \alpha_f^q = \sum_{q=1}^{Q} \frac{m_f^q}{m_h} = \frac{1}{m_f} \sum_{q=1}^{Q} \sum_{v=1}^{A_F^q} \lambda_{fh}^q = \frac{m_f}{m_f} = 1. \tag{1.32}$$

Under **Condition 1.3**, parameter θ_F may be reformulated as the result of Q sum totals, θ_F^q, (q=1,…,Q) with reference to the frames A_F^q:

$$\theta_F = \sum_{q=1}^{Q} \sum_{f=1}^{F} \alpha_f^q \, Y_f = \sum_{q=1}^{Q} \sum_{f=1}^{F} \sum_{v=1}^{A_F^q} \frac{\lambda_{vf}^q}{m_f} \, Y_f$$

$$= \sum_{q=1}^{Q} \sum_{v=1}^{A_F^q} Z_{vF}^q = \sum_{q=1}^{Q} \theta_F^q \,, \tag{1.33}$$

being

$$Z_{vF}^{q} = \sum_{f=1}^{F} \frac{\lambda_{vf}^{q}}{m_f} Y_f \text{ and } \theta_F^q = \sum_{v=1}^{A_F^q} Z_{vF}^{q}. \quad (1.34)$$

Parameter $\boldsymbol{\theta_H}$

The parameter of interest θ_H can be reformulated as

$$\theta_H = \sum_{q=1}^{Q} \sum_{h=1}^{H} \alpha_f^q \, Y_f \quad (1.35)$$

where

$$\alpha_h^q = \sum_{f=1}^{F} \frac{\lambda_{hf}}{m_h} \alpha_f^q \quad (1.36)$$

Theorem 1.5 If Conditions 1.3 and 1.4 hold true, the sum over q of α_h^q equals 1.

Proof:

$$\sum_{q=1}^{Q} \alpha_h^q = \sum_{f=1}^{F} \frac{\lambda_{hf}}{m_h} \sum_{q=1}^{Q} \alpha_f^q = 1$$

since

$$\sum_{q=1}^{Q} \alpha_f^q = 1 \text{ under } \textbf{Condition 1.3}, \text{and } \sum_{f=1}^{F} \frac{\lambda_{hf}}{m_h} = \frac{m_h}{m_h} = 1 \text{ under } \textbf{Condition 1.4}.$$

Under **Conditions 1.3** and **1.4**, parameter θ_H can be reformulated as the result of Q sum of totals, θ_H^q, (q=1,…,Q):

$$\theta_H = \sum_{q=1}^{Q} \sum_{h=1}^{H} \alpha_h^q \, Y_h = \sum_{q=1}^{Q} \sum_{h=1}^{H} \sum_{f=1}^{F} \frac{\lambda_{hf}}{m_h} \alpha_f^q \, Y_h$$

$$= \sum_{q=1}^{Q} \sum_{h=1}^{H} \sum_{f=1}^{F} \frac{\lambda_{hf}}{m_h} \sum_{v=1}^{A_F^q} \frac{\lambda_{hf}^{q}}{m_f} \, Y_h = \sum_{q=1}^{Q} \sum_{v=1}^{A_F^q} Z_{vH}^{q} = \sum_{q=1}^{Q} \theta_H^q, \quad (1.37)$$

with

$$Z_{vH}^{q} = \sum_{f=1}^{F} \sum_{h=1}^{H} \frac{\lambda_{hf}}{m_h} \frac{\lambda_{vf}^{q}}{m_f} Y_h \text{ and } \theta_H^q = \sum_{v=1}^{A_F^q} Z_{vH}^{q} \qquad (1.38)$$

1.4.3. ESTIMATION

Unified Estimator for Multiplicity

The UME estimators for θ_F is

$$\hat{\theta}_F = \sum_{q=1}^{Q} \sum_{f=1}^{S_F^q} w_f^q \, Y_f \qquad (1.39)$$

where

$$w_f^q = \frac{1}{m_f} \sum_{v=1}^{\Omega_F^q} \frac{\lambda_{vf}^{q}}{\pi_v^q} \qquad (1.40)$$

Introducing the sample indicator variable δ_v^q, it is possible to reformulate the weights as

$$w_f^q = \frac{1}{m_f} \sum_{v=1}^{A_F^q} \frac{\delta_v^q \; \lambda_{vf}^{q}}{\pi_v^q}$$

Highlights

To compute sampling weights w_f^q , the following operations are required.

1. **Sampling.** The *sample designer* must select Q samples Ω_F^q (q =1,...,Q). The inclusion probabilities π_v^q of units v in sample Ω_F^q must be recorded and stored.
2. **Tracing.** This operation is performed by the *field enumerator*, and aims to identify the farm u_f that can be linked to record v in the frame. For example:
   - If the frame is a list of points in a geographic grid, and v is a specific point, then with this operation, the farm including this point of area is identified.
   - If the frame is a register of economic farms (units v are specific farms), then this operation leads to a sample of economic farms.
3. **Collecting the data for the sample links.** In this step, the *field enumerator* must collect the data for each farm u_f (described in Step 2), to compute the variable λ_{vf}^q. This variable indicates the amount of potential links with the frame. Its value depends mainly on the frame's characteristics, e.g.:
   - If the frame is a list of points in a geographic grid, and v is a given point, then λ_{vf}^q is equal to the ratio between the farm area and the conventional area around the point.
   - If the frame is a register of economic farms and v is a specific farm, then λ_{vf}^q can be either the proportion of the area of farm u_f shared with farm v, or the proportion of the workers of farm u_f shared with farm v.
4. **Collecting the data for potential links.** The *field enumerator* must collect the data for the potential links of the farms with each Q frame. For farm u_f, Q variables must be collected: $m_f^q = \sum_{v=1}^{A_F^q} \lambda_{vf}^q$ (q= 1,..., H). Let us consider the example in which Q=2 and there are two different frames, A_F^1 and A_F^2; A_F^1 is a list of points in a geographic grid and A_F^2 is a register of economic farms. In this case,
   - m_f^1 is the area of farm u_f;
   - m_f^2 is either the area of the farm or the number of workers of farm u_f who were included in economic farms v when the register was constructed.
5. **Computing the weights.** The weights are computed according to Formula 1.40. Note that variable π_v^q is defined by the *sample designer*. The field operator collects the values of variables λ_{vf}^q and m_f^q (q=1,...,Q). Variable m_f is obtained during the computation phase, by adding Q values m_f^q.

The UME estimator for θ_H is

$$\hat{\theta}_H = \sum_{q=1}^{Q} \sum_{h=1}^{S_H^q} w_h^q \, Y_h \qquad (1.41)$$

where

$$w_h^q = \sum_{f=1}^{S_F^q} \frac{\lambda_{hf}}{m_h} \, w_f^q \, . \qquad (1.42)$$

Highlights

To compute sampling weights w_h^q , the following operations are required.

1. **Sampling and tracing**. The *sample enumerator* must select Q samples of households S_H^q (q=1,...,Q) by means of an indirect sampling mechanism. All the households linked to a farm u_f in S_F^q are surveyed.
2. **Collecting the data for the sample links.** In this step, the *field enumerator* must collect the ratios λ_{hf}/m_h for each household surveyed. The number of these ratios u_{hf} is equal to the number of households in S_H^q to which a farm u_f is linked. Therefore, if a given household is linked to 3 farms sampled in Step 1, then $u_{hf} = 3$.
3. **Computing the weights**. The weights are computed according to Formula 1.42. Note that the variables w_f^q are defined in the *first step*. The field operator must collect the values of ratios $\lambda_{hf}/\, m_h$.

Let us highlight the following aspects:

- A farm u_f may have been selected in more than one sample S_F^q. In this case, the final weight assigned to the farm is equal to the sum of the weights assigned in each sample containing u_f.
- The household u_h may have been selected in more than one sample S_H^q. In this case, the final weight assigned to the household is equal to the sum of the weights assigned in each sample containing u_h.

Bias

Theorem 1.6 If Condition 1.3 holds, the estimator $\hat{\theta}_F$ is design-unbiased.

Proof:

The expectation under the sample design, $E_P(\cdot)$, of w_h^q is equal to α_h^q:

$$E_P\left(w_f^q\right) = \frac{1}{m_f}\sum_{v=1}^{A_F^q} \frac{\lambda_{vf}^q\, E_P\left(\delta_v^q\right)}{\pi_v^q} = \frac{1}{m_f}\sum_{v=1}^{A_F^q} \lambda_{vf}^q = \frac{m_f^q}{m_f} = \alpha_f^q$$

being $E_P\left(\delta_v^q\right) = \pi_v^q$.

Therefore, for **Theorem 1.4** it is

$$E_P\left(\hat{\theta}_F\right) = \sum_{q=1}^{Q}\sum_{f=1}^{F} E_P(w_f^q)\, Y_f \;\; = \sum_{q=1}^{Q}\sum_{f=1}^{F} \alpha_f^q\, Y_f \;\; = \theta_F.$$

Highlights

The only mandatory condition for ensuring the unbiasedness of estimates $\hat{\theta}_F$ is that **Condition 1.3** must be fulfilled. Each farm u_f of *F* must have at least 1 non–null link with one of the *Q* frames.
The unbiasedness of the strategy is not based on the fact that the links are correct, but only on the fact that tracing and survey operations enable reaching all the farms of the target population, starting from the existing frames.

Theorem 1.7 If Conditions 1.3 and 1.4 hold true, estimator $\hat{\theta}_H$ is design-unbiased.

Proof:

The expectation under the sample design of w_h^q is equal to

$$E_p(w_h^q) = \sum_{f=1}^{F} \frac{\lambda_{hf}}{m_h} E_p(w_f^q) = \sum_{f=1}^{F} \frac{\lambda_{hf}}{m_h} \alpha_f^q = \alpha_h^q.$$

Therefore, for **Theorem 1.5**, it is

$$E_P(\hat{\theta}_H) = \sum_{q=1}^{Q} \sum_{h=1}^{H} E_P(w_h^q)\, Y_h = \sum_{q=1}^{Q} \sum_{h=1}^{H} \alpha_h^q\, Y_h = \theta_H.$$

Highlights

The only mandatory condition for ensuring the unbiasedness of estimates $\hat{\theta}_H$ is that **Conditions 1.3** and **1.4** are simultaneously fulfilled:

- each farm u_f of *F* must have at least 1 non–null link with one of the *Q* frames;
- each household u_h must have at least 1 non–null link with one of the *F* farms.

Exact knowledge of the links between units of different populations is not necessary. The unbiasedness of the strategy is not based on the fact that the links are correct, but only on the fact that tracing and survey operations enable reaching all farms and all households of the target population, on the basis of the existing frames.

However, exact knowledge of the links between units of different populations is necessary when the focus is on the estimation of parameters related to the integrated parameters defined in Part 1, Chapter 3.

Variance

Estimator $\widehat{\boldsymbol{\theta}}_{\boldsymbol{F}}$

The UME $\hat{\theta}_F$ may be expressed as

$$\hat{\theta}_F = \sum_{q=1}^{Q} \sum_{f=1}^{S_F^q} \frac{1}{m_f} \sum_{v=1}^{\Omega_F^q} \frac{\lambda_{vf}^q}{\pi_v^q} Y_f = \sum_{q=1}^{Q} \sum_{v=1}^{\Omega_F^q} \frac{Z_{vF}^q}{\pi_v^q} = \sum_{q=1}^{Q} \hat{\theta}_{F,ht}^q \quad (1.43)$$

in which variables Z_{vF}^q are expressed in Formula 1.34 and $\hat{\theta}_{F,ht}^q$ is the *Horvitz-Thompson* estimator of the total θ_F^q (see Formula 1.34). If, as in most cases, the samples Ω_F^q (q=1,…,Q) are selected independently, then the sampling variance of $\hat{\theta}_F$ is given by the sum of the variances of $\hat{\theta}_{F,ht}^q$:

$$\begin{aligned} V_P(\hat{\theta}_F) &= \sum_{q=1}^{Q} V_P(\hat{\theta}_F) \\ &= \sum_{q=1}^{Q} \sum_{v \in A_F^q} \sum_{v' \in A_F^q} (\pi_{vv'}^q - \pi_v^q \pi_{v'}^q) \frac{Z_{vF}^q}{\pi_v^q} \frac{Z_{v'F}^q}{\pi_{v'}^q}, \end{aligned} \quad (1.44)$$

where $\pi_{vv'}^q = E_P(\delta_v^q \, \delta_{v'}^q)$ is the joint inclusion probability of the units v and v' of frame A_F^q. The sample estimator of the sampling variance is

$$\begin{aligned} \hat{V}_P(\hat{\theta}_F) &= \sum_{q=1}^{Q} \hat{V}_P(\hat{\theta}_F) \\ &= \sum_{q=1}^{Q} \sum_{v \in \Omega_F^q} \sum_{v' \in \Omega_F^q} \frac{(\pi_{vv'}^q - \pi_v^q \pi_{v'}^q)}{\pi_{vv'}^q} \frac{Z_{vF}^q}{\pi_v^q} \frac{Z_{v'F}^q}{\pi_{v'}^q}. \end{aligned} \quad (1.45)$$

Estimator $\widehat{\boldsymbol{\theta}}_{\boldsymbol{H}}$

As for $\hat{\theta}_F$, the UME $\hat{\theta}_H$ can be expressed as

$$\hat{\theta}_H = \sum_{q=1}^{Q} \sum_{v=1}^{\Omega_H^q} \frac{Z_{vH}^q}{\pi_v^q} = \sum_{q=1}^{Q} \hat{\theta}_{H,ht}^q \quad (1.46)$$

where the variables Z_{vH}^q are expressed in Formula 1.38 and $\hat{\theta}_{H,ht}^q$ is the *Horvitz-Thompson* estimators of the population total θ_H^q (see Formula 1.38). Therefore, the variance of $V_P(\hat{\theta}_H)$ and its estimate $V_P(\hat{\theta}_H)$ can be obtained through expressions similar to 1.44 and 1.45 above, by substituting variable Z_{vF}^q with variable Z_{vF}^q.

1.4.4 SUB-SAMPLING THE HOUSEHOLDS OF LARGE FARMS

With regard to economic farms, the number of households to be interviewed is often too large. In this case, a simple solution is to sub-sample the households of the large farms.

Sampling

The following procedure is proposed:

- **Establishing a threshold**. A threshold defining the maximum number of households to be observed per farm is established, having due consideration of the survey's budget and relevant operational constraints. This number is denoted by $\bar{n}_{2H}$, e.g. $\bar{n}_{2H} = 10$.
- **Identifying the number of households for each farm.** The households linked to a given farm are counted. This number of households is denoted with H_f. For example, if farm u_f includes 100 workers, belonging to 15 different households, then $H_f = 15$.
- **Sub-sampling the farms**. If, for a farm $u_f \in S_F^q$, $H_f > \bar{n}_{2H}$, then $\bar{n}_{2H}$ households are selected out of the H_f, with simple random sampling without replacement. This represents a second stage of sampling. Therefore, each household in the farm is selected, with an inclusion probability equal to

$$\pi_{2h}^q = E_p\left(\delta_{2h}^q \middle| (u_f \in S_F^q) \wedge (\lambda_{hf} > 0)\right)$$

$$= \begin{cases} \dfrac{H_f}{\bar{n}_{2H}} \text{ if } H_f > \bar{n}_{2H} \\ 1 \quad \text{if } H_f \leq \bar{n}_{2H} \end{cases}, \qquad (1.47)$$

where δ_{2h}^q is a dummy variable, with $\delta_{2h}^q = 1$ if the household is included in the second stage sampling and $\delta_{2h}^q = 1$ otherwise.

Estimation

The UME estimator for θ_H is

$$\hat{\theta}_H = \sum_{q=1}^{Q} \sum_{h=1}^{S_H^q} w_h^q \, Y_h \qquad (1.48)$$

where

$$w_h^q = \sum_{f=1}^{S_F^q} \frac{\lambda_{hf}}{m_h} \frac{1}{\pi_{2h}^q} \, w_f^q \, . \qquad (1.49)$$

Theorem 1.8 If Conditions 1.3 and 1.4 hold true, the estimator $\hat{\theta}_H$, with weights defined by Formula 1.49, is design-unbiased.

Proof: The expectation under the sample design of w_h^q is equal to

$$\begin{aligned} E_p(w_h^q) &= \sum_{f=1}^{S_F^q} \frac{\lambda_{hf}}{m_h} \frac{1}{\pi_{2h}} \, E_p(\delta_{2h} \, w_f^q) \\ &= \sum_{f=1}^{F} \frac{\lambda_{hf}}{m_h} \, \frac{1}{\pi_{2h}} \alpha_f^q E_p\left(\delta_{2h}^q | \, (u_f \in S_F^q) \wedge (\lambda_{hf} > 0\right) \\ &= \sum_{f=1}^{F} \frac{\lambda_{hf}}{m_h} \, \frac{1}{\pi_{2h}} \alpha_f^q \, \pi_{2h} = \alpha_h^q. \end{aligned}$$

Therefore,

$$E_P\left(\hat{\theta}_H\right) = \sum_{q=1}^{Q} \sum_{h=1}^{H} E_P(w_h^q) \, Y_h \; = \sum_{q=1}^{Q} \sum_{h=1}^{H} \alpha_h^q \, Y_h \; = \theta_H.$$

Variance

To define the variance of $\hat{\theta}_H$, we must break down variance in a two-stage sampling:

$$V_P(\hat{\theta}_H) = V_{1P} E_{2P}(\hat{\theta}_H) + E_{1P} V_{2P}(\hat{\theta}_H) \qquad (1.50)$$

where V_{1P} and E_{1P} are the variance and the expectation of sampling from frames A_{vF}^q; V_{2P} and E_{2P} are, respectively, the variance and the expectation of sub-sampling concerning the households from large farms. It is trivial to demonstrate that

$$V_{1P} E_{2P}(\hat{\theta}_H) = \sum_{q=1}^{Q} \sum_{v \in A_F^q} \sum_{v' \in A_F^q} (\pi_{vv'}^q - \pi_v^q \pi_{v'}^q) \frac{Z_{vH}^q}{\pi_v^q} \frac{Z_{v'H}^q}{\pi_{v'}^q} \quad (1.51)$$

Resorting to Expression 11.12 formulated by Cochran (1977, p. 301), the second addendum of Equation 1.50, $E_{1P}V_{2P}(\hat{\theta}_H)$, may be expressed as:

$$E_{1P}V_{2P}(\hat{\theta}_H) = \sum_{q=1}^{Q}\sum_{f=1}^{F} E_{1p}\ V_{2p}\left(\hat{Y}_{fH}\right) \qquad (1.52)$$

in which

$$\hat{Y}_{fH} = \sum_{h=1}^{H_f} w_h^q\ \delta_{2h}^q Y_h\,, \quad E_{1p}V_{2p}\left(\hat{Y}_{fH}\right)$$

$$= \sum_{h=1}^{H_f} E_{1p}(w_h^q)^2\ \frac{H_f\ (H_f - \bar{n}_{2H})}{\bar{n}_{2H}}\ \sigma_{fH}^2\,,$$

$$\sigma_{fH}^2 = \frac{1}{H_f}\left[\sum_{h=1}^{H_f} Y_h - \left(\sum_{h=1}^{H_f} Y_h \Big/ H_f\right)\right]^{2.}$$

Expectation $E_{1p}(w_h^q)^2$ is given by

$$E_{1p}(w_h^q)^2 \ = \sum_{f=1}^{F}\left(g_{fh}^q\right)^2 E_{1p}\left(\frac{1}{m_f}\sum_{v=1}^{A_F^q}\ \frac{\delta_v^q\ \lambda_{vf}^q}{\pi_v^q}\right)^2 +$$

$$+\sum_{f=1}^{F}\sum_{f\prime\neq f} g_{fh}^q\ g_{f\prime h}^q\ E_{1p}\left(\frac{1}{m_f}\sum_{v=1}^{A_F^q}\ \frac{\delta_v^q\ \lambda_{vf}^q}{\pi_v^q}\right)\left(\frac{1}{m_{f\prime}}\sum_{v=1}^{A_F^q}\ \frac{\delta_v^q\ \lambda_{vf\prime}^q}{\pi_v^q}\right),$$

in which $g_{fh}^q = \dfrac{\lambda_{hf}}{m_h}\dfrac{1}{\pi_{2h}^q}$ and

$$E_{1p}\left(\frac{1}{m_f}\sum_{v=1}^{A_F^q}\ \frac{\delta_v^q\ \lambda_{vf}^q}{\pi_v^q}\right)^2 = \frac{1}{m_f^2}\left[\sum_{v=1}^{A_F^q}\pi_v^q\ \left(\frac{\lambda_{vf}^q}{\pi_v^q}\right)^2 + \sum_{v=1}^{A_F^q}\sum_{v\prime=v}\pi_{vv\prime}^q\ \ \frac{\lambda_{vf}^q}{\pi_v^q}\frac{\lambda_{v\prime f}^q}{\pi_{v\prime}^q}\ \right],$$

$$E_{1p}\left(\frac{1}{m_f}\sum_{v=1}^{A_F^q}\ \frac{\delta_v^q\ \lambda_{vf}^q}{\pi_v^q}\right)\left(\frac{1}{m_{f\prime}}\sum_{v=1}^{A_F^q}\ \frac{\delta_v^q\ \lambda_{vf\prime}^q}{\pi_v^q}\right)$$

$$= \frac{1}{m_f\ m_{f'}}\sum_{v=1}^{A_F^q}\ \frac{\lambda_{vf}^q\ \lambda_{vf'}^q}{\pi_v^q} + \sum_{v=1}^{A_F^q}\sum_{v\prime\neq v}\lambda_{vf}^q\ \lambda_{vf'}^q\ .$$

1.5 USING AN EXISTING SURVEY AS A FRAME

In many countries, a given type of survey may be conducted on a regular basis and provide detailed information on the units of one of the target populations. In these cases, it is good practice to leverage the survey and use it as a frame for the sample selection.

This topic is developed below, by considering such a survey as a frame.

Sampling

Let us suppose that one of the Q frames, denoted with $\overline{q}$, contains the records of a survey on households, and let $A_I^{\overline{q}}$ be the dataset including the records of this survey, in which v denotes the generic individual in dataset ($v = 1, ..., A_I^{\overline{q}}$).

In addition, let us denote with $K_I^{\overline{q}}$ the sampling weight attached to an individual *v* of the dataset. Let us also suppose that these weights enable the production of unbiased estimates for population *H,* including corrections for non-response and different factors of calibration.

The sample selection is achieved by means of the following operations:

1. **Sampling from the frame.** A sample of individuals, $\Omega_I^{\overline{q}}$, is selected from frame $A_I^{\overline{q}}$ with a generic sample design, where the individual $v \in A_I^{\overline{q}}$ is selected with an inclusion probability equal to $\pi_v^{\overline{q}}$. In most cases, $\Omega_I^{\overline{q}}$ is a cluster sample of households: all individuals of a given household are observed in the survey, and often have equal sample weight.

2. **From individuals to households.** All households u_h of individuals belonging to $\Omega_I^{\overline{q}}$ are included in sample $S_H^{\overline{q}}$. Thus, it is possible to note any changes in the households' structure that may have occurred in the period between the survey and the current observation.

3. **From households to individuals.** Sample $S_I^{\overline{q}}$ includes all individuals living in the households of sample $S_H^{\overline{q}}$.

Estimation

The estimator of θ_H is:

$$\hat{\theta}_H = \sum_{q=1}^{Q} \sum_{h=1}^{S_H^q} w_h^q \, Y_h \qquad (1.54)$$

with

$$w_h^q = \begin{cases} \dfrac{1}{m_h} \displaystyle\sum_{v=1}^{\Omega_I^q} K_v^q \, \dfrac{\lambda_{vh}^q}{\pi_v^q} & \text{if } q = \bar{q} \\ \dfrac{1}{m_h} \displaystyle\sum_{v=1}^{\Omega_I^q} \dfrac{\lambda_{vh}^q}{\pi_v^q} & \text{if } \quad q \neq \bar{q} \end{cases} \tag{1.55}$$

where $\bar{q}$ is the frame containing the records of the survey.

$$\hat{\theta}_F = \sum_{q=1}^{Q} \sum_{h=1}^{S_f^q} w_f^q \, Y_f \tag{1.56}$$

Variance

The estimator variance in Formula 1.54 is:

$$V_P(\hat{\theta}_H) = \sum_{q=1}^{Q} V_P(\hat{\theta}_{H,ht}^q). \tag{1.57}$$

The expression of variance $V_P(\hat{\theta}_{H,ht}^q)$ for $q \neq \bar{q}$ can be defined with reference to the well-known results of the two-phase sampling design (Särndal, 1993, p. 348):

$$V_P(\hat{\theta}_{H,ht}^q) = V_{1p} E_{2p}(\hat{\theta}_{H,ht}^q | \, A_I^{\bar{q}}) + E_{1p} V_{2p}(\hat{\theta}_{H,ht}^q | \, A_I^{\bar{q}}) \tag{1.58}$$

where V_{1P} and E_{1P} are the variance and the expectation of the first sampling, in which sample $A_I^{\bar{q}}$ was selected, and V_{2P} and E_{2P} are the variance and the expectation of sub-sampling from $A_I^{\bar{q}}$.

The variance $V_P(\hat{\theta}_F)$ may be obtained with expressions similar to Formulas 1.57 and 1.58 above, by substituting $\hat{\theta}_H$ and $\hat{\theta}_{H,ht}^q$ with $\hat{\theta}_F$ and $\hat{\theta}_{F,ht}^q$.

1.6 CALIBRATED ESTIMATES

Usually, there is a variety of auxiliary information for producing estimates. Benchmarking on external counts enables ensuring the overall consistency of a country's statistical system.

In the case described in this Chapter, the auxiliary information can be derived from the following sources:

- **Auxiliary information from frames** A_j^q (J=I or F ; and q=1,…,Q). Let $\mathbf{X}_v^q$ be the vector of auxiliary variables available for the record $v = 1, \dots, A_j^q$ and let

$$T^q = \sum_{v=1}^{A_j^q} \mathbf{X}_v^q$$

be the benchmarking population vector of totals.

For instance, if A_j^q is a geographic register of land parcels, covering the whole country,

  - $\mathbf{X}_v^q$ can be a scalar denoting the area of land parcel v employed for agricultural use, and
  - T^q can be a scalar indicating the total area of a country employed for agricultural use.

- **Auxiliary information on households.** Let $\mathbf{X}_h$ be the vector of auxiliary variables available for household h and let

$$\mathbf{T}_H = \sum_{h=1}^{H} \mathbf{X}_h$$

be the benchmarking population vector of the totals.

We highlight the following aspects:

- information on $\mathbf{X}_h$ is usually collected during the survey operations;
- information on $\mathbf{T}_H$ commonly derives from external sources. For example,
  - $\mathbf{T}_H$ can be extracted from demographic statistics which often publish counts by population, sex and age. In this case, vector $\mathbf{X}_h$ indicates the households' counts by sex and age.
  - The vector of total $\mathbf{T}_H$ can be estimated by means of a widespread survey on households conducted on a regular basis (e.g. a labour force survey).

- **Auxiliary information on farms.** Let $\mathbf{X}_f$ be the vector of auxiliary variables available for farm f and let

- $\mathbf{T}_F = \sum_{f=1}^{F} \mathbf{X}_F$

the benchmarking population total. We highlight the following aspects:

- information on $\mathbf{X}_f$ is usually collected during the survey operation;
- information on $\mathbf{T}_H$ is commonly derived from external sources, such as national accounts.
- In light of this amount of auxiliary information, various strategies for constructing calibrated weights can be proposed:

  a) The first strategy is to perform 3 different and separate calibration processes: (i) the first step involves Q separate calibrations for each of the Q frames available; (ii) in the second step, the calibration is performed on the totals $\mathbf{T}_H$ referred to households, starting with the weight established in the former calibration; (iii) the third process develops a calibration of the totals $\mathbf{T}_F$ with reference to households, starting with the weight established in step (i) of the process of calibration.

  b) The second strategy performs a simultaneous calibration, considering all available auxiliary information.

  c) Other mixed strategies may also be proposed.

Strategy (a) is less computer-intensive, but fails to leverage all existing information. Strategy (b), described below, may present some computational problems if applied to a very large dataset, but exploits all existing auxiliary information and ensures the coherence of the estimates to all benchmarking totals.

The calibrated weights for strategy (b) are constructed in the following steps:

1. Constructing the stack vectors of auxiliary variables derived from different sources.
2. Computing the calibrated weights for the sampled units of frames A_j^q (J=I or F ; and q=1,…,Q).
3. Computing the calibrated weights for the sampled households and farms.

These steps will be detailed below, in relation to the chain from households to farms.

Step 1: Constructing the stack vectors of auxiliary variables

The vector of auxiliary variables for record v selected in sample Ω_I^q is obtained by stacking all the vectors with the auxiliary information available.

$$\mathbf{t}_v^q = \left((\mathbf{X}_v^1)', \cdots, \lambda_{vh}^q (\mathbf{X}_v^q)', \ldots, \lambda_{vh}^q (\mathbf{X}_v^Q)', \ldots, (\mathbf{Z}_{vH}^q)', (\mathbf{Z}_{vF}^q)'\right)'. \quad (1.59)$$

in which

$$\mathbf{Z}_{vH}^q = \sum_{h=1}^{H} \frac{\lambda_{vh}^q}{m_h} \mathbf{X}_h \text{ and } \mathbf{Z}_{vF}^q = \sum_{h=1}^{H} \sum_{f=1}^{F} \frac{n_{hf}}{m_f} \frac{\lambda_{vh}^q}{m_h} \mathbf{X}_f. \quad (1.60)$$

Some components of vector $\mathbf{t}_v^q$, for the record v selected in sample Ω_I^q, will be equal to zero. In particular, denoting with $\tilde{q}$ ($\tilde{q} = 1, \ldots, Q$) a given index for specifying the frame:

- $\lambda_{vh}^q (\mathbf{X}_v^{\tilde{q}}) = \mathbf{0}$ if $\tilde{q} \neq q$
- $\mathbf{Z}_{vH}^{\tilde{q}} = \mathbf{0}$ if $\tilde{q} \neq q$
- $\mathbf{Z}_{vF}^{\tilde{q}} = \mathbf{0}$ if $\tilde{q} \neq q$,

where **0** denotes a vector **of** zeroes. The stack vectors of auxiliary benchmarking totals are given by:

$$\mathbf{T} = \left(\left(\mathbf{T}^1\right)', \cdots, \left(\mathbf{T}^q\right)', \ldots, \left(\mathbf{T}^Q\right)', (\mathbf{T}_H)', (\mathbf{T}_F)'\right)'. \qquad (1.61)$$

Step 2: Computing the calibrated weights for the frames' sampled units

The calibrated weight for unit *v* selected in sample Ω_I^q is expressed by:

$$w_{v,cal}^q = \frac{1}{\pi_v^q}\left[1 + \left(\mathbf{T} - \widehat{\mathbf{T}}\right)' \mathbf{A^{-1}}\right] \mathbf{t}_v^q \qquad (1.62)$$

where

$$\widehat{\mathbf{T}} = \sum_{q=1}^{Q}\sum_{v=1}^{\Omega_I^q} \frac{1}{\pi_v^q}\mathbf{t}_v^q \quad , \quad \mathbf{A} = \sum_{q=1}^{Q}\sum_{v=1}^{\Omega_I^q} \frac{1}{\pi_v^q}\mathbf{t}_v^q \left(\mathbf{t}_v^q\right)'. \qquad (1.63)$$

Step 3: Computing the calibrated weights for the households and farms sampled

The Unified Calibrated Multiplicity Estimator (UCME) for the totals θ_H and θ_F is given by

$$\hat{\theta}_{H,cal} = \sum_{q=1}^{Q}\sum_{h=1}^{S_H^q} w_{h,cal}^q \, Y_h \ , \qquad (1.64)$$

$$\hat{\theta}_{F,cal} = \sum_{q=1}^{Q}\sum_{h=1}^{S_F^q} w_{f,cal}^q \, Y_f \ , \qquad (1.65)$$

with

$$w_{h,cal}^q = \frac{1}{m_h}\sum_{v=1}^{\Omega_I^q} w_{v,cal}^q \ \lambda_{vh}^q \, , \qquad (1.66)$$

and

$$w_f^q = \sum_{h=1}^{S_H^q} \frac{n_{hf}}{m_f} \, w_{h,cal}^q \ . \qquad (1.67)$$

Theorem 1.9 The UCME is benchmarked with respect to the totals $\mathbf{T}_H$ and $\mathbf{T}_F$. Indeed,

$$\sum_{q=1}^{Q}\sum_{h=1}^{S_H^q} w_{h,cal}^q\,\mathbf{X}_h\; = \mathbf{T}_H \quad , \sum_{q=1}^{Q}\sum_{f=1}^{S_F^q} w_{h,cal}^q\,\mathbf{X}_f\; = \mathbf{T}_F. \qquad (1.68)$$

Proof. The following transformation holds for estimator $\hat{\theta}_{H,cal}$:

$$\sum_{q=1}^{Q}\sum_{h=1}^{S_H^q} w_{h,cal}^q(\,\mathbf{X}_h)' = \sum_{q=1}^{Q}\sum_{h=1}^{S_H^q}\frac{1}{m_h}\sum_{v=1}^{\Omega_I^q}\; w_{v,cal}^q\;\; \lambda_{vh}^q(\,\mathrm{X}_h)'$$
$$= \sum_{q=1}^{Q}\sum_{h=1}^{S_H^q}\frac{1}{m_h}\sum_{v=1}^{\Omega_I^q}\;\frac{1}{\pi_v^q}\left[1+\left(\mathbf{T}-\widehat{\mathbf{T}}\right)'\mathbf{A}^{-1}\right]\,\mathrm{t}_v^q\;\; \lambda_{vh}^q(\,\mathrm{X}_h)'.$$

Let us consider the result below

$$\sum_{h=1}^{S_H^q}\sum_{q=1}^{Q}\sum_{v=1}^{\Omega_I^q}\;\frac{1}{\pi_v^q}\mathbf{A}^{-1}\,\mathbf{t}_v^q\;\lambda_{vh}^q(\,\mathrm{X}_h)' = \mathbf{B}_H \qquad (1.69)$$

being $\mathbf{B}_H = \begin{bmatrix}\mathbf{0}_{(\widetilde{Q},dimH)}\\ \mathbf{I}_{(dimH,dimH)}\\ \mathbf{0}_{(dimF,dimH)}\end{bmatrix}$,

where $\mathbf{0}_{(\widetilde{Q},col\mathrm{X}_h)}$ is a matrix of zeroes with $\widetilde{\boldsymbol{Q}}$ rows and $dimH$ columns, $\mathbf{I}_{(colH,colH)}$ is $(dimH \times dimH)$ dimension identity matrix and $\mathbf{0}_{(dimF,dimH)}$ is a $(dimF \times dimH)$ dimension matrix of zeroes, in which $\widetilde{\boldsymbol{Q}}$ is the dimension of vector $\left(\mathbf{T^1}\right)',\cdots,\left(\mathbf{T^q}\right)',\ldots,\left(\mathbf{T^Q}\right)'$, $dimH$ is the dimension of vector $\mathbf{T_H}$ and $dimF$ is the dimension of vector $\mathbf{T_F}$. Thus, considering that

$$\sum_{q=1}^{Q}\sum_{h=1}^{S_H^q}\frac{1}{m_h}\sum_{v=1}^{\Omega_I^q}\;\frac{1}{\pi_v^q}\;\lambda_{vh}^q(\,\mathrm{X}_h)' = \sum_{q=1}^{Q}\sum_{v=1}^{\Omega_I^q}\frac{1}{\pi_v^q}\sum_{h=1}^{S_H^q}\frac{\lambda_{vh}^q}{m_h}\mathrm{X}_h = \widehat{\mathbf{T}}_H \quad ,$$

it can be derived that

$$\sum_{q=1}^{Q}\sum_{h=1}^{S_H^q} w_{h,cal}^q\,\mathbf{X}_h\; = \widehat{\mathbf{T}}_H + \mathbf{T}_H - \;\widehat{\mathbf{T}}_H.$$

The proof of $\sum_{q=1}^{Q}\sum_{f=1}^{S_F^q} w_{h,cal}^q \mathbf{X}_f = \mathbf{T}_F$, can be obtained using the results just described.

1.7. DATASETS FOR THE INTEGRATED ANALYSIS

A dataset for an integrated analysis exploits the possibility of developing cross-sectoral analyses by merging the data on households and those on farms in a single dataset, with a unique sampling weight for estimation. This will be shown in relation to the sampling chain from households to farms.

A dataset (DS) on households can be thus built:

- each DS record corresponds to a sampled household;
- the weight for the estimation is given by

$$w_h = \sum_{q=1}^{Q} w_h^q,$$

or, alternatively, the calibration weight can be used:

$$w_{h,cal} = \sum_{q=1}^{Q} w_{h,cal}^q.$$

- The variables for the households are the original ones, Y_f .
- The variables for farms are the transformed ones,

$$Z_{hF} = \sum_{q=1}^{Q}\sum_{f=1}^{S_F^q} \frac{n_{hf}}{m_f} Y_f \ .$$

2

Optimal sampling

2.1 INTRODUCTION

The sampling strategy for estimating a set of parameters of interest is defined by the couple sampling design and estimator. If a randomization approach to the inference is adopted, the design must entail a random selection scheme of the sample, in accordance with a given set of inclusion probabilities. Standard selection schemes are one-stage or multi-stage sampling designs, and simple or stratified random sampling designs. When the inclusion probabilities in the strata are homogeneous, stratified simple random sampling (SSRS) design is defined. When the probabilities are proportional to a given size measure, Probability Proportional to Size (PPS) sampling designs are implemented. In SSRS design, the set of the inclusion probabilities fix the sample size in each stratum, thus obtaining the allocation of the sample within the strata.

In the SSRS, the optimal allocation for a univariate population is well-known in sampling literature (Cochran, 1977). In the multivariate scenario, where more than one characteristic is to be measured on each sampled unit, the optimal allocation for individual characteristics do not have much practical use, unless the characteristics under study are highly correlated. This is because an allocation that is optimal for one characteristic will generally be far from optimal for others. In the real survey context, the search for the optimal allocation also depends on other relevant dimensions of the problem: the estimates must be produced at the subpopulation or domain levels. Different sets of domains can define a partition of the population, and some domains can be aggregations of other partitions. Domains that belong to a given partition of the population may cut across domains of another partition.

The literature on sampling design has devoted much attention to sample allocation, seeking to satisfy certain criteria of optimality in terms of efficiency and budget constraints. Nevertheless, the criteria established for the problem's multidimensionality leads to a definition of an allocation that loses precision, compared to the individual optimal allocation. For these reasons, the methods are sometimes referred as compromise allocation methods (Khan *et al*., 2010).

Several authors have discussed various criteria for obtaining a useable compromise allocation. Among these see Neyman (1934), Dalenius (1953), Dalenius (1957), Kokan and Khan (1967), Chatterjee (1967), Chatterjee (1968), Chromy (1987), Bethel (1989), Falorsi and Righi (2008) and Choudhry *et al*. (2012).

2.2 BASIC CONCEPTS AND ADDITIONAL NOTATION FOR THE OPTIMAL SAMPLE ALLOCATION

Let J be the overall target population in which J is either I, H, L or F. We consider the unbiased Horvitz-Thompson (H-T) estimator. In a general sampling design, the H-T estimator of the parameter $\theta_{J,v} = \sum_{j \in J} Y_{j,v}$ is given by $\hat{\theta}_{J,v} = \sum_{j \in s_J} (1/\pi_j) Y_{j,v}$, where s_J is a random sample, π_j is the inclusion probability of unit j, and $Y_{j,v}$ is the value variable for the unit $j \in J$. The variance of $\hat{\theta}_{J,v}$ is given by

$$V(\hat{\theta}_{j,v}) = E_p[\sum_{j \in s_J} \frac{Y_{j,v}}{\pi_j} - E_p(\sum_{j \in s_J} \frac{Y_{j,v}}{\pi_j})]^2, \qquad (2.1)$$

where the $E_p(\cdot)$ operator is the expectation under the random sampling design. In a general sampling design, Equation 2.1 becomes

$$V(\hat{\theta}_{j,v}) = \sum_{j \in J} (\frac{1}{\pi_j} - 1) Y_{j,v}^2 + 2 \sum_{(j<j') \in J} (\frac{\pi_{jj'} - \pi_j \pi_{j'}}{\pi_j \pi_{j'}}) Y_{j,v} Y_{j',v}, \qquad (2.2)$$

where $\pi_{jj'}$ is the second order inclusion probability of the (j, j') pair. Below, we will focus on the stratified simple random sampling design. Let J_g be the population of size N_{Jg} of stratum $g (g = 1, \ldots, G)$. The parameter of interest is reformulated as $\theta_{J,v} = \sum_g N_{Jg} \theta_{J,vg}$ where $\theta_{J,vg}$ is the total computed at stratum level. The H-T estimator is given by $\hat{\theta}_{J,vg} = \sum_g \sum_{s_{J,g}} (N_{Jg}/n_{Jg}) Y_{j,v}$, where $s_{J,g}$ indicates the random sample in stratum g, and n_{Jg} the stratum sample size.

The variance of $\hat{\theta}_{J,v}$ is given by

$$V(\hat{\theta}_{j,v}) = \sum_g (N_{Jg})^2 \sigma_{J,vg}^2 / n_{Jg} - \sum_g N_{Jg} \sigma_{J,vg}^2, \qquad (2.3)$$

in which $\sigma_{J,vg}^2$ is the variance of $Y_{j,v}$ in stratum g.

Equation 2.3 can be reformulated as

$$V(\hat{\theta}_{j,v}) = \sum_g (N_{Jg})^2 \sigma^2_{J,vg}/n_{Jg} + V_{0J,v} = \sum_g (V_{J,vg}/n_{Jg}) + V_{0J,v}\,, \quad (2.4)$$

where apparently only the first part of the variance depends on the sample size n_{Jg}. Table 2.1 below shows the basic notation when populations I, H, L or F are examined.

	Population				
Description	Generic	Individuals	Households	Land parcels	Farms
Population	J	I	H	L	F
Population Size	J	I	H	L	F
Population Size stratum	N_{Jg}	N_{Ig}	N_{Hg}	N_{Lg}	N_{Fg}
Value of the variable	$Y_{j,v}$	$Y_{i,v}$	$Y_{h,v}$	$Y_{l,v}$	$Y_{f,v}$
Stratum variance of the variable of interest	$\sigma^2_{J,vg}$	$\sigma^2_{I,vg}$	$\sigma^2_{H,vg}$	$\sigma^2_{L,vg}$	$\sigma^2_{F,vg}$
Parameter of interest	$\theta_{J,v}$	$\theta_{I,v}$	$\theta_{H,v}$	$\theta_{L,v}$	$\theta_{F,v}$
Stratum parameter	$\theta_{J,vg}$	$\theta_{I,vg}$	$\theta_{H,vg}$	$\theta_{L,vg}$	$\theta_{F,vg}$
Sample size	n_J	n_I	n_H	n_L	n_F
Stratum sample size	n_{Jg}	n_{Ig}	n_{Hg}	n_{Lg}	n_{Fg}
Estimate	$\hat{\theta}_{J,v}$	$\hat{\theta}_{I,v}$	$\hat{\theta}_{H,v}$	$\hat{\theta}_{L,v}$	$\hat{\theta}_{F,v}$
Stratum estimate	$\hat{\theta}_{J,vg}$	$\hat{\theta}_{I,vg}$	$\hat{\theta}_{H,vg}$	$\hat{\theta}_{L,vg}$	$\hat{\theta}_{F,vg}$
Stratum variance estimate depending on the sample size	$V_{J,vg}$	$V_{I,vg}$	$V_{H,vg}$	$V_{L,vg}$	$V_{F,vg}$

TABLE 2.1. Additional basic notation for the sample allocation

2.3 SAMPLE ALLOCATION AND THE OPTIMIZATION PROBLEM

A reasonable sample allocation depends on several elements that define the sampling surveys: the inferential approach, the parameters of interest, the domains of interest, the sampling design and the estimator. This chapter focuses on a standard scenario in which totals or proportions must be estimated in the randomization (or design-based) framework by means of the H-T estimator for different types of domains. We refer mainly to an integrated sampling framework. The sampling design will be shared by two different sampling schemes. The first is a direct sampling design, while the second is based on an indirect sampling design (Lavallèe, 2007).

The optimization problem for the sample allocation is based on two pillars: the accuracy of the estimates, and the costs of carrying out the survey. Greater costs enable a larger sample to be sampled, thus reducing the variances of the estimates and viceversa. Hence, the optimal solution must address a trade-off between variance and budget constraints.

The concept of budget constraints defines a cost function for implementing the survey. Let us assume the following linear cost function: $C = \sum_g c_{J,g} n_{Jg} + C_0$, where $c_{J,g}$ denotes the per unit cost of measuring the variables in stratum g, and C_0 is a fixed cost. Given the variance Expression 2.3 above and the total cost C, we face a dually constrained optimization problem: fixing the variances by minimizing the costs, or fixing the costs by minimizing the variances.

In the following pages, we gradually introduce the optimization problem. First, the allocation methods applied to the direct samples are shown; next, the approaches covering direct and indirect samples are proposed. In the step-by-step introduction, we assume one parameter of interest for one domain, and then, more than one domain of interest are considered in the sample allocation; the allocation eventually takes into account several parameters for several domains. The SSRS sampling design and PPS design are discussed.

As far as direct sampling is concerned, Sections 2.3.1, 2.3.2 and 2.3.3 cover the sample allocation for the SSRS design. With multi-stage sampling designs, the sample allocation becomes more complex. Section 2.3.4 shows rational and practical allocation strategies applicable when two-stage sampling or cluster sampling designs are carried out.

When different inclusion probability designs are implemented, the determination of sample size is an elaborate task, due to the complexity of Formula 2.2. Section 2.3.5 addresses the optimimization problem for PPS design. Section 2.4 introduces the sample allocation with an integrated sampling framework. The approaches proposed consider direct and indirect sampling designs to perform a global sample allocation for direct and indirect samples. In particular, Section 2.4.1 assumes a Master Sampling Frame in which units of two or more populations are linked according to a deterministic record linkage process, while Section 2.4.2 assumes a probabilistic record linkage between the units. Section 2.4.3 describes the optimal sample allocation of two non-linked populations with common variables in their sampling frame. The approach avoids the use of probabilistic record linkage, and exploits the transitivity properties of indirect sampling design.

Finally, Section 2.4.5 provides some suggestions on the optimal allocation of direct samples, taking into account also indirect samples when neither a Master Sampling Frame nor transitivity properties can be used.

2.3.1 UNIVARIATE AND UNIDOMAIN CASE

Let us suppose that the target of the survey is a total for one domain, e.g. the Country total. Two allocation approaches may be used. Having established a variance threshold $\overline{V}_{J,v}$ for the variance in Equation 2.3, the cost C is minimized. The optimal sample size, denoted as the precision-constrained optimal allocation, is given by

$$n_J = \sum_g (W_g \sqrt{\sigma^2_{J,vg}} / \sqrt{c_{J,g}}) \frac{\sum_g W_g \sqrt{\sigma^2_{J,vg} c_{J,g}}}{\overline{V}_{J,v} + N_J^{-1} \sum_g W_g \sigma^2_{J,vg}}, \quad (2.5)$$

W_g being a measure of the importance of stratum g, and usually $W_g = J_g/J$ with

$$n_{Jg} = (W_g \sqrt{\sigma^2_{J,vg}} / \sqrt{c_{J,g}}) \frac{\sum_g W_g \sqrt{\sigma^2_{J,vg} c_{J,g}}}{\overline{V}_{J,v} + N_J^{-1} \sum_g W_g \sigma^2_{J,vg}} \quad (2.6)$$

The converse problem of fixing the cost and minimizing the variance, denoted as cost-constrained optimal allocation, leads to

$$n_J = (C - C_0) \frac{\sum_g W_g \sqrt{\sigma^2_{J,vg}} / \sqrt{c_{J,g}}}{\sum_g W_g \sqrt{\sigma^2_{J,vg} c_{J,g}}}, \quad (2.7)$$

with

$$n_{Jg} = (C - C_0) \frac{W_g \sqrt{\sigma^2_{J,vg}} / \sqrt{c_{J,g}}}{\sum_g W_g \sqrt{\sigma^2_{J,vg} c_{J,g}}}. \quad (2.8)$$

Usually, sizes (2.6) and (2.8) are not integers; they are rounded up to the next integer.

Specific choices of the parameters of the problem define the well-known sample allocations. In Equation 2.7, when $c_{J,g} = c_J$, the Neyman allocation is achieved, with $n_{Jg} = N_{Jg} \sqrt{\sigma^2_{J,vg}} / (\sum_g N_{Jg} \sqrt{\sigma^2_{J,vg}})$. Furthermore, when $\sigma^2_{J,vg} = \sigma^2_{J,v}$ and $c_{J,g} = c_J$, the proportional (to population size) sample allocation is achieved, $n_{Jg} = N_{Jg} / sum_g N_{Jg}$. Finally, adding the condition $W_g = W$, the equal sample allocation $n_{Jg} = n_J / G$ is performed.

Remark 2.1: let us note that the optimization problem requires the $\sigma^2_{J,vg}$. However, this is unknown and, in practice, the sampling allocation must use proxies or estimates of the variances of interest, denoted as $\hat{\sigma}^2_{J,vg}$. These can be obtained in several ways. Usually, they are estimated by means of (i) data of previous surveys on the same or similar population; (ii) considering the variance of an auxiliary variable (of administrative source) that is highly correlated with the variable of interest; (iii) the results of a pilot survey; (iv) guesswork on the structure of the population, with the assistance of some mathematical results. In general, they may be assumed as model-based estimates.

2.3.2 THE UNIVARIATE AND MULTIDOMAIN CASES

The Neyman sample allocation is optimal, as it minimizes the variance of the subject to $\sum_g n_{Jg} = n_J$. However, the Neyman allocation may cause certain strata to suffer great variation, because the strata are not domains of estimation. If we consider the $\theta_{J,vg}$ too as parameters of estimate, the Neyman allocation is no longer optimal and a better rough sample allocation can be the equal sample allocation. Equal allocation is efficient for estimating stratum (or specific domain) parameters, but it may lead to a much larger variance of the estimator, compared to that of the Neyman allocation for the overall population. In many practical applications, a compromise allocation can be performed. When the Neyman allocation coincides with the proportional allocation, the compromise allocation with fixed costs assumes the Expression

$$n_{Jg}^{C} = \alpha \; n_J (N_{Jg}/J) + (1-\alpha)(n_J/G) \qquad 0 \le \alpha \le 1. \qquad (2.9)$$

Costa *et al*. (2004) describe some issues that may arise with this type of allocations. Bankier (1988) proposes a compromise allocation denoted as a power allocation. When each stratum is a domain of interest, the power allocation is given by

$$n_{Jg}^{B} = n_J \frac{\sqrt{\sigma_{J,vg}^2}\,(N_{Jg}/\theta_{J,vg})W_g^t}{\sum_g \sqrt{\sigma_{J,vg}^2}\,(N_{Jg}/\theta_{J,vg})W_g^t} \qquad (2.10)$$

where t is a tuning constant. The $\theta_{J,vg}$ being unknown, a proxy must be used. The allocation in Formula 2.10 is obtained by minimizing $\sum_g [W_g^t \sqrt{V(\hat{\theta}_{j,v})/\theta_{J,vg}^2}\,]^2$. The choice $t = 1$ and $W_g = \hat{\theta}_{J,vg}$ defines the Neyman allocation.

Chromy (1987), Bethel (1989) and Choudhry *et al*. (2012) give a mathematical formalization to the compromise allocation, according to an optimization problem. We introduce the method of determining the strata sample sizes subject to specified reliability requirements, concerning both the domain and population estimates (precision-constrained optimal allocation). Usually the cost-constrained approach is better for a National Statistical Institute. However, shifting the focus on precision, obtaining the costs, and iterating the process until a good compromise between estimate precision and costs is reached, can be interesting. On the other hand, the cost-constrained optimal allocation often leads to definition of a sample size without a critical analysis of the statistical output (see Chromy, 1987).

The problem is formulated as follows:

$$\begin{cases} \min \sum_g c_{J,g} n_{Jg} \\ \sum_g V_{J,vg}/n_{Jg} + V_{0J,v} \le \overline{V}_{J,vg} \quad \forall g = 1,\ldots,G \\ 0 < n_{Jg}/N_{Jg} \le 1 \\ n_{Jg} \ge 2 \qquad\qquad (when N_{Jg} \ge 2), \end{cases} \qquad (2.11)$$

in which $\overline{V}_{J,vg}$ is the fixed threshold of variance tolerance for the estimate $\hat{\theta}_{J,vg}$. The optimization problem is solved by treating the variance constraints as equality, and applying the Lagrange multipliers approach (or Cauchy Schwarz solution). The final solution is $n_{Jg} = \sqrt{\phi V_{J,vg}/c_{J,g}}$, with $\phi = \sum_g [\sqrt{V_{J,vg} c_{J,g}}/\overline{V}_{J,vg}]^2$, and ϕ being the Lagrange multiplier.

2.3.3 MULTIVARIATE AND MULTIDOMAIN CASE

In large-scale surveys, when a sample allocation must be performed, it is realistic to take into account more than one target parameter for different sets of domains estimated. A straightforword generalization of the univariate case (see Section 2.3.2) is achieved if the concept of domain-specific variables is introduced. Let us assume that the totals $\theta_{J,v}$ of V variables must be estimated for D domains. The domains are strata or aggregations of strata (planned domains). For each variable of interest, let us consider the domain-specific study variable Y_q, with q being a pair of (v,d), and $d = 1,\ldots,D$. The domain-specific value variable $Y_{j,q}$ is defined as

$$Y_{j,q} = \begin{cases} Y_{j,v}, & \text{if unit } u_j \in J_d \\ 0, & \text{otherwise} \end{cases} \qquad (2.12)$$

in which J_d denotes the d th domain. Let us indicate $\theta_{J,q} = \sum_j Y_{j,q}$. The optimization problem (see Equation 2.11) in the multivariate and multidomain cases becomes

$$\begin{cases} \min \sum_g c_{J,g} n_{Jg} \\ \sum_g V_{J,qg}/n_{Jg} + V_{0J,q} \le \overline{V}_{J,qg} \quad \forall q = 1,\ldots,V \times D \\ 0 < n_{Jg}/N_{Jg} \le 1 \\ n_{Jg} \ge 2 \qquad \text{(when } N_{Jg} \ge 2\text{)}, \end{cases} \qquad (2.13)$$

with $\overline{V}_{J,qg} = \sum_g (N_{Jg})^2 \sigma^2_{J,qg}/n_{Jg}$, and where $\sigma^2_{J,qg}$ is the variance of the domain-specific variable Y_q in stratum g.

Remark 2.2: Several algorithms proposed in the literature solve Problem 2.13. Among these, Chromy (1987) developed an algorithm that is suitable for automated spreadsheets. The algorithm is implemented by means of the MAUSS-R package (which may be downloaded from the Istat website: www.istat.it). Bethel (1989) proposed an iterative method and also illustrated the convergence. More recently, Choudhry *et al*. (2012) noted that the objctive function is a convex separable function. They deal with the problem as Non-Linear Programming, and use the SAS proc NLP with the N-R option.

2.3.4 OPTIMAL SAMPLE ALLOCATION IN MULTI-STAGE SAMPLING DESIGN

When the sample selection follows a two-stage or multi-stage sampling design, the optimal sample allocation becomes more complex, because the variance expression is more intricate and the number of final units could be obtained at random (one-stage cluster design).

In two-stage or multi-stage sampling design, the optimal allocation depends on the type of inclusion probabilities. When Primary Sample Units (PSUs) are not homogeneous in terms of size (the number of final elementary sampling units), it may be more efficient to select the PSU having inclusion probabilities that are proportional to size. Another possibility is to define a PSU-stratified design, in which the strata are correlated with PSU size (stratification encompasses the size classes). Thus, a PSU SSRS design is implemented.

Here, we investigate the multi-stage or cluster sampling designs when the target population is the elementary unit population. In Section 2.4.1, a sort of cluster design is introduced, in which the cluster population and the elementary unit population are both integrated target populations.

Below, for the sake of brevity, we consider a two-stage sampling design with a homogeneous PSU.

Let M be the cardinality of the PSU population. The number of elementary units of the i th PSU is denoted as $N_i \cong \overline{N}$, where $\sum_i N_i = J$. A two-stage sample is selected, with m as the number of sampled PSU and $\overline{n}$ the sample size of elementary units belonging to the PSU u_i.

The following conditions are set up:

- m/M is small;
- $\overline{n}/\overline{N}$ is small.

Under these conditions,

$$V_{2ST}(\hat{\theta}_{J,v}) \cong \frac{\widetilde{V}_{J,v}}{m\overline{n}} k[1+\rho_v(\overline{n}-1)] \qquad (2.14)$$

(Valliant *et al*., 2013) with

$$\widetilde{V}_{J,v} = [J^2\sigma^2_{J,v}] \qquad (2.15)$$

with $\sigma^2_{J,v} = \sum_i \sum_{j\in u_i} [Y_{ij,v} - (1/J)\sum_i \sum_{j\in u_i} Y_{ij,v}]^2/(J-1)$ as the population variance; $k = [V(B_v)+V(W_v)]/\widetilde{V}_{J,v}$;

$$V(B_v) = \frac{\sum_{i}^{M}(Y_{i,v} - \frac{1}{M}\sum_{i}^{M}Y_{i,v})^2}{M-1} \qquad (2.16)$$

with $Y_{i,v} = \sum_{j=1}^{N_i} Y_{ij,v}$, where $Y_{ij,v}$ is the value of variable Y_v in unit $j \in u_i$ and

$$V(W_v) = \frac{M\sum_{i} N_i^2 \sum_{j=1}^{N_i}(Y_{ij,v} - (1/N_i)\sum_{j=1}^{N_i}Y_{ij,v})^2}{N_i - 1} \qquad (2.17)$$

Finally,

$$\rho_v = \frac{V(B_v)}{V(B_v) + V(W_v)}. \qquad (2.18)$$

A population of homogeneous PSUs is not usual in reality; rather, skewed population distributions are found. For example, the distribution of farms may be defined by a large number of very small farms, and a limited number of very large farms. The PSU stratification, with strata correlated with the PSU size enables definition, at the stratum level, of the condition that $N_i \cong \overline{N}_g$ in most strata, and the sample in each PSU is fixed to $\overline{n}_g$.

In this case, Expressions 2.14, 2.15, 2.16, 2.17 and 2.18 are computed for each stratum g.

Expression 2.14 or the stratified version $V_{2STssrs}(\hat{\theta}_{J,vg})$ are interesting, because $\widetilde{V}_{J,v}/[m\overline{n}]$ or $\widetilde{V}_{J,vg}/[m\overline{n}g]$, where $\widetilde{V}_{J,vg} = [J_g^2\sigma_{J,vg}^2]$ is the variance of the estimated total in a simple random sampling or SSRS design without replacement, respectively of size $n_J = m\overline{n}$ and $n_{J,g} = m_g\overline{n}_g$, and when the sampling fractions are small.

Then, terms $k[1+\rho_v(\overline{n}-1)]$ or $k_g[1+\rho_{vg}(\overline{n}_g-1)]$ (for the stratified design) represent the design effect. A practical approach to achieve an optimal allocation accounts for Equation 2.13 inflating the stratum variances through the design effect. In particular, for the stratified version, the application of the optimization problem 2.13 holds when the estimates computed on the elementary units are defined on domains defined by aggregation of the PSU strata.

In case of two-stage SSRS of PSUs, conditions (a) and (b) are generally not completely fulfilled, especially when the population has a skewed distribution. For example, in the

strata with smaller farms, the ratio $\overline{n}_g/\overline{N}_g$ cannot be small. On the right tail of the distribution, m_g/M_g cannot be small, with M_g and m_g being, respectively, the PSU stratum population and sample size.

In developed countries, several farms are individual or have two or three workers; in these cases, $n_i = N_i$. For strata with small PSUs, the variance is approximated by the variance of a simple cluster sampling design:

$$V_{CLstr}(\hat{\theta}_{J,vg}) \cong \frac{J_g^2 \sigma_{J,vg}^2}{m_g \overline{n}_g}[1+\rho_{CLvg}(\overline{n}_g - 1)] \qquad (2.19)$$

where $\rho_{CLvg} = (V(B_{vg}) - \sigma_{J,vg}^2)/[(\overline{n}_g - 1)\sigma_{J,vg}^2]$ (Cochran, 1977).

When the m_g/M_g small condition is not fulfilled, an efficient sampling strategy can certainly include all the PSUs, and a one-stage sampling design is performed.

In farm population examples, this means that large farms are certainly selected in the sample, and PSUs are referred to as self-representative units.

A practical approach for defining the optimal sample size of elementary units is to use the system established by Formula 2.13 to inflate the variance of SSRS design, $V(\hat{\theta}_{J,vg}) = (J_g^2 \sigma_{J,vg}^2)/(m_g \overline{n}_g)$, according to ρ_{CLvg}.

2.3.5 OPTIMAL SAMPLE SIZE DETERMINATION, WITH VARYING INCLUSION PROBABILITY SAMPLING DESIGNS

A direct sample can be selected in accordance with a set of inclusion probabilities π, varying among population units. The literature distinguishes between PPSs with replacement sampling and those without replacement, denoted by πPS. Generally, this type of design can be efficient for estimating totals or means when the population sizes are known, but does not increase accuracy when non-linear functions such as ratio parameters must be estimated.

When a survey planner opts for a varying inclusion probability sampling design, he/she bears in mind a superpopulation model ξ in describing the variable of interest.

Let us consider a general linear superpopulation model,

$$\begin{cases} Y_{j,v} = \mathbf{X}'_j \beta + \varepsilon_{j,v} \\ V_\xi(\varepsilon_{j,v}) = \varsigma_{j,v}^2 \\ E_\xi(\varepsilon_{j,v}, \varepsilon_{j',v}) = 0 \quad j \neq j', \end{cases} \qquad (2.20)$$

where $\boldsymbol{X}_j$ is a vector of auxiliary variables, and β defines a vector of regression slopes of the same dimension of $\boldsymbol{X}_j$. Let us consider a general estimator $\hat{\theta}_{J,v}$, unbiased under model ξ and probability sampling design $p(\cdot)$, $E_\xi E_p(\hat{\theta}_{J,v} - \theta_{J,v}) = 0$. Godambe and Joshi (1965) showed that

$$E_\xi E_p(\hat{\theta}_{J,v} - \theta_{J,v})^2 \geq \sum_{j \in J}(\frac{1}{\pi_j} - 1)\varsigma_{j,v}^2. \qquad (2.21)$$

with the expectation both under the model and sampling design $E_\xi E_p(\cdot)^2$ for unbiased estimators denoted as Anticipated Variance (Isaki and Fuller, 1982). Expression 2.21 provides the lower bound of the variance of $\hat{\theta}_{J,v}$.

Let the sampling design be such that the expected sample size is n_J, and let the parameter $\hat{\theta}_{J,v}$ be estimated by the regression estimator (Särndal *et al*., 1992). An optimal design for which Anticipated Variance is minimized is one in which the inclusion probabilities are determined by

$$\pi_j = \pi_{j,opt} = \frac{n_J \varsigma_{j,v}}{\sum_{j \in J} \varsigma_{j,v}}, \qquad (2.22)$$

in which it is assumed that $\pi_{j,opt} \leq 1$ for all $j \in J$. With $\pi_{j,opt}$, the Anticipated Variance is equal to $\sum_J [(1/\pi_j) - 1]\varsigma_{j,v}^2$.

In agricultural surveys, PPS or πPS designs are particularly useful for farm populations. A superpopulation model that reasonably fits the farm population has a variance of $V_\xi(Y_{j,v}) = \varsigma^2 \boldsymbol{k}' f(\boldsymbol{X}_j)$, $\boldsymbol{k}$ being a vector of the same dimensions of $\boldsymbol{X}$ and $f(\cdot)$ being a function. For example, a suitable model for farm quantitative variables uses $V_\xi(Y_{f,v}) = \varsigma x_f^\gamma$, with $x_f = N_f$ (the number of workers), and γ as a power. Generally, the terms $\varsigma, \boldsymbol{k}$ and γ are unknown, and must be estimated. In case $\gamma = 1$, $\pi_{f,opt} = n_F \sqrt{N_f} / [\sum_{f \in F} \sqrt{N_f}]$.

Therefore, the optimal size determination must establish the sample size n_J, using the inclusion probabilities $\pi_j \propto \varsigma_{j,v}$.

If a πPS design is adopted, as is commonly the case, the computation of the optimal sample size must deal with a general expression of the variance given in Formula 2.2, where the second-order inclusion probabilities are involved. The problem can be overcome readily, by approximating the variance expression with the variance of the PPS design. Having established the variance threshold $\overline{V}_{J,v}$, the sample size is

$$n_j = \frac{V_{1,v}}{N_j^2 \overline{V}_{J,v}}, \qquad (2.23)$$

where $V_{1,v} = \sum_{j \in J} p_j \left(\frac{Y_{j,v}}{p_j} - \theta_{J,v}\right)^2$ and $p_j = \varsigma_{j,v} / \sum_J \varsigma_{j,v}$.

Remark 2.3: The π_{opt} depends on the unknown variables $\varsigma_v, \boldsymbol{k}$ and γ. The final remark of Section 2.3.1 suggests a way to estimate ς_v. Model analysis enables definition of the heteroschedastic terms $\boldsymbol{k}$ and γ: these can be estimated by using data obtained previously or data from a pilot survey. Usually, γ varies in the interval $[0,2]$ (Särndal *et al*., 1992). Let us consider the variance $\hat{V}_\xi(\hat{\theta}_{J,v}) = \varsigma_v^2 X_j^\gamma$, where γ can be obtained by means of an iterative procedure. Starting from a homoschedastic regression model where $\varsigma_{j,v} = \varsigma_v$ the model is fit by ordinary least squares, and the residuals calculated. We use the squared residual $\hat{\varepsilon}_j^2$ as an approximate estimate of $V_\xi(Y_{j,v})$. Then, the equation $\hat{V}_\xi(\hat{\theta}_{J,v}) = \varsigma_v^2 X_j^\gamma$ is established and the regression $log(\hat{\varepsilon}_j^2) = \gamma log(X_j)$ is estimated, where log is the natural logarithm. The estimate of the regression slope is an estimate of the power. The procedure is iterated until convergence.

Remark 2.4: the determination of an optimal sample size defined by Formulas 2.22 and 2.23 is univariate and unidomain. Therefore, it may be inefficient for certain estimates, especially in the case of parameters joined with variables not correlated to the variable leading to the $\pi_{j,opt}$.

2.4 OPTIMAL SAMPLE ALLOCATION IN MASTER SAMPLE FRAMES

A Master Sample Frame (MSF) for Agriculture defines a multidimensional population involving the farms, land parcels, households and individuals (supported by agricultural activities). The target populations are distinctive due to the phenomena of interest. These can be classified as economic, environmental and social dimensions of a country. The MSF is based on the conceptual framework requiring linkage between the four populations (FAO, 2010). By definition, the MSF includes the Integrated Data Set (IDS). In the IDS, the relationship between the units of two populations can be of cluster-elementary units type (one-to-many linkage), or with partial overlapping between units (many-to-many linkage). The presence of these links enables definition of the Integrated Survey Framework (ISF). The ISF is based on sample selections from the four target populations, and the links between the units enable analysis of the data across surveys. The MSF's added value is the availability of information on the links between units of different but connected populations.

A reasoned sample allocation in the ISF must take into account the sampling process globally. The following observational strategy is adopted. The following two-step scheme is required if four populations are involved:

- the first step consists of choosing one out of the couples of populations $I-H$ (Individuals - Households) or $L-F$ (Land parcels – Farms). Then, the data on the units of the populations of the chosen couple are collected;

- in the second step, the units of the populations not observed in the first steps are reached through indirect sampling, using the links with the units in the sample selected in the first step.

If populations $I-H$ are observed first, then populations $L-F$ are observed in the second step. In this second step, data are collected on all farms where the individuals observed in the first step work. For these farms, information related to L is also collected.

If $L-F$ are observed first, then populations $I-H$ are observed in the second step. In particular, in this second step data are collected on all households with workers who work in the farms observed in the first step. In these households, information related to I is also collected.

Note that the first step represents an ISF, with either $I-H$ or $L-F$ being two integrated populations.

Assume that the observation starts from H; all individuals in the households selected are observed. The observational strategy is a cluster sampling design. The PSU inclusion probabilities are established by the designer. The design can be simple/stratified or PPS/multivariate compromised.

When the observation starts from I, if the individual i is observed, then the household u_h is observed, such that $i \in u_h$. The variables $Y_{h,v}$ are collected in u_h. Since the household variables are an aggregate of the variables measured at an individual level, the observational process is such that all individuals of the u_h selected are observed too. Then, a cluster sampling is performed, but unlike in the previous observational strategy, the design is not simple/stratified nor PPS/multivariate compromised.

Although the two observational strategies adopt a cluster sampling design, Section 2.4.1 shows that the set of inclusion probabilities of the u_h is different.

The same conclusions are reached when the first step is perfomed on the $L-F$ pair. We are implementing a PPS cluster sampling design, where the farms are the PSUs.

Intermediate ISF contexts are also possible. An interesting example considers the ISF only between I and F. In this case, the direct and indirect samples correspond to a cluster sample.

2.4.1 OPTIMAL SAMPLE ALLOCATION IN CASE OF DETERMINISTIC RECORD-LINKED MASTER SAMPLING FRAMES

In this Section, we consider a general example of a MSF involving populations F and H, with individuals working on farms, and with linkage matrix Λ^{HF} available. The farm and household samples are selected to study the economic dimension of agriculture within a given geographic area. Without any loss of generality, let the parameters of the overall populations be the parameters of interest. Besides $\theta_{F,v}$ and $\theta_{H,v}$ the presence of links opens the way to estimate the interrelationship between the economic and social dimensions, such as between farm income and household well-being, or between farm structure and household income. For the sake of simplicity, let the ratios measured over the two different populations be the target parameters $\theta_{FH,vv'} = \sum_f Y_{f,v} / \sum_h Y_{h,v'}$.

The sample allocation techniques seek to jointly select a sample of farms and a sample of households, such that the samples are optimal for the estimates of populations F and H and the integrated parameters. Here, we propose a simple strategy adopting the indirect sampling approach (Lavallèe, 2007; see also Chapter 1, Part 4 of this publication). The Generalised Weight Share Method (GWSM) for the estimation process completes the sampling strategy (Lavallèe and Caron, 2001; Lavallèe, 2007).

The basic assumption for preserving the estimates from bias is that each household is linked to at least one farm. We show the case of a direct sample drawn in accordance with the SSRS design.

Example A - From households to individuals: the observational strategy coincides with a cluster sampling design. We focus on the allocation of the households. Note that in Section 2.3.4, the sample allocation concerns the elementary units, i.e. the individuals. To set up the optimization problem, the Z variables are built:

$$Z_{h,v} = \sum_i \tilde{\lambda}_{hi} Y_{i,v}, \qquad (2.24)$$

with $\tilde{\lambda}_{hi}$ being the element of the standardized link matrix $\tilde{\Lambda}^{HI}$, where

$$\tilde{\lambda}_{hi} = \begin{cases} 1, & \text{when } i \text{ is linked with } u_h \\ 0 & \text{otherwise} \end{cases}.$$

Then, $Z_{h,v} = \sum_{i \in u_h} Y_{i,v}$, and using the variance of the Z s in the system (Equations 2.13 or 2.27), we obtain the households' sample size. The optimization problem must be refined by a suitable cost function.

We assume a general cost function $C = C_0 + \sum_g \sum_h \widetilde{C}_{1h,g}$, where the overall cost to observe the h th PSU is $\widetilde{C}_{1h,g} = C_{1H,g} + C_{2hI,g}$, with $C_{1H,g}$ being the constant cost per PSU in stratum g, and $C_{2hI,g}$ the overall cost of observing the elementary units in the h th PSU. A simple expression for $C_{2hI,g}$ is

$$C_{2hI,g} = c_{i,g} n_h, \tag{2.25}$$

with $c_{i,g}$ being the cost per element i in unit u_h, being constant for all households in stratum g. Any monotone non-decreasing function can describe $C_{2hI,g}$. An interesting cost function assumes that the average cost of observing the elementary units decreases when n_i increases. For instance, the average interview cost of n_i individuals in a large PSU household is lower than the average interview cost for a small household. The cost function set out below is an example of this type of cost function:

$$C_{2hI,g} = log[c_{i,g} n_h]. \tag{2.26}$$

Here, $log(\cdot)$ is the natural logarithm. Jessen (1942) added the cost of traveling between PSUs, $\widetilde{C}_0$, to the cost function. The complete cost function is therefore

$$C = C_0 + \sum_g [\widetilde{C}_0 \sqrt{n_{H,g}} + C_{1H,g} n_{H,g} + \sum_h C_{2hI,g}]. \tag{2.27}$$

It is assumed that the travel cost between PSUs varies, approximately, as the square root of the number of PSUs.

Using Equations 2.24 or 2.25, Equation 2.26 gives varying costs per PSU; this is not suitable for Equation 2.13. Therefore, in practice, in Equations 2.24 or 2.25, term n_h is replaced with $\overline{n}_{h,g}$, the average size of the households in stratum g. In this case, $C_{2hI,g} = \overline{C}_{H,g}$ and it can be included in $C_{1H,g}$.

When System 2.22 is used, multivariate optimization enables taking into account the cost per specific PSU. Furthermore, the optimal allocation when PSUs are selected with varying inclusion probabilities is also discussed. However, the dimension of the optimization problems must be carefully controlled.

Example B - From individuals to households: to set up the optimization problem, the Z variables are established as

$$Z_{i,v} = \sum_h \widetilde{\lambda}_{ih} Y_{h,v}, \tag{2.28}$$

with $\widetilde{\lambda}_{ih}$ being the element of the standardized link matrix $\widetilde{\Lambda}^{IH}$, where

$$\widetilde{\lambda}_{ih} = \begin{cases} 1 & \text{when } u_h \text{is linked with } i \\ 0 & \text{otherwise} \end{cases}.$$

and n_h is the number of individuals in u_h. Then, $Z_{i,v} = Y_{h,v}/n_h$ with $i \in u_h$. When the population I is directly sampled, the simple or SSRS designs are implemented.

Deville and Lavallèe (2006) shows that the variance of $\hat{\theta}_{H,v}$ is given by

$$V(\hat{\theta}_{H,v}) = \sum_g (V^Z_{I,vg}/n_{I,g}) + V^Z_{0I,v} = V(\hat{\theta}^Z_{I,v}) \tag{2.29}$$

where $V^Z_{I,vg} = (N_{I,g})^2[\sigma^Z_{I,vg}]^2$, and

$$[\sigma^Z_{I,vg}]^2 = \left\{ \sum_{f \in g} [Z_{i,v} - (1/N_{I,g} \sum_{i \in g} Z_{i,v})]^2 \right\} / (N_{I,g} - 1);$$
$$V^Z_{0I,v} = \sum_g N_{I,g}[\sigma^Z_{I,vg}]^2 \text{ and } \hat{\theta}^Z_{I,v} = \sum_g \sum_{i \in s_{ig}} (N_{I,g}/n_{I,g}) Z_{i,v},$$

where s_{ig} is the farm sample in stratum g.

Given the Z variables, the optimization problem is formalized according to System 2.13, where the variance inequalities of the estimates of $\hat{\theta}_{H,v}$ use $V^Z_{I,vg}$. A general cost function is the following:

$$C = C_0 + \sum_g [\widetilde{C}_0 \sqrt{n_{I,g}} + C_{1I,g} n_{I,g} + \sum_h C_{2hI,g}]. \tag{2.30}$$

with $C_{1I,g} = c_{i,g}$ being the constant cost per individual in stratum g, $C_{2iH,g}$ the overall cost of observing the u_h, such that $i \in u_h$. We may set $C_{2iH,g} = C_{1I,g} n_h$. As for Example A using System 2.13, we can approximate $C_{2iH,g} \cong \overline{C}_{I,g} = c_{i,g} \overline{n}_{h,g}$.

Example C - From farms to individuals: see Example A, replacing population H with F.

Example D - From individuals to farms: see Example B, replacing population H with F.

Example E - From households to farms: to set up the optimization problem, the Z variables are established as:

$$Z_{h,v} = \sum_f \widetilde{\lambda}_{hf} Y_{f,v},$$

and $\widetilde{\lambda}_{hf}$ is the element of the standardized link matrix $\widetilde{\Lambda}^{HF}$, where

$$\widetilde{\lambda}_{hf} = \begin{cases} n_{hf}/n_f, & \text{when } u_h \text{ is linked with } u_f \\ 0 & \text{otherwise} \end{cases}$$

and n_f is the number of workers in u_f.

The variance of $\hat{\theta}_{F,v}$ is

$$V(\hat{\theta}_{F,v}) = \sum_g (V^Z_{H,vg}/n_{H,g}) + V^Z_{0H,v} = V(\hat{\theta}^Z_{H,v}) \qquad (2.31)$$

where $V^Z_{H,vg} = (N_{H,g})^2[\sigma^Z_{H,vg}]^2$, and

$$[\sigma^Z_{H,vg}]^2 = \left\{ \sum_{h \in g} [Z_{h,v} - (1/N_{H,g} \sum_{h \in g} Z_{h,v})]^2 \right\} / (N_{H,g} - 1);$$
$$V^Z_{0H,v} = \sum_g N_{H,g}[\sigma^Z_{H,vg}]^2 \text{ and } \hat{\theta}_{H,v} = \sum_g \sum_{h \in s_{hg}} (N_{H,g}/n_{H,g}) Z_{h,v},$$

in which s_{hg} is the household sample in stratum g.

Given the Z variables, the optimization problem is defined according to System 2.13, where the variance inequalities of the estimates of $\theta_{F,v}$ are based on $V^Z_{H,vg}$.

The cost function is

$$C = C_0 + \sum_g [\widetilde{C}_0 \sqrt{n_{H,g}} + C_{1H,g} n_{H,g} + \sum_h C_{2hF,g}]. \qquad (2.32)$$

In this case, we can set $C_{2hF,g} = c_f n^F_h$, with c_f being the cost per farm and n^F_h the number of farms where the workers of u_h work.

Example F - From farms to households: to set up the optimization problem, the Z variables are built as

$$Z_{f,v} = \sum_h \widetilde{\lambda}_{fh} Y_{h,v}, \qquad (2.33)$$

$\tilde{\lambda}_{fh}$ being the element of the standardized link matrix $\tilde{\Lambda}^{FH}$, where

$$\tilde{\lambda}_{fh} = \begin{cases} n_{hf}/n_{hF} & \text{when } u_f \text{ is linked with } u_h \\ 0 & \text{otherwise} \end{cases}$$

and n_{hf} is the number of workers in $u_h \cap u_f$ and n_{hF} the number of workers in u_h.

Deville and Lavallèe (2006) show that the variance of $\hat{\theta}_{H,v}$ is given by

$$V(\hat{\theta}_{H,v}) = \sum_g (V^Z_{F,vg}/n_{F,g}) + V^Z_{0F,v} = V(\hat{\theta}^Z_{F,v}) \qquad (2.34)$$

where $V^Z_{F,vg} = (N_{F,g})^2[\sigma^Z_{F,vg}]^2$, and

$$[\sigma^Z_{F,vg}]^2 = \left\{\sum_{f\in g}[Z_{f,v} - (1/N_{F,g}\sum_{f\in g} Z_{f,v})]^2\right\}/(N_{F,g}-1);$$
$$V^Z_{0F,v} = \sum_g N_{F,g}[\sigma^Z_{F,vg}]^2 \text{ and } \hat{\theta}_{F,v} = \sum_g \sum_{f\in s_{fg}} (N_{F,g}/n_{F,g}) Z_{f,v},$$

where s_{fg} is the farm sample in stratum g.

Given the Z variables, the optimization problem is formalized according to System 2.13, where the variance inequalities of the estimates of $\hat{\theta}_{H,v}$ use $V^Z_{F,vg}$.

The cost function is

$$C = C_0 + \sum_g [\tilde{C}_0 \sqrt{n_{F,g}} + C_{1F,g} n_{F,g} + \sum_h C_{2fH,g}]. \qquad (2.35)$$

In this case, we may set $C_{2fH,g} = c_h n^H_f$, with c_H being the cost per household and n^H_f the number of households with workers working in u_f.

Remark 2.5: Deville and Lavallèe (2006) and Lavallèe and Labelle-Blanchet (2013) propose the use of optimal weighted links that reduce the variance of an indirectly-investigated variable. The linkage weights are optimal for one parameter of estimate, and their use can be attractive when tackling a univariate indirect sample allocation.

Remark 2.6: as far the integrated parameters are concerned, a Taylor linearization is used. We highlight that $\theta_{FH,vv'} = \theta_{F,v}/\theta_{H,v'}$ can be defined as a ratio between totals of the same population: $\theta_{H,v'} = \sum_f Z_{f,v'} = \theta^Z_{F,v'}$.

2.4.2 OPTIMAL SAMPLE ALLOCATION IN CASE OF PROBABILISTIC RECORD-LINKED MASTER SAMPLING FRAMES

For the four populations, we assume that some Λs matrices were not available, but that the units were linked in a probabilistic record linkage procedure (Part 3). Let the linkage between each pair of the populations be completed. The linkages between the units are not necessarily one-to-one, or one-to-many. For example, it is reasonable to expect many-to-many linkage for populations L and F or I and H, although many-to-one linkages are possible in reality. Record linkage uses a decision rule to decide whether there is a link between unit j of population J and unit j' of population J'. Once the links are established, the two populations J and J' are thus obtained, linked to each other (Lavallèe and Caron, 2001).

In several practical contexts, when estimation must be performed, the only file available is often the linked file obtained at the end of the linkage process.

However, it is necessary to plan the sample, and the linkage procedure supports construction of the IDS. When the direct sample from population J is drawn, the indirect sampling procedure on population J' selects units with real links, and not units with links defined by the decision rule.

From the point of view of sample allocation, it is important to take into account any randomness in the linkage; we will illustrate how this uncertainty can be encompassed in the optimal allocation.

Uncertainty in the linkage changes the relationship between the populations $I-H$ and $L-F$, because the linkage is no longer one-to-many (respectively, (u_h or u_f) and (i or l)), but many-to-many. Therefore, in the planning phase, we must consider a classic indirect sampling, while in the observational phase a cluster sample is drawn.

Let us assume that the configuration C matrices with linkage weights are available. The linkage weights between two units of two populations must not necessarily be between 0 and 1, but need only represent the relative likelihood that there exists a link (Lavallèe and Caron, 2001).

Below, we provide some examples to illustrate the strategy for obtaining the optimal sample allocation.

Example 2.4.2.1 – From households to individuals: the linkage process defines the C^{HI} matrix of order $(H \times I)$, with linkage weights c_{hi} for each pair (h,i). We consider the standardized linkage weights $\tilde{c}_{hi} = c_{hi} / \sum_h c_{hi}$, with $\sum_h \tilde{c}_{hi} = 1$. To set up the optimization problem, the Z^c variables are established as:

$$Z^c_{h,v} = \sum_i \tilde{c}_{hi} Y_{i,v}.$$

In the Z^c variables, the missing values $\tilde{\lambda}_{hi}$ are replaced with predicted values $\tilde{c}_{hi}$. Following the indirect sampling strategy, when the indirect sample of individuals is

drawn, each $Z_{h,v}$ will conflict with the $Z^c_{h,v}$ due to the uncertainty of the probabilistic linkage model. To properly account for the model's uncertainty, the input variance must be inflated, including the variability of $\widetilde{c}_{hi}$. To obtain the variance when a probabilistic record linkage procedure has been used, we begin with the variance $V(\hat{\theta}^{Z^c}_{H,v})$ in accordance with Expression 2.1, when the Λ^{HI} matrix is known:

$$V(\hat{\theta}^{Z}_{H,v}) = E_p[\sum_{h=1}^{n_H} \frac{Z_{h,v}}{\pi_h} - E_p(\sum_{h=1}^{n_H} \frac{Z_{h,v}}{\pi_h})]^2.$$

When we use $\widetilde{C}^{HI}$, the variance must also take into account expectations under the probabilistic linkage model $E_{plm}(\cdot)$. Therefore, the variance expression becomes

$$V(\hat{\theta}^{Z^c}_{H,v}) = E_p E_{plm}[\sum_{h=1}^{n_H} \frac{Z^c_{h,v}}{\pi_h} - E_p E_{plm}(\sum_{h=1}^{n_H} \frac{Z^c_{h,v}}{\pi_h})]^2. \qquad (2.36)$$

with Equation 2.41 being the general expression of an Anticipated Variance (see also Section 2.3.5), where the uncertainty of the predicted values $Z^c_{h,v}$ depends on the probabilistic linkage model.

We note that $E_p E_{plm}(\sum_{h=1}^{n_H} \frac{Z^c_{h,v}}{\pi_h}) = \sum_{h\in H}\sum_{i\in I} \overline{\widetilde{c}}_{hi} Y_{i,v}$, where $E_{plm}(\widetilde{c}_{hi}) = \overline{\widetilde{c}}_{hi}$. We also denote with $V_{plm}(\widetilde{c}_{hi})$ the variance under the probabilistic linkage model of variable $\widetilde{c}_{hi}$ and we assume that $E_{plm}(\widetilde{c}_{hi}\widetilde{c}_{hi'}) = \overline{\widetilde{c}}_{hi}\overline{\widetilde{c}}_{hi'}$ In this model, the variance in Equation 2.41 is equal to

$$V(\hat{\theta}^{Z^c}_{H,v}) = V(\hat{\theta}^{Z^{\bar{c}}}_{H,v}) + \sum_{h=1}^{N_H} \frac{1}{\pi_h} \sum_i V_{plm}(\widetilde{c}_{hi}) Y^2_{i,v}. \qquad (2.37)$$

$V(\hat{\theta}^{Z^{\bar{c}}}_{H,v})$ is the variance of $Z^{\bar{c}}_{h,v} = \sum_i \overline{\widetilde{c}}_{hi} Y_{i,v}$. In practice, $\overline{\widetilde{c}}_{hi}$ is unknown and $\widetilde{c}_{hi}$ must be used. The maximum likelihood variance estimation $\widetilde{c}_{hi}(1-\widetilde{c}_{hi})$ replaces $V_{plm}(\widetilde{c}_{hi})$.

The proposed measure of variability is based on the following linkage process: together with the probabilistic linkage model, the linkage weights are also defined, and the 0 or 1 links between households and individuals are chosen, with probabilities proportional to the linkage weights $\widetilde{c}_{hi}$. This can be achieved by Bernoulli trials where, for each pair (h,i), we decide to give links equal to 1 or 0 by generating a random number from uniform distribution $U(0,1)$, i.e. compared to a quantity that is proportional to the linkage weight.

Lavallèe and Caron (2001) obtain a similar result in Expression 4.16. An individual can be linked with more than one household.

In the stratified sampling, the variance of $\hat{\theta}_{H,v}^{Z^c}$ is given by

$$V(\hat{\theta}_{H,v}^{Z^c}) = \sum_g (\frac{V_{H,vg}^{Z^c} + N_{H,g} V_{h,plm}}{n_{H,g}}) + V_{0H,v}^{Z^c}$$

where $V_{H,vg}^{Z^c} = (N_{H,g})^2 [\sigma_{H,vg}^{Z^c}]^2$, and

$$[\sigma_{H,vg}^{Z^c}]^2 = \left\{ \sum_{h \in g} [Z_{h,v}^c - (1/N_{H,g} \sum_{h \in g} Z_{h,v}^c)]^2 \right\} / (N_{H,g} - 1);$$

$$V_{0H,v}^{Z^c} = \sum_g N_{H,g} [\sigma_{H,vg}^{Z^c}]^2 .$$

Example 2.4.2.2 – From individuals to households: See **Example 2.4.2.1** and reverse notation *H* with *I*.

Remark 2.7: Note that the IDSs with deterministic links (see Section 2.4) and probabilistic links potentially store the same amount of information (i.e. the links). The main difference is that with deterministic links, we have a one/zero identifying flag, while if the links derive from a probabilistic linkage process, the linkages are identified by real numbers and the frequency of zeroes is probably much lower than that observed in relation to deterministic links. This can create problems for storing the IDSs after the record linkage procedure. An efficient approach to produce an IDS similar to an MSF is to maintain a few links above a given threshold, and set the remaining linkage weights to zero. It must be considered that all indirect-sampled populations must be linked with some units of a direct-sampled population (unbiasedness), by enabling a small downward approximation of the variance (see Formula 2.36) to be obtained.

2.4.3 OPTIMAL SAMPLE ALLOCATION: USE OF THE TRANSITIVITY PROPERTY

Probabilistic record linkage is a statistical instrument which enables definition of an ISF. Let us consider the following example: there is a sampling frame of land parcels and farms, and the two datasets have an Enumeration Area (EA) code variable. The EA generally establishes administrative areas into which the land of a given country is partitioned. For the sake of brevity, we assume that EAs are aggregations of land parcels that can be cut across by farms. We can use the EA as a block variable for the linkage procedure, and use the approach shown in Section 2.4.2.

An alternative way to integrate the two datasets applies the GSWM's transitivity property (Deville and Lavallèe, 2006).

The use of indirect sampling by transitivity is as follows: a sample s_L from population L is selected and the indirect sample s_J of the EAs is identified, the j th EA being in s_J, if the link $\lambda_{lj} > 0$. Then, we identify the indirect sample of farms s_F, such that

$u_f \in s_F$ if $\lambda_{jf} > 0$. In the two indirect sampling selections, the standardized link matrices $\widetilde{\Lambda}^{LJ}$ and $\widetilde{\Lambda}^{JF}$ are used.

Deville and Lavallèe show that the product matrix $\widetilde{\Lambda}_t^{LF} = \widetilde{\Lambda}^{LJ} \times \widetilde{\Lambda}^{JF}$ is also a standardized link matrix, and can be used to compute the Z variables.

2.4.4 OPTIMAL SAMPLE ALLOCATION USING AUXILIARY INFORMATION CONCERNING THE INDIRECTLY SAMPLED POPULATION

In many real situations, data integration is not possible: the record linkage process does not lead to a good linkage, or, simply, a single frame population does not exist.

Let us consider the following example. There is a frame list of farms, and household population is to be investigated. The observational strategy will be from F to I and from I to H. The aim is to globally consider the observational strategy applied in the farm sample allocation.

In Sections 2.4 and 2.4.2, the problem is solved by including the Z variables in the optimization problem 2.13, where

$$Z_{f,v} = \sum_h \widetilde{\lambda}_{fh} Y_{h,v}.$$

In this context, the λ and $Y_{h,v}$ variables are absent, but we suppose that some auxiliary farm variables are available, such as the number of workers, and that from statistical or administrative sources some aggregate statistics related to H population may be found, e.g. the average household size, the number of workers per household or employment rate by area (by district, province, region, etc.), the rural and urban populations of individuals and/or households, etc. Finally, statistics on phenomena of interest for households can also be important.

The goal is to predict $Z_{f,v}$ through a superpopulation model based on the auxiliary variables. A pilot survey on a small sample of farms can be of use in defining the model.

The procedure is the following:

1. Using the number of workers n_f and another auxiliary variable, the model predicts the number of households having one member working in farm u_f. A very simple model ξ_1 is

$$\begin{cases} n_{Hf} = \beta n_f + \varsigma\sqrt{(n_f)} \\ V_{\xi_1}(\varsigma\sqrt{(n_f)}) = \varsigma^2 n_f \\ E_{\xi_1}(\varsigma\sqrt{(n_f)}, \varsigma\sqrt{(n_{f'})}) = 0 \quad f \neq f', \end{cases} \tag{2.38}$$

where n_{Hf} is the number of households with a member working in farm u_f. Without a pilot survey, n_{Hf} is generally unknown, and the model cannot be estimated. Aggregate statistics can provide an estimate for β.

2. Let us assume, from statistical or administrative data sources, that the number of workers in farms $n_F = \sum_f n_f$ and the number of rural households H are known. Then,

$$\hat{n}_{Hf} = n_f \frac{n_F}{H}.$$

3. Also, if an estimate for the variable of interest (or correlated variables) is known at individual or household levels, e.g. the household average $\hat{\bar{Y}}_{H,v}$, then

$$\bar{Z}_{f,v} = \hat{n}_{Hf} \hat{\bar{Y}}_{H,v},$$

is used for the optimal allocation. In this case, no model variance is taken into account, and a downward variance is used in the optimization problem.

Part 4B: Application

3

Simulation study for the assessment of the integrated sampling of different populations

3.1 INTRODUCTION

This Chapter illustrates the main results of a simulation study, performed to assess the properties of the integrated sampling strategy presented in Chapter 1.

The study was carried out with three main objectives:

1. Validate the proposed Unified Multiplicity Estimator (UME) by assessing some of its empirical properties
2. Outline empirical situations in which the proposed observational strategy is typically suitable
3. Investigate the empirical situations in which the UME is more efficient and accurate than classic direct estimators.

For this simulation, we initially consider only two target populations: the population of households and the population of farms.

The implementation phase is briefly described in Section 2, which illustrates the simulated data, the available frames and the observational strategy proposed. The study's results are given in Section 3, where the UME is compared with some classic direct estimators. Finally, in Section 4, the results of a sensitivity analysis are described, showing how the estimators can perform differently depending upon the observational context being considered.

3.2 GENERAL DESCRIPTION

To meet the objectives outlined above, the simulation study was performed with resort to an artificial small-size dataset. The results obtained were naturally affected, to a certain extent, by the observational context under consideration. Thus, a particularly useful extension of this study could consist in an evaluation of the differences in outcomes due to changes in the observational framework.

We began by simulating the values of certain variables related to the populations under consideration (number of members, number of workers and total income, for the population of households; number of individuals employed and total monetary production, for the population of farms). To simplify the computations, we introduced the hypothesis that all members of a given family work in the same farm. Therefore, a household was to be linked to only one farm, and the number of workers of a given farm could be obtained as the sum of the number of workers in all households linked to this particular holding. This assumption is clearly not necessary for the application of the observational strategy under consideration; all results obtained also hold in situations presenting more complex links.

Here, the populations studied were grouped in clusters and, in this preliminary simulation, we considered the aggregated data for each cluster, without taking into account the detailed data concerning the elemental units.

3.2.1 HOUSEHOLDS AND FARMS DATA

We considered a universe of $H = 1000$ households. For each household, a number from 1 to 10, indicating the total number of members, was simulated; lower probabilities of being extracted were assigned to the extreme values (1, 2, 9 and 10). An additional vector, the elements of which indicated the number of employed individuals of each family, was simulated from binomial distributions, where the probabilities of these distributions depended on the number of individuals in each family. Finally, a variable indicating the income of each household was generated. Therefore, the income of each household was also made dependent on the number of workers.

The size of the population of farms was fixed as equal to 200. Then, a vector of 1000 numbers from 1 to 200 was simulated. Each number was the identifier (ID) of a given farm, indicating the farm to which a given household was associated. Starting from the IDs of the households and of the farms, the link matrix Λ^{HF}, used in the indirect sampling procedure, was computed. The resulting linking matrix was also used to aggregate the values of the variable related to the population of farms. Indeed, as mentioned above, we assumed that the number of workers of a given farm is the aggregation of the number of individuals employed in all households linked to this specific holding. Finally, a variable indicating the total production of each farm was generated, hypothesizing that its values also depended on the income of the households linked.

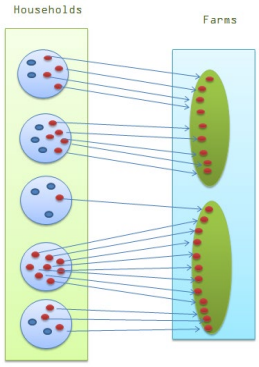

FIGURE 3.1. Relationship between the population of households and the population of farms. The number of workers of a given farm is the sum of the number of workers in all households linked with this particular farm.

Households' variables	
Number of members	Number from 1 to 10, indicating the total number of members. Simulated assigning lower probabilities to the extreme values (1,2,9 and 10).
Number of workers	Simulated from binomial distributions. The probabilities of these distributions depend on the number of people in each family.
Income	Generated from log-normal distributions. The parameters of these distributions depend on the number of workers of each household.
Farms' variables	
Number of workers	Given by the aggregation of the number of employed in all the households linked to a specific holding
Monetary Production	Generated from log-normal distributions. The parameters of these distributions depend on the total income of the linked households.

TABLE 3.1. Simulated variables for the populations of households and farms

3.2.2 OBSERVATIONAL STRATEGY

Let us consider the case in which a sample of farms and a sample of households are selected, to perform an integrated analysis on both of these populations. Moreover, suppose that the sampling frame is available only for the population of farms, while the frame for the population of households is missing or inadequate. In this context, we considered a further complication, namely the availability of two different frames for the farms – one covering the whole population and the other covering only a part of it. Thus, by selecting two samples (one from each frame), for the population of farms, a multiple frames approach was used for the population of farms to estimate the total monetary production and the UME results in the Multiplicity Estimator. Then, following the links between farms and households, two indirect samples of households were collected, computing the estimates of the total income by means of the UME; this became the GWSM estimator in the context of multiple frames, and was computed considering the multiplicity deriving from the multiple frames and from the related links.

We considered these two types of frames because multiple frames surveys may be convenient, even if a complete population list does exist. For example, in an agricultural survey, it is possible to use an area frame of farms under consideration. However, area frames are usually expensive to sample, as they require local visits and a team of interviewers for data collection. Suppose that a partial list of farms located in the same area is also available from an independent source, such as a list of email addresses or a farm register. This list would be partial and overlapping with the area frame – indeed, wholly included – and would also be cheaper to sample. Thus, an efficient sampling design should contemplate a combined design, using both lists available, under-sampling from the expensive complete frame and over-sampling from the cheaper partial one.

In this specific context, we sought to estimate the total income in the population of households and the total monetary production in the population of farms. In constructing the UME for these two population totals, we had to consider two types of unit multiplicity:

- The multiplicity related to the fact that a given farm can be contained in more than one frame;
- The multiplicity related to the fact that a given household can have more than one member working on the farm.

For the computation of the UME for farms' population, only the first type of unit multiplicity is considered; however, in computing the UME of the total income, we take both types of multiplicity into account.

3.2.3 INTEGRATED VS. INDEPENDENT DESIGNS

In relation to the population of farms, a sample is selected independently from each frame, iteratively for 30,000 simulation runs, with two different sampling designs (the simple random sampling without replacement design and the probability-proportional-to-size design).

To compare the behaviour of the UME to that of the Horvitz-Thompson estimator, we repeated the operation of selecting 30,000 samples through simple random sampling without replacement and with probability-proportional-to-size sampling, exploiting only the frame that provided a complete coverage of the target population.

For the population of households, we selected 30,000 samples, resorting to the links with the farms included in the samples collected in the first step, by means of the two sampling designs mentioned above. Thus, an indirect sampling mechanism in a context of multiple frames was used. As already mentioned, in this case the UME corresponds to the GWSM estimator, adjusted for Multiple Frames.

To compare the behaviour of the estimates of the total income obtained with the UME with the estimates obtained through the Horvitz-Thompson estimator, we again selected 30,000 samples, independent from those selected for the farms, again by resorting to a simple random sampling design without replacement and to a probability-proportional-

to-size design. The sampling designs considered for this application are summarized in Table 3.2 below.

Integrated Sampling Design: From F to H	Independent sampling design	
	F	H
SRSWOR for F IS for H	SRSWOR for F (from one frame)	SRSWOR for H (from one frame)
PPS for F IS for H	PPS for F (from one frame)	PPS for H (from one frame)

TABLE 3.2. Integrated sampling designs vs Independent sampling designs

3.3. ASSESSMENT OF RESULTS

3.3.1 ESTIMATES FOR THE POPULATIONS OF HOUSEHOLDS AND FARMS

Given the Unified Multiplicity Estimator $\hat{\theta}_J$ $(J = H, F)$, the collection of values $\{\hat{\theta}_{F,p}; p = 1, \dots, 30.000\}$ was assumed as its Monte Carlo distribution, and the empirical mean $E_{mc}(\hat{\theta}_F) = \sum_p \hat{\theta}_{F,p}/30.000$, the empirical standard deviation $SD_{mc}(\hat{\theta}_F) = \sqrt{\sum_p \frac{[\hat{\theta}_{F,p} - \theta_F]^2}{30.000}}$, the empirical coefficient of variation $CV_{mc}(\hat{\theta}_F) = SD_{mc}(\hat{\theta}_F)/|E_{mc}(\hat{\theta}_F)|$ and the square root of the empirical mean squared error $\sqrt{MSE_{mc}(\hat{\theta}_F)} = SD_{mc}(\hat{\theta}_F) + Bias(\hat{\theta}_F)$ were calculated. The same features were also computed for the empirical distributions of the estimates obtained with the H-T estimator, related to both sampling designs considered.

All the estimators considered were design-unbiased. Furthermore, with resort to the simple random sampling design without replacement, the UME efficiency was similar to that obtained with the H-T estimator; however, when using the πps sampling design, there was a significant increase in efficiency using the H-T estimator (see **Table 3.3**). This feature depended on the fact that, in this specific context, the πps design from one individual frame corresponds to the optimal sampling design.

The graphical representation of the empirical distributions of the estimates, corresponding to different estimators and sampling designs, confirms all of these considerations (see **Figure 3.2** below).

	Empirical Mean	Relative bias (%)	Empirical Standard Deviation	Empirical CV	$\sqrt{MSE}$
SRSWOR with Multiple Frames	13960555	0.04	3243567	0.23	3243571
SRSWOR with one Frame	13975581	0.14	3262373	0.23	3262435
πpr with Multiple Frames	13942271	0.09	2838810	0.2	2838841
πpr with one Frame	13956845	0.01	474822	0.03	474824
Total Monetary Production	13955459				

TABLE 3.3. Estimates of total monetary production in the population of farms

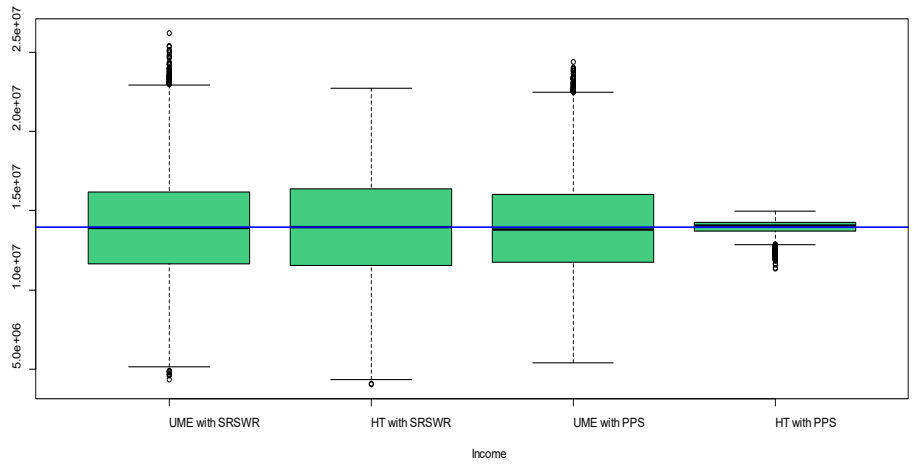

FIGURE 3.2. Empirical distributions of the estimates of total monetary production, for the population of farms

Similar considerations hold for the estimates of the total income of the population of households. In this case too, the UME, i.e. the GWSM estimator adjusted for the availability of multiple frames, is design-unbiased. Furthermore, as in the case of the estimates relating to the farms' population, the efficiency of the UME is similar to that of the H-T estimator, when a simple random sampling design without replacement is used; but in the case of the πps sampling design, there is a significant gain in terms of efficiency when the H-T estimator is used.

	Empirical Mean	Relative bias (%)	Empirical Standard Deviation	Empirical CV	$\sqrt{MSE}$
SRSWOR with Multiple Frames	13958128	0.043	3230164	0.23	3230170
SRSWOR from one Frame	13973550	0.153	3261087	0.23	3264298
πpr from Multiple Frames	13946500	0.040	3059107	0.21	3059412
πpr from one Frame	13951995	0.001	454436	0.03	454474
Total Income	13952142				

TABLE 3.4: Estimates of the total income in the population of households

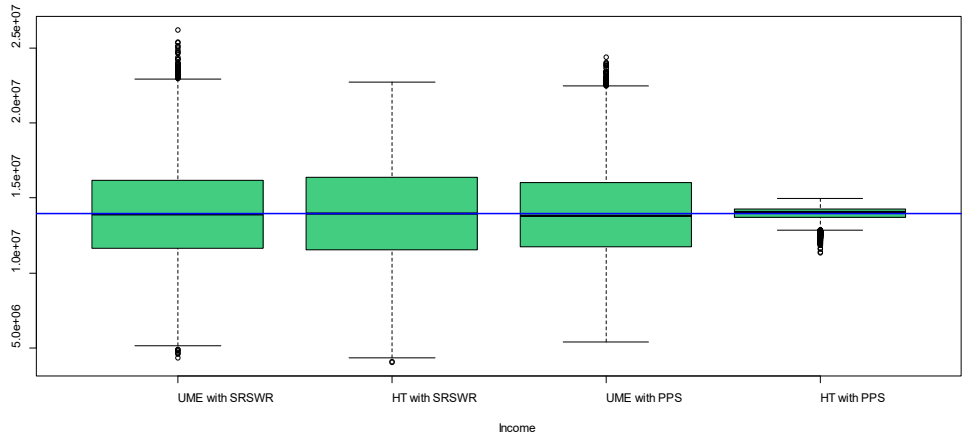

FIGURE 3.3. Empirical distributions of the estimates of total income, for the population of households

3.3.2 ESTIMATE OF A RATIO

Part 1, Chapter 4 above showed that the need to develop an integrated analysis leads to the definition of the Integrated Population and its related parameters. In particular, an integrated parameter $\theta_{JJ'}$ refers to two different target populations, J and J', and is defined as a function f of variables related to both populations. In this simulation study, the focus is on integrated parameters θ_{HF}, which refer to the populations of households and farms.

We will examine the case in which the ratio between the total income of households and the total monetary production of farms is to be estimated. As seen above, the integrated parameter is expressed as $\theta_{HF} = \frac{\theta_H}{\theta_F}$, where θ_H and θ_F are, respectively, the total income in the population of rural households, and the total monetary production of the population of farms.

Here too, the results obtained by means of the proposed integrated observational strategy were compared with those deriving from the application of a classic direct design.

Specifically, the results obtained by means of the two following observational strategies were compared:

1) The ratio was estimated by computing the ratio between the direct estimate of the total income and the direct estimate of the total monetary production, obtained from two independent samples selected with a simple random sample design without replacement.

2) The estimates for the populations of farms and households were computed with the UME. Given the two samples of farms, selected from both frames available by means of a simple random sample without replacement, the estimate of the total income for the households linked with the selected farms was calculated.

Finally, the ratio between these estimates was computed. Therefore, in this case, the samples selected for the two populations are not independent.

	Empirical Mean	Relative Bias (%)	Empirical Standard Deviation	Empirical CV	$\sqrt{MSE}$
UME	0.999739	0.002	0.0002	0.002	0.0002
HT	1.067836	6.809	0.39	0.37	0.3956
Ratio	0.9997623				

TABLE 3.5. Estimates of the ratio between the households' total income and farms' total monetary production

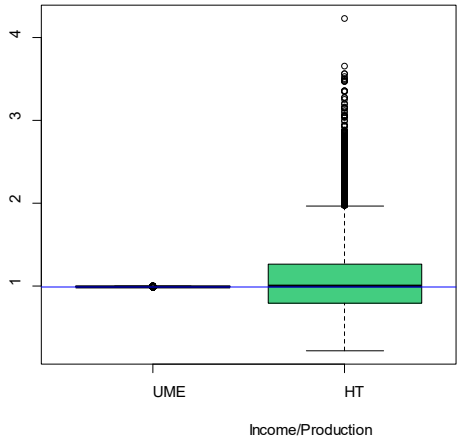

FIGURE 3.4. Empirical distributions of the estimates of the ratio between households' income and farms' monetary production

The results obtained (see **Table 3.5** and **Figure 3.4** above) illustrate that the observational strategy proposed brings a remarkable gain in terms of solution efficiency and accuracy. Indeed, in this type of situations, where the two populations are interlinked (i.e. the units of one population are also the units of the other), the observational strategy proposed enables computation of the ratio between corresponding values, i.e. totals computed from dependent samples.

3.4 A SENSITIVITY ANALYSIS

One of the main objectives of this simulation study is to identify empirical situations in which the observational strategy proposed could be more suitable than others, providing an integrated coverage of different populations and estimates with an efficiency that is at least comparable to that deriving from the application of other classic direct sampling designs.

As seen above in Section 3.3, this sampling strategy is particularly useful when it is sought to estimate a parameter related to an integrated population as a ratio between two

variables, the first related to the population of households and the second to the population of farms.

Therefore, we begin by simulating four pairs of variables (the first variable referring to the population of households and the second to that of farms), with different values for the correlation coefficient ρ, and we estimate the values of the ratio for each pair, with the two sampling designs considered in Section 3.2.

The results obtained show that, with the observational strategy proposed, instead of directly sampling from both populations and then computing the ratios between the estimated totals, there is a remarkable gain in terms of efficiency. Furthermore, the higher the absolute value of ρ, the greater is this gain (see **Table 3.6** and **Figure 3.5** below).

	Correlation coefficient	Population Ratio	Empirical Mean	Relative Bias (%)	Empirical Standard Deviation	Empirical CV	$\sqrt{MSE}$
UME	≈ 0.9	0.8297333	0.8297665	0.004	0.0026976	0.00325243	0.002698965
HT			0.8307133	0.118	0.0294558	0.03545817	0.02947188
UME	≈ 0	1.388085	1.388209	0.009	0.02031599	0.0146347	0.02031637
HT			1.389673	0.114	0.04963526	0.03571722	0.04966066
UME	≈ -0.9	-1.713101	-1.71321	0.006	0.01322255	0.007717995	0.01322299
HT			-1.715351	0.131	0.06535055	0.03809747	0.06538929
UME	≈ 0.5	0.9631746	0.9631894	0.001	0.007926945	0.008229892	0.007926958
HT			0.964336	0.12	0.03380664	0.03505691	0.03382659

TABLE 3.6. Estimates of the ratio between variables related to different populations

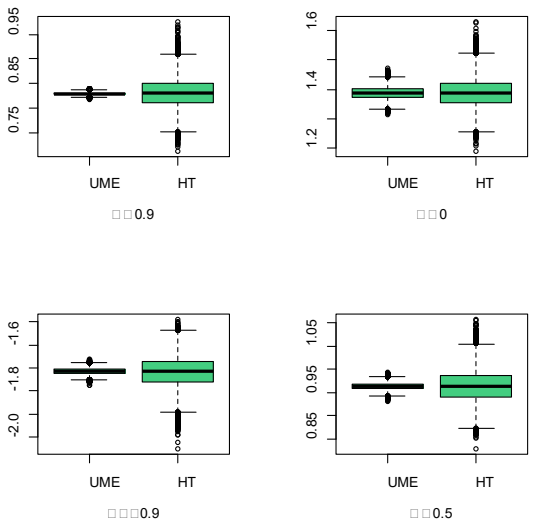

FIGURE 3.4: Empirical distributions of the estimates of the ratio, with different values of the correlation coefficient

3.5 CONCLUDING REMARKS

This study has described some of the main strengths of the proposed integrated sampling technique, which are, briefly, the following:

- the Unified Multiplicity Estimator is design-unbiased;
- the loss in terms of efficiency when estimating a population total is often negligible;
- when it is sought to estimate an integrated parameter, the observational strategy proposed provides a more accurate and efficient solution and the efficiency of the estimator increases with the absolute value of the correlation coefficient ρ.

In addition, from a computational point of view, a great advantage of this method lies in the fact that the estimation weights are calculated with resort only to the inclusion probabilities of the units selected in the first direct sample, thus overcoming the need to know the inclusion probabilities of all units of all target populations considered.

4

Experiments on optimal sampling

This Chapter shows four applications based on real survey data relating to Districts 7, 8, and 9 of Gaza Province, Mozambique. Two integrated populations are taken into account: one of individuals (heads of households) and one of farms. This simulation study used the MAUSS-R software, implemented in R language (the software can be downloaded from the following website: http://www.istat.it/it/strumenti/metodi-e-software/software/mauss-rdownload; to access the user-friendly interface, JAVA is required). The software produces an optimal sample allocation, in accordance with the optimization problem defined by Equation 2.13.

4.1 DESCRIPTION OF THE ARTIFICIAL POPULATION AND ESTABLISHMENT OF THE IT PROCEDURE

The first population's sampling frame is the household census database (for 2007), hereinafter denoted I : the database records concern individuals (heads of households) involved in agricultural, fishing, or forestry activities. The database's original dimensions are of about 54,000 records, and includes several socio-demographic, enviromental, and economic variables. Table 4.2 below lists the variables used in our experiments.

The second database gathers information on large and medium farm censuses, and features a sample of the small farm survey (for 2009), hereinafter denoted as F . Records are kept on about 890 farms and enviromental and economic variables are included (see Table 4.1).

We merged the two databases, creating an artificial link between individuals and farms. The exercise does not seek to predict the actual links between the two populations, but attempts to define a realistic ISF context. The merging procedure exploited the following variables: for individuals, job type and residence district; for farms, the sector, district and the number of persons employed by type. Before merging, the I frame was cleaned, and the records which did not feature a job type variable were removed (these amounted to about 9,000 records).

Variable type	Individuals
Socio-demographic	Gender and age of head of household, Rural/Urban household
Environmental	Residence District (7, 8, 9), Enumeration Area
Economic	Cattle, Pigs and Small Ruminants, Trees, Fishing (yes/no), Type of Employment (worker in private or public/governmental farm, worker in family farms, farmholder, farmholder without employed persons, etc.)
	Farms
Environmental	District (7, 8, 9), Enumeration Area
Economic	Cattle, Pigs and Small Ruminants, Poultry, Employment by type (number of workers, family workers, etc.) Farm Sector (private or public/governmental)

TABLE 4.1. The simulation's variables

Subsequently, we note that about 36,000 records within I declare themselves to be farmholders without employed persons. In these cases, we define a one-to-one farm-individual link. The remaining individuals were linked to the 890 farms according to the following hierarchical rules:

- when possible, each farm is linked to a number of individuals equal to the number of workers, according to their employment type;
- first, individuals and farms in the same district are linked. Several residents of District 9 are linked to farms in the other two districts, because the number of workers of the farms in District 9 is lower than the number of individuals residing in the same district;
- individuals are linked to private/public governmental farms when the *type of employment* and *farm sector* are compatible.

The software for the optimal sample allocation must define two specific input datasets. The first dataset includes $N_{F,g}$, variances $\sqrt{\sigma^2_{F,qg}}$ and $\sqrt{\sigma^Z_{I,qg}}$ (the allocation procedure considers the domain-specific variables), and the estimates of $Y_{q,g}$ and $Z_{q,g}$. These are the estimation targets, and can be obtained from the ISF database. Finally, the dataset includes the variable cost per farm; this must be constant throughout the stratum.

The second dataset establishes the precision thresholds, in terms of the coefficient of variations for estimation parameters at domain level.

4.2 SIMULATION 1: COMPARISON OF SOME DIRECT SAMPLE ALLOCATION METHODS

The first simulation investigated the sample allocation method for Population I.

The sampling design is SSRSWOR, with 91 given by districts. The variables of interest are: the number of cattle, of pigs and small ruminants, and of trees, and whether or not

fishing activities are performed. Table 4.2 below summarizes the statistics of the variables driving the sample allocation.

Statistics	Cattle	Pigs and small ruminants	Trees	Fishing(yes/no)
Median	0.0	1.0	12.0	-
Mean	1.5	3.1	30.0	5.1%
CV	451.1%	208.6%	226.3%	433.1

TABLE 4.2. Summary statistics of the variables of interest for Population *I*

The allocation was performed to obtain about 3,500 units in the sample; the cost per unit is equal to 1. We then performed a cost-constrained allocation. We considered the Neyman allocation for each of the four variables, the Proportional-to-Stratum-Population-Size (PSPS) allocation and the Multivariate – Multidomain (MM) optimal sample allocation given by System 2.13. We fixed the CV thresholds of the estimates of the totals for the four variables, so that the sample size is about 3,500. The domains of the estimates are the provinces (all districts) and each district individually. Figures 4.1 and 4.2 below illustrate the expected CV with the sample allocation obtained.

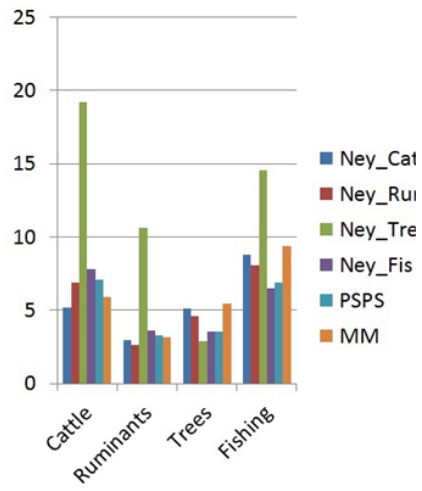

Figure 4.1: *Expected CV(%) at province level, with sample size of 3,500 units*

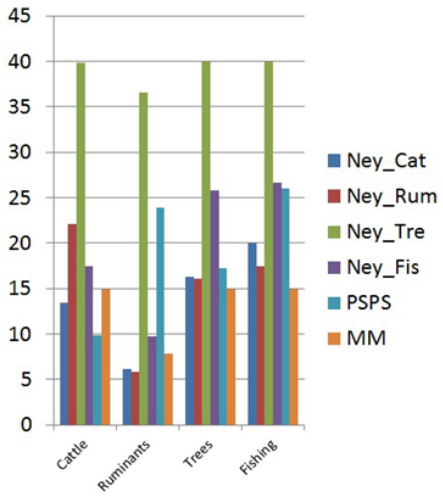

Figure 4.2: *Maximum expected CV(%) at district level, with sample size of 3,500 units*

The evidence indicates that at the province level, a Neyman allocation for the cattle (Ney_Cat) appears to be the best method. The MM method also performs well. As far district estimates are concerned, only the MM allocation yields estimates with CVs lower than 15%.

4.3 SIMULATION 2: COMPARISON BETWEEN SAMPLE SIZE ALLOCATIONS, WITH AND WITHOUT CONSIDERATION OF THE INDIRECT SAMPLED POPULATION

This simulation focuses on the MM method, and two scenarios are compared. The first scenario does not assume an ISF. A sample of farms is directly planned in the MM

method, and the expected CV of the indirect sample of individuals is computed. The second scenario assumes a sample allocation that properly takes into account the direct sample of farms and the indirect sample of individuals (see Section 2.4.1).

For the farms, we consider an SSRS design. The strata are defined as districts by size class (1, 2, 3-4, 5-9, 10-19, 20-49, 50-99, 100+), such that 21 strata are obtained.

The parameters of interest concern both farms and individuals. For the farms, we consider cattle, small ruminants and pigs, and poultry; for individuals, cattle, small ruminants and pigs, trees, and fishing.

In the first scenario with a non-Integrated Observation (IO), at province and district level, the CV thresholds are fixed only for the farm parameters.

In the second scenario with IO, the thresholds for the CV referred to individual estimates are fixed at 10%, at province level, and 15% at district level. The cost per farm is fixed as equal to 1. Figures 4.3 and 4.4 below show the expected CV. The analysis focuses on the performance of the sample allocation with respect to the estimates for individuals. In the two scenarios, the farm sample size is of about 4,000 records. In the first scenario, the expected CV for the farm estimates is, of course, lower than the expected CV observed when the allocation is based on the second scenario. This is because in the second allocation, a part of the farm sample is required to define the indirect sample of individuals (Simulation 3 studies this inefficiency issue in detail). The individual sample size expected is of about 5,300 records.

Figure 4.3 below shows that in the two scenarios, the CVs are similar. However, in the district domains, the CVs of Scenario I are extremely high (Figure 4.4), exceeding 30% in two cases. The integrated approach to allocation enables the CVs to be controlled.

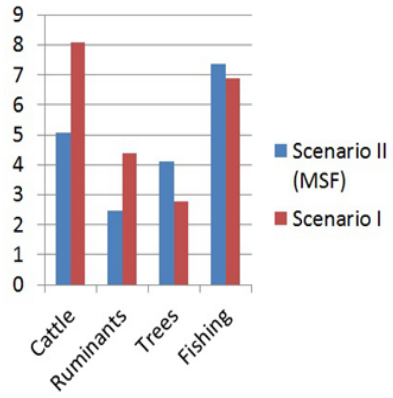

Figure 4.3: *Expected CV(%) of individul province estimates in Scenario I (no IO) and II (IO)*

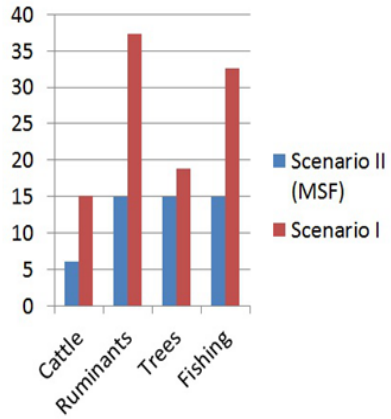

Figure 4.4: *Maximum expected CV(%) individual district estimates in Scenario I (no Io) and II (IO)*

4.4 SIMULATION 3: COMPARISON BETWEEN TWO INDEPENDENT SAMPLE ALLOCATIONS AND AN INTEGRATED SAMPLE ALLOCATION

This simulation assumes that the sampling frames of F and I are available, and that it is possible to build an MSF. From a statistical point of view, approaching estimation with an ISF presents several potential advantages. However, the costs of the approach must also be considered.

The simulation focused on two observational strategies: the first planned two independent samples, one for farms and one for individuals. Therefore, a truly integrated analysis could not be performed. The second observational strategy applied an integrated sampling design that drew a direct sample of farms and an indirect sample of individuals (the workers of the farms sampled).

For both direct sampling designs, the stratification was given by district and by size classes. Then, Population I was stratified according to the size of the farm in which the relevant individuals work. The target parameters are, for farms, the cattle, small ruminants and pigs, and poultry; for individuals, the small ruminants and pigs, trees and fishing. The simulation establishes different CV thresholds and costs for farms and individuals respectively.

Experiment A – High CV thresholds for farms

In this experiment, the CV thresholds for farms are higher than the CVs for individual estimates (Table 4.3 below).

	Farms			Individuals		
Domains	Cattle	Pigs and small ruminants	Poultry	Pigs and small ruminants	Trees	Fishing
Province	20%	20%	20%	10%	10%	10%
District	25%	25%	25%	15%	15%	15%

TABLE 4.3. CV thresholds for Experiment A.

Furthermore, interview costs were established. If individuals were drawn with a independent sample, the cost was ($c_{i,g} = c_i = 1$). If the individuals were observed by means of an indirect sample, the overall cost of interviewing the farms' workers is given by Equations 2.29 or 2.30. In this case, $c_{i,g} = c_i$. Moreover, we replace n_f with the stratum average size $\overline{n}_{f,g}$.

When two independent samples are planned, the different costs do not affect the sample sizes (precision-constrained optimal allocation). The sample size of farms is 1,551, while the number of individuals within the sample is 4,042. In the integrated sample allocation, the costs do affect the allocation, essentially because the number of sampled farms

decreases as the farm interview costs increase. To maintain the individuals' sample size, the allocation increases the inclusion probability of the larger farms.

Table 4.4 below shows the sample sizes of farms and the expected sample sizes of individuals when Expression 2.29 is used to calculate the costs of individual interviews in the integrated allocation. We can see that the farm sample is more than double the sample size, considering farms alone (1,551). The increase in size is due to precision constraints on the individuals.

Cost per farm interview	1	2	5	10
Farms	3,495	3,357	3,191	3,098
Individuals	6,083.9	6,103.1	6,714.8	7,355.8

TABLE 4.4. Sample sizes for the integrated sample allocation, when the overall individual costs are given by Equation 2.29

Table 4.5 below shows the allocation when Equation 2.30 is used for the cost of individual interviews in the integrated allocation.

Cost per farm interview	1	2	5	10
Farms	3,603	3,046	3,036	3,035
Individuals	7,757.8	7,911.8	8,109.2	8,213.3

TABLE 4.5. Sample sizes for the integrated sample allocation, when the overall individual costs are given by Equation 2.30

Figures 4.5 and 4.6 below show that in terms of the overall costs of performing the surveys, the integrated observational strategy is generally more expensive, except when the cost per farm interview is equal to 1 and the cost of the interviews of farm workers is given by Equation 2.30.

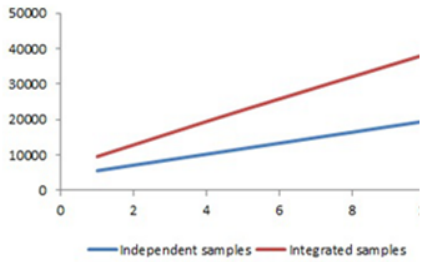

Figure 4.5: *Overall costs - integrated vs two independent allocation using the (2.29)*

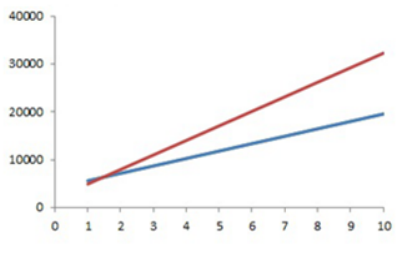

Figure 4.6: *Overall costs - integrated vs two independent allocation using the (2.30)*

Experiment B - Low CV thresholds for farms.

The above example emphasizes that in the integrated allocation, the farm sample size is very large due to the thresholds for the individual estimates, which require a large sample, thus leading to the definition of a large farm sample.

In this experiment, we set lower CV thresholds for the farms (Table 4.6 below).

	Farms			Individuals		
Domains	Cattle	Pigs and small ruminants	Poultry	Pigs and small ruminants	Trees	Fishing
Province	7%	7%	7%	10%	10%	10%
District	15%	15%	15%	15%	15%	15%

TABLE 4.6. CV thresholds, Experiment B

With these CVs, when the allocation is performed independently, the farm sample size is equal to 5,283 units. Table 4.7 below illustrates the integrated allocation with the cost function given by Equation 2.29.

Cost per farm interview	1	2	5	10
Farms	5,559	5,463	5,374	5,353
Individuals	7,810.9	7,868.7	7,870.8	8,272.1

TABLE 4.7. Sample sizes for the integrated sample allocation, when the overall individual costs are given by Equation 2.29

Table 4.8 below illustrates the allocation when Equation 2.30 is used for the cost of individual interviews in the integrated allocation.

Cost per farm interview	1	2	5	10
Farms	5,324	5,319	5,318	5,319
Individuals	8,967.4	8,976.4	9,034.7	9,065.3

TABLE 4.8. Sample sizes for the integrated sample allocation, when the overall individual costs are given by Equation 2.30

In this case, the performance of the integrated sample allocation and of the two independent allocations are similar.

Observing the overall costs reached (cost per farm inteview and cost per individual interview), the integrated approach is viable when the cost function represented by Equation 2.30 is used.

Figures 4.7 and 4.8 below show that the two approaches are rather similar.

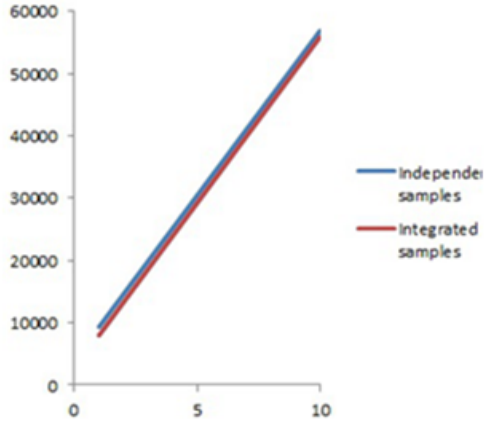

Figure 4.7: *Overall costs - integrated vs two independent allocation using the (2.30)*

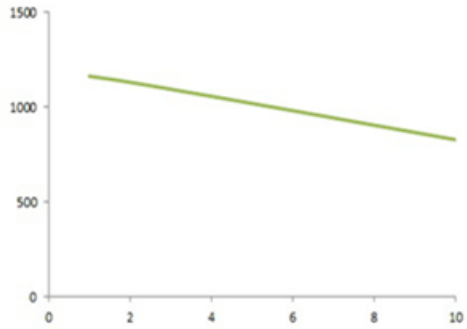

Figure 4.8: *Overall costs - the difference between integrated vs two independent allocation using the (2.30)*

4.5 SIMULATION 4: OPTIMAL INTEGRATED SAMPLE ALLOCATION WITH A PROBABILISTIC RECORD LINKAGE MSF

This simulation measures the increase in sample size arising when the IO is obtained by means of a probabilistic record linkage procedure. Expression 2.42 is innovative, and the simulation provides empirical proof that the expression holds. A Monte Carlo simulation is performed to verify whether the expected CV given by the allocation is confirmed by the CV computed on 500 Monte Carlo estimates.

The definition of the IO by means of probabilistic linkage begins with the previous MSF. Now, for each individual employed in a farm (excluding farmholders without employed persons) a probabilistic linkage was given. The probability was given by a random value extracted from the uniform distribution $\tilde{c}_{fi} = U(0,1) + 0.3$. When $\tilde{c}_{fi} > 1$, it was fixed as equal to 1. The individual is then linked with another farm f' (selected randomly), with a standardized linkage weight equal to $\tilde{c}_{f'i} = 1 - \tilde{c}_{fi}$. For individuals who are not farmholders without employed persons, the linkage weight $\tilde{c}_{fi}$ average was 0.85. For the farms, we considered an SSRSWOR design. The strata were defined as districts by size class (1, 2, 3-4, 5-9, 10-19, 20-49, 50-99, 100+); thus, 21 strata were obtained.

The parameters of interest at the province and district levels are for farms, cattle, small ruminants and pigs, and poultry, and for individuals, cattle and fishing. Table 4.9 below shows the CV thresholds. Note that the allocation constraints are rather similar to those proposed in Experiment B (see Section 4.5).

	Farms			Individuals	
Domains	Cattle	Pigs and small ruminants	Poultry	Pigs and small ruminants	Fishing
Province	7%	7%	7%	10%	10%
District	15%	15%	15%	15%	15%

TABLE 4.9. CV thresholds with the probabilistic record linkage MSF

We fix a constant cost for farm interviews. To implement the sample allocation, the correct input stratum variance must be used. According to Formula 2.42, we replace the stratum population variance with the expression

$$\sqrt{[\sigma^{Z^c}_{F,qg}]^2 + (1/N_{F,g})[\tilde{c}_{fi}(1-\tilde{c}_{fi})]}.$$

The optimal sample size allocation is equal to 6,117 farms, an increase from the optimal sample allocation obtained with a deterministic linkage (Experiment B, Section 4.5).

We will now investigate the relationship between the expected CVs, indicated by the solution to the optimization problem, and the empirical Monte Carlo CVs computed on 500 samples. For the sake of brevity, we will focus on the estimates at the province level for the total of pigs and small ruminants, and for fishing, referred to the I population.

We emphasize that the CV thresholds are 10%, but the allocation procedure defines a sample size that guarantees an expected CV equal to 6.38% for pigs and small ruminants, and equal to 6.00% for fishing. The empirical CVs of the Monte Carlo simulation are, respectively, 6.24% for pigs and small ruminants and 5.62% for fishing, confirming that Expression 2.42 correctly indicates variability.

PART 5

Integrated estimation combining different sources of information

Part 5 focuses on the integrated estimation strategies that can be used to combine different sources of information (surveys or registers). The ultimate aim is to improve the efficiency of the estimates currently produced, while at the same time ensuring that the estimates computed within national statistical systems are consistent. The set of methods proposed enables integrated statistics to be produced even when surveys are not carried out in an integrated manner. Indeed, data produced by both industrialized and developing countries are usually collected by different government authorities, each carrying out statistical surveys, on the basis of a stovepipe model. This implies that different surveys are carried out independently of each other, with the application of separate sampling designs and different specific estimation techniques. Therefore, the stovepipe model cannot guarantee that the estimates of the same target variables in different surveys will be consistent. Furthermore, the estimates' efficiency is not optimized. To solve these problems, a set of methods are proposed and applied in a very general framework, in which the information processed is retrieved from registers and from two or more sample surveys. The goal is to obtain accurate and coherent estimates, using the information available in an integrated and efficient manner.

Typically, the information on target variables is available only for a small sample, while auxiliary information may be observed in other data sources. Under a design-based framework, a projection estimator and a repeated weighting estimator are proposed. The former method enables the complete information available to be boosted, from a small sample to a larger sample or to a register, to obtain accurate domain estimates; the latter guarantees coherence between different sets of estimates, through the calibration property of the regression estimator.

In the context of domain estimation, small area methods are considered. These are presented as the model-based counterpart of the projection estimator, derived under a model-assisted paradigm. In particular, model-based unit-level predictors are examined. The projection estimator, proposed by Kim and Rao (2012), is an asymptotically-unbiased model-assisted estimator that combines information from different sources, using common unit-level auxiliary information. A working model is fitted to the units of a smaller sample, and synthetic values are then obtained for the units of a larger sample or, if available, for the units of the register.

The corresponding projection estimator, derived by means of the model-based approach, is also defined and evaluated; model-based synthetic values are obtained for the larger survey or for the register. In this context, the goal is to obtain synthetic values by means of the standard linear mixed model, and of unweighted or weighted model parameter estimation. We show that the model-based projection estimator is equivalent to EBLUP and pseudo-EBLUP predictors, in which the known totals or means of covariates, necessary for computing the estimates, are obtained with larger survey or register data. Consequently, model-based projection estimates and MSEs coincide with the corresponding expressions from EBLUP and pseudo-EBLUP procedures. The correspondence between the model-based projection estimator and the EBLUP and pseudo-EBLUP predictors is due to the linearity of the working model.

A case study on the data from the 2007 Mozambique Population Census is provided. As for the design-based methods, both projection and repeated weighting estimators demonstrate good performance against quality indicators (AARE and ASE). Moreover, both methods enable more precise estimates to be obtained, which in turn allows the CVs of the estimates to be reduced with respect to direct estimators (H-T and GREG). The projection method displays a better performance, because the predictive capabilities of the working model improve.

For a more widespread application of the projection method, it is advisable to establish an IDS, as this may improve the quality of the estimates. This may be done in a master sample frame that links the units of each population of interest, using a unique identification key variable, or more generally, using linkage methods. The repeated weighting method enables both coherence among estimates and their improved efficiency. When the working model's predictive poweris high, model-based projection estimators outperform the corresponding design-based projection estimator. Otherwise, when the model's predictive power is mild or poor, the design-based projection displays a better performance. Under poor predictive models, the H-T and GREG estimators both perform better than model-based estimators. Thus, in terms of model failures, the design-based projection estimator is more robust.

Part 5A: Theory

1

Estimation methods for combining different sources of information

1.1 INTRODUCTION

This Report focuses on integrated estimation strategies that can be used to combine information from different sources – gathered from surveys or stored in registers. The ultimate aim is to improve the efficiency of the estimates currently produced, while also seeking to ensure consistency among the estimates computed by National Statistical Systems (NSS). The set of methods proposed enables integrated statistics to be produced even when the surveys are not performed in an integrated manner.

Indeed, NSS data produced by both industrialized and developing countries are usually collected by different government authorities, each carrying out statistical surveys, on the basis of a stovepipe (SP) model. This implies that different surveys are carried out independently from one another, with the application of separate sampling designs and different specific estimation techniques. Consequently, for each survey, the final estimates are based on a specific set of weights associated to the corresponding units. This method of elaborating data does not guarantee that the estimates of the same target variables will be consistent in different surveys, thus making it difficult to measure and analyze the impact of policies from one sector to another. Also, in light of the inconsistency generated by the lack of coherence between surveys, the stovepipe approach does not enable similar information observed in other surveys to be exploited; therefore, the accuracy of estimates may be lower than in studies where available information is used in an integrated manner.

For example, in agricultural statistical surveys, data on crop and livestock production are often drawn from separate samples. The separate data sets do not provide an adequate basis for an in-depth analysis of the characteristics of farms that produce both crops and livestock, or that specialize in one or the other. The conduction of household surveys is

rarely coordinated with farm surveys, and farm surveys generally focus on economic aspects, with very little information on socio-economic features relating to agricultural and non-farm activities. Moreover, the small dimensions of sample sizes do not enable more disaggregated estimates to be computed by cross-classifying the statistic units with reference to important structural information, e.g. classifying farms as rural or non-rural.

Furthermore, agricultural statistics are often compiled by different institutions (some surveys are carried out by local authorities and/or donor agencies) that are not part of the national statistical system. Consequently, few data sets contain integrated information on socio-demographic and economic variables. For example, both poverty and agricultural information could be useful in analyzing the relationships between factors, but sometimes, estimates for the same set of indicators, for the same country and for the same reference period, may differ dramatically; this generates confusion among users and hinders the formulation of policies for an adequate allocation of funds.

Estimates are traditionally computed by using separate sets of weights for each survey, and performing separate calibration processes (Deville and Särndal, 1992); however, this method can be affected by several types of inefficiency. Indeed, this approach guarantees internal coherence between estimates within the same survey, but coherence across different surveys is not guaranteed. Recently, Knotterus *et al*. (2003), Knotterus and Coen (2006) and Särndal and Traat (2010) have proposed weighting techniques that enable coherent and unbiased estimates to be produced from different surveys on the same populations by using adjusted sampling weights. For the target population and the main sub-populations of interest, the estimates produced are efficient and design-consistent. The method is based on the notion of calibrating or re-calibrating the sampling weights of a given sample S_2, for which the target variable Y is observed with respect to a set of population totals estimated from a sample S_1 of a larger survey. This implies that S_1 and S_2 must have a common set of variables, which will be called C-variables, to be used in the calibration step. In the same framework, Kim and Rao (2012) proposed another method that enables the information actually available to be increased. This method, called the projection estimator, is a model-assisted technique based on the generation of synthetic values, $\widetilde{Y}$, of the target variable Y on the sampling units of S_1. Bearing this aim in mind, a model on S_2 is fitted using C-variables as covariates. Then, the final estimates are calculated by applying S_1 sampling weights to the $\widetilde{Y}$ values. Both samples S_1 and S_2 may derive from a single survey, or may be obtained by joining the samples of two or more surveys as appropriate. Furthermore, the set of weights of S_1 and S_2 may be calibrated on a set of common variables, the A-variables, which can be retrieved from a register R. All of these techniques produce valid inferences under design-based or model-assisted approaches. However, there are some other methods, within the model-based paradigm, that enable efficient and coherent estimates to be produced. Elbers *et al*. (2003), and Kim and Rao (2012), proposed a model-based type of projection estimator. In this scenario, small area estimation methods (see Rao, 2003) can be useful when small domain sample sizes are observed in S_2. In addition, these provide estimates even when no sampling units are observed in one or more small domains. Small area model-based methods are obtained by means of Empirical Best Linear Unbiased Predictors (EBLUP), based on mixed models. In particular, area-level

EBLUP and pseudo-EBLUP derived from the unit-level linear mixed model guarantee some consistency with design-based estimates. Indeed, to return to the issue of estimate coherence, when small area estimates are computed, it is important to recall that small area estimates are benchmarked with respect to direct estimates published for larger domains.

2

Design-based methods for integrating information

2.1 DATA STRUCTURE AND WEIGHTING STRATEGIES

This Section examines design-based methods for computing coherent and efficient estimates, using multiple sources of information in an integrated fashion. An important preliminary step for organizing all the information available from the different surveys is the construction of an Integrated Data Set (IDS), by linking all the elementary information available from different sources of information, e.g. administrative registers and surveys carried out by different institutions. This organization of statistical data can be useful to facilitate the application of the proposed methods and to achieve the established objectives, but also to control costs and to reduce the response burden. The efforts to construct an IDS, and its informative capability in terms of the coverage of the populations of interest, depend on each country's NSS. For example, a set of agricultural data of interest may be collected with different strategies: multi-subject sampling surveys covering different contents, relating to different target populations; registers in which data on land usage, crops and production are stored; and agricultural and population censuses. Furthermore, geo-referenced information may be integrated in an IDS, to enable a comprehensive spatial analysis to be conducted. In Part 3 above, two types of integrated populations were introduced: the Integrated Data Set of Individuals (IDSI), linking e.g. individuals and households, and the Integrated Population of Land parcels (IPL), which linked, for example, land parcels and farms. Referring to each integrated population, it is possible to define the corresponding IDS on the basis of which editing and imputation of missing values can be performed.

Figure 2.1 below displays an example of IDS developed from three different sources of information: a register, a large survey and a small survey. Variables $R_1,\ldots,R_q$ are observed for all units of the register, variables $V_1,\ldots,V_p$ are observed for the larger survey S_1 and values for $Z_1,\ldots,Z_r$ are available from the small survey S_2.

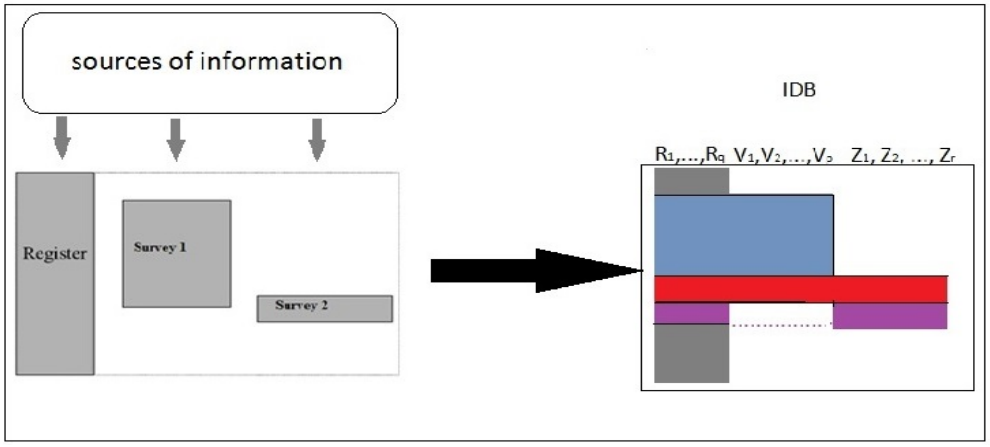

FIGURE 2.1. Example of an IDS

After performing the record linkage of the three different sources of information, we obtain an IDS which features several different *blocks*, defined by subsets of units and available variables. More precisely, as shown in Figure 2.2 below:

- the first block BLK_1 is defined considering all units in the register, and includes the correspondent values of variables $R_1,\ldots,R_q$;
- the second block BLK_2 is defined as the intersection between R and S_1. Therefore, it contains the variables $R_1,\ldots,R_q$ and $V_1,\ldots,V_p$;
- the third block BLK_3 is the intersection of R and S_2. It includes information on $R_1,\ldots,R_q$ and $Z_1,\ldots,Z_r$;
- the fourth block BLK_4 is the intersection of S_1, S_2 and R. All variables $R_1,\ldots,R_q$, $V_1,\ldots,V_p$ and $Z_1,\ldots,Z_r$ are included in the block.

On the basis of the IDS structure introduced above, the following calibration strategies may be defined:

1. A-calibration of BLK_2 on BLK_1 totals for variables $R_1,\ldots,R_q$;
2. A-calibration of BLK_3 on BLK_1 totals for variables $R_1,\ldots,R_q$;
3. A-calibration of BLK_4 on BLK_1 totals for variables $R_1,\ldots,R_q$;
4. C-calibration of BLK_4 on BLK_2 totals for variables $V_1,\ldots,V_p$;
5. AC-calibration of BLK_4 on BLK_2 totals for the variables $V_1,\ldots,V_p$ and on BLK_1 totals for variables $R_1,\ldots,R_q$;
6. C-calibration of BLK_4 on BLK_3 totals for variables $Z_1,\ldots,Z_r$;

7. AC-calibration of BLK_4 on BLK_3 totals for variables $Z_1,\ldots,Z_p$, and on BLK_1 totals for variables $R_1,\ldots,R_q$;

8. C-calibration of BLK_4 on BLK_3 totals for variables $Z_1,\ldots,Z_p$ and on BLK_2 totals for variables $V_1,\ldots,V_p$;

9. AC-calibration of BLK_4 on BLK_3 totals for variables $Z_1,\ldots,Z_p$, on BLK_2 totals for the variables $V_1,\ldots,V_p$ and on BLK_1 totals for variables $R_1,\ldots,R_q$.

In the list above, the less efficient calibration strategies were not considered; an example is the calibration of BLK_3 on BLK_2 totals for variables $R_1,\ldots,R_q$, which is less efficient than the calibration carried out on BLK_1 totals for variables $R_1,\ldots,R_q$.

Moreover, there is a correspondence between the blocks used in the different calibrations and the information utilized in the calibration constrains. Indeed, A-calibration denotes the calibration process with respect to the set of variables $R_1,\ldots,R_q$ stored in the register; C-calibration indicates that the calibration is performed only considering totals estimated from sample surveys. An example to this effect is Case 4 of the above list, in which variables $V_1,\ldots,V_p$ of Block 2 are used for the calibration; similarly, AC-calibration is performed on the set of variables $\{A,C\}$ when both variables from the register and surveys are used for the calibration, as in Case 5.

It is important to highlight that the above calibration steps can be executed even if the data are not stored in an IDS, but are maintained separately, because they originate on the basis of a stovepipe production model. The elaboration of an IDS naturally enables a more efficient and straightforward use of the available data sets.

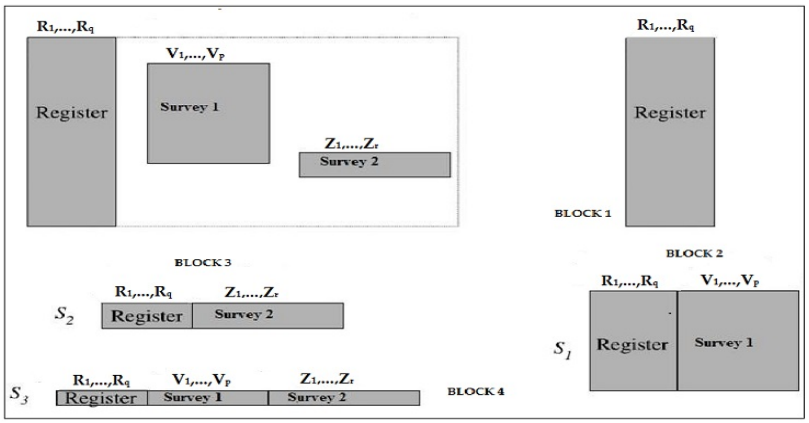

FIGURE 2.2. An example of different blocks derived from an IDS

2.2 NOTATION

In the following Sections, we consider the case in which the target population J is divided into a set of D sub-populations, or domains, of interest J_d $(d = 1, \cdots, D)$, thus forming a partition of the entire population. Then, with reference to the *j*-th statistical unit of interest belonging to the *d*-th $(d = 1, \cdots, D)$ sub-population J_d, let us denote with:

- Y_{dj} the observed value of target variable Y;
- $\mathbf{X}_{dj} = [X_{k,dj}]$ the vector of observed values, for a set of K auxiliary variables $\mathcal{X} = \{X_k\}$, in which the symbol $[\cdot]$ denotes the vector's generic element;
- w_{dj} the sampling weight, obtained as the inverse of inclusion probability π_{dj} for a correction factor δ_{dj}. When the Horvitz-Thompson (HT) estimator is used, $\delta_{dj} = 1$; when Generalized Regression (GREG) or other types of estimators are used, generally $\delta_{dj} \neq 1$.

Let us now consider the following aggregated quantities, referring to the *d*-th $(d = 1, \cdots, D)$ domain of interest for the set of data b (e.g. the generic block of information described in Figure 2.1 above):

- ${}_bT_{Y,d} = \sum_{j \in b} Y_{dj} \gamma_d$ is the total of target variable Y, with $\gamma_d = 1$ if $j \in U_d$ and zero otherwise;
- ${}_b\hat{T}_{Y,d} = \sum_{j \in b} w_{dj} Y_{dj} \gamma_d$ is the estimate of population total ${}_JT_{Y,d} = \sum_{j=1}^{J} Y_{dj} \gamma_d$;
- ${}_bT_{\bar{Y},d} = (1/{}_bN_d) \sum_{j \in b} Y_{dj} \gamma_d$ is the mean of target variable Y, with ${}_bN_d$ being the number of units belonging to set b and domain d;
- ${}_b\hat{T}_{\bar{Y},d} = (1/{}_bN_d) \sum_{j \in b} w_{dj} Y_{dj} \gamma_d$ is the estimate of population mean ${}_JT_{\bar{Y},d} = (1/J) \sum_{j=1}^{J} w_{dj} Y_{dj} \gamma_d$;
- for the set of auxiliary variables X, the corresponding aggregated quantities are therefore denoted as: ${}_bT_{x,d}$, ${}_bT_{x,d}$, ${}_bT_{\bar{x},d}$ and ${}_bT_{\bar{x},d}$.

Totals and means at the domain level may be added to obtain the correspondent overall totals and means. These are denoted in the same manner as the correspondent domain-level quantities, without the domain subscript. Finally, to distinguish between different

estimators at the domain level, the generic estimator EST (for $EST = HT, REG, \cdots$) of a vector of population totals or means ${}_J T_{*,d}$, is denoted with ${}_J T_{*,d}^{EST}$; the correspondent estimator at national level ${}_J T_*$ is denoted as ${}_J T_*^{EST}$.

2.3 INTEGRATION OF DIFFERENT SURVEYS THROUGHOUT CALIBRATION METHODS

In this Section, we consider different statistical techniques that introduce consistency between estimates by means of calibration. The first is Repeated Weighting (RW), and was proposed by the Netherlands' Central Bureau of Statistics (CBS), while the second is AC-calibration and was proposed by Särndal and Traat (2011). Different blocks for estimating specific sets of tables may be defined, as shown in Figure 2.2. The tables of interest are constructed by grouping micro-data sets, defined by means of:

- one-count variable (such as the total number of elements or the total income for the target population);
- one or more classification variables (such as sex, age-classes, farm type, land use) against which each target variable is counted.

The categories of classification variables are used to partition the population into exclusive and exhaustive sub-populations. Moreover, classification variables may consist of several hierarchical levels (r); this means that classes of a more detailed classification are nested within less detailed classes. In particular, observing Figure 2.2, different sets of tables for computing estimates of interest may be identified, by crossing the structural information $R_1, \cdots, R_q$ stored in the register, the information $V_1, \cdots, V_p$ collected through the larger sample S_1 and the $Z_1, \cdots, Z_r$ variables collected through the smaller sample S_2. Combining this information, the estimates may be computed by using all relevant information available, with the constraint that each block's margin must be consistent with the estimates or counts related to the bigger blocks.

When Y is a categorical variable whose modalities are obtained as a cross-classification of these K variables, it is useful to formalize the generic table T_y as

$$T_\Lambda = T_{\Lambda_1 \times \cdots \times \Lambda_k \times \cdots \times \Lambda_K}. \qquad (2.1)$$

The generic estimator EST (EST=HT, REG, RW,...) of Table T_Λ based on block of data b is expressed as

$${}_b T_\Lambda^{EST} = {}_b T_{\Lambda_1 \times \cdots \times \Lambda_k \times \cdots \times \Lambda_K}, \qquad (2.2)$$

where Λ represents the tabulation of count variable of interest Y according to the multiway classification $\Lambda_1 \times \cdots \times \Lambda_k \times \cdots \times \Lambda_K$, while Λ_k is the generic classification variable with $0 \leq r_k \leq R_k$ hierarchical levels.

With reference to Figure 2.2 above, and denoting with $n_1 \equiv N$ the population size BLK1, with $n_2 > n_3 > n_4$ as the sample sizes of BLK2, BLK3 and BLK4 respectively, the set of target tables is $\mathbf{T}(n_1), \mathbf{T}(n_2), \mathbf{T}(n_3), \mathbf{T}(n_4)$. Each table enables estimates to be computed at a given level of disaggregation, with a gradually descending quantity of information. Setting aside for a moment the issue of estimate precision, which should decrease from Block 1 to Block 4 due to the blocks' very definition, our goal is for the margin of the more detailed table to be sequentially coherent with the corresponding table defined in a previous block.

In the example shown in Figure 2.2 above, micro-data from Block 1 contains variables in the register. Data sets from Block 2 use information collected in Survey 1, that is combined with the information stored in the register. Block 3 contains information collected in Survey 2 and combined with information from the register. From the smallest Block 4, information from all three sources may be combined, enabling the most detailed tables to be defined, by integrating and cross-classifying the information from the register and from Surveys 1 and 2. Clearly, the more detailed the classifications, the lesser is the sample information that can be used for estimation within each class. Consistency between tables of interest can be achieved by using the properties of the regression estimator, calibrating on known or estimated totals in a single step (AC-caliration) or when necessary in two successive stages (repeated weighting). Note that the calibration is performed by associating each block with the corresponding sampling weights. Finally, investigation is required on how small-area estimation can be applied within this framework, how to deal with incomplete registers and how to address probabilistic record linkage in the estimation phase.

2.3.1 REPEATED WEIGHTING

This method is based on three main steps:

1. specification and ordering of the tables of interest, considering that a different order of tables generally implies different sets of re-weighting schemes. To avoid this problem, the most detailed margins of each target table can be defined, identifying the type of coherence among estimates that must be ensured;

2. computation of the regression estimation for each individual table under consideration. Depending upon the variable of interest, each estimate is computed using as many records as possible, and the margins in common with the set of tables already subjected to the estimation are identified;

3. the relevant table is then subjected to estimation by calibrating these margins, by means of a re-weighting procedure which enables estimates to be numerically consistent with all previous estimates derived from less detailed tables.

In addition to being more consistent, the estimates may also be more accurate, because the margins may be estimated from a larger data set or even counted from registers, and thus used as auxiliary information for the estimation concerning the target table; this strengthens the actual information available for the relevant survey.

In other words, the repeated weighting is therefore a calibration procedure, which consists of two steps. In the first step, the estimates are computed using the most appropriate and relevant data set. The largest survey, or a combination of surveys, can be used to identify the available relevant data. The regression estimator type used is given by the following formula:

$$ {}_b\hat{T}_y^{REG} = {}_b\hat{T}_y^{HT} + \left({}_b\mathbf{T}_x - {}_b\hat{\mathbf{T}}_x^{HT} \right)\hat{\beta}, \qquad (2.3) $$

where b is a generic block of Figure 2.2 above, ${}_b\hat{T}_y^{HT}$ and ${}_b\hat{\mathbf{T}}_x^{HT}$ are the Horvitz-Thompson estimators of the totals of y and x respectively, ${}_b\mathbf{T}_x$ is the known population total of $\mathbf{x}$ and $\hat{\beta}$ is the estimator of the regression coefficient. The estimates computed depend on the set of initial weights associated with each element of each individual table in a specific Block b, and on a set of auxiliary variables whose population totals $R_1,....R_q$ are known from a register R or estimated from a previous block of data. The initial weights should refer to the relevant block of data.

More precisely, from BLK_2, the weight associated with each unit depends only on the sampling design used for S_1. The same is for the starting weights for BLK_3, which only depend on the sampling design of S_2; instead, the initial weights are defined considering that the units belonging to this block consist of the intersection of the sampling units selected in S_1 and S_2. Whenever the samples are independent, the initial weights are given by the product of the two associated weights. When different sampling units from two different surveys are united, the starting weights are given by a combination of the initial weights related to the corresponding survey. Therefore, the generic initial weight, denoted with a_j, associated to the *j*-th unit in a block *b* is given by:

$$ a_j = \begin{cases} \lambda_1/\pi_{1j} & \text{if } j \in \mathrm{S}_1 \\ (1-\lambda_1)/\pi_{2j} & \text{if } j \in \mathrm{S}_2 \\ \lambda_1/\pi_{1j} + (1-\lambda_1)/\pi_{2j} & \text{if } j \in \mathrm{S}_1 \cap \mathrm{S}_2 \end{cases}, \qquad (2.4) $$

where π_{1j} and π_{2j} are the first-order inclusion probabilities for the sampling units selected for Surveys 1 and 2 respectively, while λ_1 reflects the importance of Survey 1 in the generic block b.

A simple way to define λ_1 is to set $\lambda_1 = n_1/(n_1+n_2)$, where n_1 and n_2 are the sample sizes of Surveys 1 and 2 respectively. This choice is optimal if the two samples are selected independently, with a simple random sampling design with replacement, or when the two samples are two mutually disjoint sub-samples selected without replacement from a large SRS master sample (Knottnerus *et al*., 2006), as occurs when different modules on specific topics are implemented.

The regression estimator can be written in the following way:

$$ {}_b\hat{T}_y^{REG} = \sum_{j\in S_b} w_j y_j, \tag{2.5}$$

with

$$ w_j = a_j \left\{ 1 + \mathbf{x}'_j \left(\sum_{j\in S_b} a_j \mathbf{x}_j \mathbf{x}'_j \right)^{-1} \left({}_b\mathbf{T}_x - {}_b\hat{\mathbf{T}}_x^{HT} \right) \right\}, \tag{2.6}$$

and the final weights that satisfy the calibration equation $\sum_{j\in S_b} w_j \mathbf{x}_j = {}_b\mathbf{T}_x$.

Once these weights have been computed, they can be applied to any set of variables recorded in the corresponding data set. The estimates are approximately design-unbiased, and approximate formulas can be used to estimate the design variances. Moreover, the estimates are consistent with respect to all auxiliary variables used for the calibration, and then also for the estimates considered within the calibration process. In any case, not all estimates can be taken into account in one step of the calibration, mainly because the number of constraints cannot exceed the data available for the estimation.

Indeed, the more detailed are the tables of interest, more computationally difficult it becomes to introduce all constraints in a single calibration step. In these cases, only a limited and selected amount of auxiliary information can used for the calibration. This means, for example, that the margins of the tables that can defined in BLK_4 may be inconsistent with respect to the corresponding tables identified in BLK_2 or BLK_3. In these cases, an additional calibration step is required to align the present estimates with previous corresponding ones, computed from a table defined in a previous block and therefore estimated with more information. Indeed, this method is called re-weighting or repeated weighting. Denoting with m the margins that must be consistent with previous estimates, the expression of this second step of calibration is as follows:

$$ {}_b\hat{T}_y^{RW} = {}_b\hat{T}_y^{REG} + \left({}_b\hat{\mathbf{T}}_m^{RW} - {}_b\hat{\mathbf{T}}_m^{REG} \right) \hat{\beta}_m^{RW} \tag{2.7}$$

with repeated final weights w_j^{RW} given by:

$$w_j^{RW} = w_j \left\{ 1 + \mathbf{m}'_j \left(\sum_{j \in S_b} w_j \mathbf{m}_j \mathbf{m}'_j \right)^{-1} \left({}_b\hat{\mathbf{T}}_m^{RW} - {}_b\hat{\mathbf{T}}_m^{REG} \right) \right\}. \qquad (2.8)$$

The final weights computed with this second step of calibration enable consistency between different estimate tables, with $\sum_{i \in S_b} w_i^{RW} m_i = {}_b\hat{T}_m^{RW}$.

To derive the variance of RW estimators, let us suppose that the target table of interest ${}_bT_y$ is obtained by means of a cross-classification of two categorical variables $\Lambda_1 \times \Lambda_2$, synthetically denoted as $T_{\Lambda_1 \times \Lambda_2}$. Furthermore, let us suppose that, on the basis of sample S_1, a reliable estimate of marginal table T_{Λ_1} is obtained; on the basis of sample S_2, drawn independently from S_1, a reliable estimate of marginal table T_{Λ_2}, is computed; from $S_3 \left(\equiv S_1 \cap S_2 \right)$, an estimate of the two-way table $T_{\Lambda_1 \times \Lambda_2}$ must be obtained, consistently with estimates of marginal tables T_{Λ_1} and T_{Λ_2}. For samples S_1 and S_2, there is a common set x of A-variables, whose totals are retrieved from a population register; these variables are then used to produce regression estimates ${}_{S_1}T_{\Lambda_1}^{REG}$, ${}_{S_2}T_{\Lambda_2}^{REG}$ and ${}_{S_3}T_{\Lambda_1 \times \Lambda_2}^{REG}$ of the corresponding tables.

On the basis of the general formulas set out above, the RW estimator of Table $T_{\Lambda_1 \times \Lambda_2}$ is expressed as

$${}_{S_3}T_{\Lambda_1 \times \Lambda_2}^{RW} = {}_{S_3}T_{\Lambda_1 \times \Lambda_2}^{REG} + B_{\Lambda_1}^T \left(T_{\Lambda_1}^{RW} - {}_{S_3}T_{\Lambda_1}^{REG} \right) + B_{\Lambda_2^-}^T \left(T_{\Lambda_2^-}^{RW} - {}_{S_3}T_{\Lambda_2^-}^{REG} \right),$$

where Λ_2^- is the set of categories of variable $\mathbf{\Lambda}_2$ minus one, to avoid linear dependence between constraints $T_{\Lambda_1}^{RW} = {}_{S_1}T_{\Lambda_1}^{REG}$ and $T_{\Lambda_2^-}^{RW} = {}_{S_2}T_{\Lambda_2^-}^{REG}$.

It is useful to rewrite the RW estimator as a function of population residuals

$${}_{S_3}T_{\Lambda_1 \times \Lambda_2}^{RW} = T_{\Lambda_1 \times \Lambda_2} + B_{\Lambda_1}^T \left({}_{S_1}T_{e(\Lambda_1)}^{HT} - {}_{S_3}T_{e(\Lambda_1)}^{HT} \right) + B_{\Lambda_2^-}^T \left({}_{S_2}T_{e(\Lambda_2^-)}^{RW} - {}_{S_3}T_{e(\Lambda_2^-)}^{HT} \right),$$

where the random character of regression matrices B has not been considered. Now, collecting the residuals for each sample, we obtain

$${}_{S_3}T_{\Lambda_1 \times \Lambda_2}^{RW} = T_{\Lambda_1 \times \Lambda_2} + {}_{S_1}T_{\varepsilon_1, \Lambda_1 \times \Lambda_2}^{HT} + {}_{S_2}T_{\varepsilon_2, \Lambda_1 \times \Lambda_2}^{HT} + {}_{S_3}T_{\varepsilon_3, \Lambda_1 \times \Lambda_2}^{HT},$$

where

- $\varepsilon_{1j,\Lambda_1\times\Lambda_2} = B^T_{\Lambda_1} e_j(\Lambda_1)$
- $\varepsilon_{2j,\Lambda_1\times\Lambda_2} = B^T_{\Lambda_2^-} e_j(\Lambda_2^-)$
- $\varepsilon_{3j,\Lambda_1\times\Lambda_2} = e_j(\Lambda_1\times\Lambda_2) - B^T_{\Lambda_1} e_j(\Lambda_1) - B^T_{\Lambda_2^-} e_j(\Lambda_2^-)$.

Then, the table's covariance matrix is

$$Cov\left(S_3 T^{RW}_{\Lambda_1\times\Lambda_2}\right) = Cov\left(S_1 T^{HT}_{\varepsilon_1,\Lambda_1\times\Lambda_2} + S_2 T^{HT}_{\varepsilon_2,\Lambda_1\times\Lambda_2} + S_3 T^{HT}_{\varepsilon_3,\Lambda_1\times\Lambda_2}\right)$$
$$= Cov\left(\sum_{j\in S_1} w_j \varepsilon_{1j,\Lambda_1\times\Lambda_3} + \sum_{j\in S_2} w_j \varepsilon_{2j,\Lambda_1\times\Lambda_2} + \sum_{j\in S_3} w_j \varepsilon_{3j,\Lambda_1\times\Lambda_2}\right).$$

Finally, the table's covariance matrix may be expressed as

$$Cov\left(S_3 T^{RW}_{\Lambda_1\times\Lambda_2}\right) = \sum_{k=1}^{3}\left\{\frac{n_k}{n_k - 1}\sum_{j\in S_k} u_{kj}u^T_{kj} + -\frac{1}{n_k - 1}\left(\sum_{j\in S_k} u_{kj}\right)\left(\sum_{j\in S_k} u^T_{kj}\right)\right\},$$

where $u_{kj} = w_j \varepsilon_{kj,\Lambda_1\times\Lambda_2}$ $(k = 1,2,3)$.

This design-based method is applicable to the estimation of all tables under consideration, as long as the sample is sufficiently large to yield reliable estimates of the table. When the sample size associated with an excessively detailed table is not sufficiently large, small area methods should be applied. In this case, it is necessary to investigate how these methods can be used within this framework, considering the benchmarking of model-based small area estimation with respect to direct estimates published for larger domains. In any case, a definite advantage is that by integrating the information available, small-area modeling can be performed better.

The pros of the method consist in its enabling:

1. computation of consistent estimates from different data sub-sets;

2. more accurate estimates to be obtained: more auxiliary information can be used during the calibration, because known (or estimated) population totals of the register (or surveys) can be considered in the calibration process;

3. reduction of the response burden, because some information can be gathered from registers which contain information on more or less all elements of the population;

4. detection of more detailed information on subgroups within the population;

5. generation of new outputs on different populations, by linking registrations.

2.3.2 AC-CALIBRATION

Särndal and Traat (2010) have proposed an alternative method to repeated weighting. They propose different calibration estimators for domains when two surveys present a common set of variables. The smallest survey, based on a sample S_2, is termed the Present Survey (PS), while the largest survey, based on a sample S_1 or on a register R, is the Reference Survey (RS). The RS provides reliable estimates (or counts values, in case of a register) for the variables common to the two surveys at the national level. The PS provides estimates of the corresponding totals for some sub-populations or domains $d = 1,\ldots,D$ of interests; if the domain indicator variable is supposed to be absent in the RS, then the RS estimates of domain population totals cannot be obtained.

To achieve consistency between the estimates of the surveys RS and PS, the calibration weights must satisfy the constraints $\sum_{d\,PS} \hat{\mathbf{T}}_{x,d} =_{RF} \hat{\mathbf{T}}_x$. The final weights are called C-calibrated weights and denoted as w_{Cdi}.

When the RS coincides with a register, the parameter of interest is computed by simply adding the true non-random values associated to each element of the register, and ${}_{RF}\hat{\mathbf{T}}_x \equiv \mathbf{T}_x$. To distinguish this case, in which the common variables are stored in a register, from the previous one in which the common variables were collected by means of a survey, the variables $\mathbf{x}$ are called "A-variables".

In this case, a classical calibration is performed, as described in Deville and Särndal (1992). The final weights of the PS are obtained through the calibration, to satisfy the constraints $\sum_{d\,PS} \hat{\mathbf{T}}_{x,d} = \mathbf{T}_x$. Importantly, if the A-variables display a strong correlation with the target variables, there are efficiency gains for the domain estimates of A-variables, and for other target variables. This classical calibration strategy is called A-calibration.

Indeed, if the RS is a sampling survey, the final weights of the PS are calibrated to satisfy the constraints $\sum_{d\,PS} \hat{\mathbf{T}}_{x,d}] =_{RF} \hat{\mathbf{T}}_x$. It must be noted that even when the C-variables display a strong correlation with the target variables, there is no certainty that the efficiency of the domain estimates will improve. Therefore, to achieve a gain in efficiency, the C variables estimated with the RS must present a sufficiently low sample variance. In the following Sections, this calibration strategy will be denoted as

C-calibration. These are the two extreme situations; however, it is often possible to calibrate the *PS*' initial weights by examining the set of A-variables, whose totals are known from a population register, and at the same time, the set of C-variables observed in a larger survey. This occurs when multiple sources of information are available. In this case, the RS coincides with the register for the A-variables, as well as with the larger sampling survey for the C-variables. In this case, the entire calibration strategy performed, which considers both Type A and Type C variables, is called AC-calibration.

To analyze these three strategies of calibration in further detail, we will reprise the blocks representation displayed in Figure 2.2. Therefore, when:

1. Block BLK_2 is considered, the RS coincides with the population register, the PS data coincide with BLK_2, the A-variables are $R_1,\cdots,R_q$ and the target variables are $V_1,\cdots,V_p$. In this case, the A-calibration is performed with $X_{k,dj} \equiv R_{k,dj}$, and the final A-calibrated weights w_{Adj} are obtained, under the constraint that $X_{k,d} = \sum_j X_{k,dj} w_{Adj}$. Observing Figure 2.2, in this case $\mathsf{Y} \equiv V_k$ ($k = 1,\cdots,p$). Therefore, the A-calibrated domain estimate for the target variable Y is $\hat{Y}_d = \sum_j Y_{dj} w_{Adj}$;

2. Indeed, when BLK_3 is considered, it is evident that $RS \equiv R$ and $PS \equiv BLK_3$. In this case, the A-variables are $R_1,\cdots,R_q$ and the target variables are $Z_1,\cdots,Z_r$. Moreover, the same A-calibration as in the previous situation can be performed, considering that in the present case, the generic target variable Y is obtained by letting $\mathsf{Y} \equiv Z_k$ ($k = 1,\cdots,r$);

3. Finally, when BLK_4 is considered, all the calibration strategies described above are possible, i.e. A-calibration, C-calibration and AC-calibration.

The two methods proposed ensure the consistency of the estimates of variables when multiple sources of information are available. Coherence between sets of estimates is a crucial objective if the estimates computed are to be validated, but it can also be useful for avoiding confusion in interpreting results.
The construction of an Integrated Data Base for a specific system of surveys is a key issue that may consistently and efficiently advance the exploitation of all information available. This is also true for small area estimation. In addition, once an IDS for each population of interest is available, more outputs of interest may be generated by linking the units belonging to each IDS, thus increasing countries' capacity to produce statistical information.

2.4 INTEGRATION OF SURVEYS THROUGH THE PROJECTION ESTIMATOR

The Projection estimator enables production of synthetic values of a target variable Y, that is not contained within the larger sample S_1 or the register R. This is done by exploiting all the information (auxiliary information and target variables) collected in the small survey S_2. The application of the method requires different blocks to possess a set of common auxiliary information, as displayed in Figure 2.2 above. Within this framework, a working model that links the target variable with the set of these auxiliary variables is fitted onto data collected by S_2. This model is then applied to the elementary units collected with a larger sample S_1, or stored in register R. Thus, the information available for the sample S_1 or the register R can be expanded, possibly enabling gains in the efficiency of the estimates produced.

The projection estimator is an asymptotically-unbiased model-assisted estimator proposed by Kim and Rao (2012); its operation is illustrated in Figure 2.3 below.

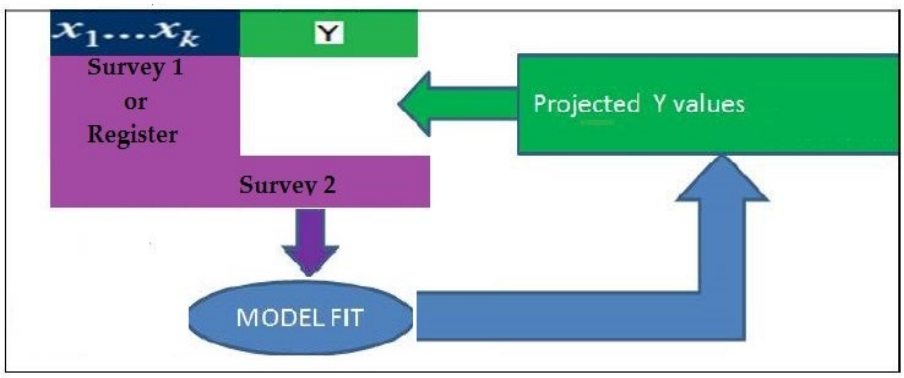

FIGURE 2.3. Projection estimator

Observing Figures 2.2 and 2.3, the steps for implementing the method are the following:

- for each target variable, the most relevant data set containing the information on this variable and on a set of auxiliary information is identified. For example, from Figure 2.2, this is Survey S_2;
- a larger block B, containing the same set of auxiliary information, is identified. This block can be another survey ($B \equiv S_1$) or a register ($B \equiv R$);
- the working model is fitted using S_2, and the regression coefficients are estimated;
- synthetic values $\tilde{Y}$ are projected onto B, through the estimated values of the regression coefficients, applied to the auxiliary information in B that is common to the auxiliary information available in S_2, and used to specify the working model;

- when the synthetic values are projected onto a larger sample, the final estimator is obtained using the sampling weights associated to each elementary unit belonging to S_1;

- when the synthetic values are projected onto the register R, the final estimates are computed by adding the synthetic values.

This method aims to improve the efficiency of the estimates under consideration, using unit-level auxiliary information of the small sample that is also available in a larger sample or a register. The working model is fitted by using data from a survey in which the target variable is observed. This model is then used to project, onto a larger survey or register, proxy or synthetic values of the target variable that were not observed in the bigger survey or stored in a register. This method is defined under a model-assisted framework; therefore, the inference does not depend on the validity of the working model, but can affect the efficiency of the final estimates if the auxiliary information used in the model is not sufficiently correlated with the target variable. The method's advantage lies in the fact that when more than one synthetic target variable is generated, each final estimate can be easily obtained, by using the unique set of sampling weights associated to the unit collected through S_1 units, or simply by summing up the corresponding values in the register.

This estimator can be applied when:

- a large first-phase sample (master sample) is selected, to collect information on auxiliary variables, and a second-phase sub-sample is drawn to gather information on both auxiliary variables and a particular target variable (e.g. modules associated to the mother survey);

- when two independent surveys are carried out on the same population, where the larger Survey 1 enables collection of some auxiliary information, and a smaller Survey 2 provides information on both target and auxiliary variables (as occurs when the two samples are selected from different frames, i.e. a list frame for Survey 1 and an area frame for Survey 2).

The method enables:

1. augmenting Survey S_1 or register R information with one or more target variables that were not observed in S_1 or stored in R;

2. more accurate estimates to be obtained, since a larger database is used;

3. integrating information, when an *ad-hoc* survey does not enable computation of an efficient estimation for particular subgroups of population;

4. in the case of IDS, the methods may also be applied to integrate some missing information.

The projection method enables the information available to be increased, thus allowing more precise estimates to be computed. However, the method's ability to improve estimate precision depends on the correlation between the target variable and the set of common auxiliary information. Indeed, as soon as the model's predictive capability increases, this estimation method performs better, in terms of the efficiency of the final estimates. Also, in this case the construction of an Integrated Data Base is a key issue for a wider application of the method, since more auxiliary information can be used to add further detail to the working model.

2.4.1 PROJECTION OF THE SMALL SAMPLE INFORMATION ONTO A LARGER SAMPLE

Let us denote with Y_j the value of the target variable of interest related to the j_{th} unit of a given population of interest, with a_{1j} and a_{2j} as the sampling weights associated to the generic unit j observed in S_1 and S_2 respectively. The aim of the projection method is to predict certain synthetic values $\widetilde{Y}_j$ associated with each unit of S_1. These values are used to estimate population or domain totals of the target parameter of interest, using the auxiliary information $\mathbf{X}_i$ collected in S_1. If $\hat{\beta}$ is the weighted estimation of the regression coefficient β, estimated by S_2 information, the synthetic values are given by the following formula:

$$\widetilde{Y}_j = \mathbf{X}_j \hat{\beta} \qquad (2.9)$$

where $\hat{\beta}$ is:

$$\hat{\beta} = (\sum_{j \in S2} a_{2j} \mathbf{X}_j \mathbf{X'}_j)^{-1} \sum_{j \in S2} a_{2j} \mathbf{X}_j Y_j. \qquad (2.10)$$

The projection estimator $\hat{\theta}$ of the total $T_y = \sum_{j=1}^{N} Y_j$ is based on the synthetic values (obtained with Formula 2.9), and is given by

$$\hat{\theta}^{proj} = \sum_{j \in S1} a_{1j} \widetilde{Y}_j + BC, \qquad (2.11)$$

where $BC = \sum_{j \in S2} a_{2j} (Y_j - \hat{Y}_j)$ is the bias correction term, $\hat{Y}_j$ being the predicted values of the target variable in S_2. When $BC = 0$, the projection estimator is given by

$$\hat{\theta}^{proj} = \sum_{j \in S1} a_{1j} \widetilde{Y}_j, \qquad (2.12)$$

which is asymptotically design-unbiased. The constraint $BC = 0$ holds when the first element of $\mathbf{X}_t$ is equal to one.

The projection estimator set out in Formula 2.11 can be written as:

$$\hat{\theta}^{proj} = \sum_{j \in S1} a_{1j} \widetilde{Y}_j + \sum_{j \in S2} a_{2j} (Y_j - \hat{Y}_j) = \sum_{j \in S1} a_{1j} \widetilde{Y}_j + \sum_{j \in S2} a_{2j} \hat{e}_j , \quad (2.13)$$

so that the variance of the projection estimator is given by:

$$var(\hat{\theta}^{proj}) = var_1(\widetilde{Y}_j) + var_2(\hat{e}_j), \quad (2.14)$$

where, using an operator notation (see Hartley, 1959), $var_1(\cdot)$ and $var_2(\cdot)$ are the proper design-based variances for S_1 and S_2 respectively. Therefore, $var_1(\widetilde{Y}_j)$ is based on n_1 sample elements, while $var_2(\hat{e}_j)$ is based on n_2 sample elements. Thus, if $n_2 << n_1$, $var_1(\widetilde{Y}_j) << var_2(\hat{e}_j)$. Moreover, when the working model is sound, the squared error terms $e_j^2 = (Y_j - \hat{Y}_j)^2$ are small and, consequently, also $var_2(\hat{e}_j)$ is small. This is why the auxiliary information common to the different sources of information must be as correlated as possible with the target variable.

In the case of domain estimation, we will denote with $d = 1, \ldots, D$ the D domain of interest, which can derive either from a geographical partition of the entire territory, or by specific cross-classification of some categorical variables. The total of the target variables of interest in each domain is $\theta_d = \sum_{j=1}^{N} \delta_j(d) Y_j$, where $\delta_j(d)$ is an indicator variable equal to one if j belongs to domain d, and zero otherwise. The projection estimator is given by:

$$\hat{\theta}_d^{proj} = \sum_{j \in S1} a_{1j} \delta_j(d) \widetilde{Y}_j . \quad (2.15)$$

This estimator $\hat{\theta}_d^{proj}$ is based on a domain sample, belonging to S_1, that is much larger than an estimator based on a domain sample belonging to S_2, even though it may be asymptotically design-biased. Therefore, its efficiency would be greater if its relative bias were small. A bias-corrected domain estimator is:

$$\hat{\theta}_d^{proj} = \sum_{j \in S1} a_{1j} \delta_j(d) \widetilde{Y}_j + \hat{B}C_d . \quad (2.16)$$

Here, $\hat{B}C_d = \sum_{j \in S2} a_{2j}\delta_j(d)(Y_j - \hat{Y}_j)$ is the bias-correction term. The estimator defined by Formula 2.16 above satisfies the internal consistency property, because the domain estimates possess the benchmarking property. Moreover, $\hat{B}C_d = 0$ when, for each domain, an intercept term is added to the auxiliary information; in these cases, the matrix X is augmented by including a domain-incident matrix Z. Consequently, all the domains of interest must be included in S_2. Indeed, the estimator defined by Formula 2.21 can be asymptotically unbiased if

$$\sum_{j \in S2} a_{2j}\delta_j(d)(Y_j - \hat{Y}_j) = 0. \tag{2.17}$$

Considering Expression 2.14, the variance of the domain estimator is given by

$$var(\hat{\theta}_d^{proj}) = var_1(\delta_j(d)\tilde{Y}_j) + var_2(\delta_j(d)\hat{e}_j)\,. \tag{2.18}$$

2.4.2 PROJECTION OF THE SAMPLE INFORMATION ONTO THE REGISTER

The synthetic values given by Formula 2.9 can be directly computed for each element belonging to a register R. In this case, the estimator of the total of the target parameter is given by:

$$\hat{\theta}_R^{proj} = \sum_{j \in S2} a_{2j}Y_j + \sum_{j \in R}\tilde{Y}_j - \sum_{j \in S2} a_{2j}\tilde{Y}_j \tag{2.19}$$

with variance

$$var(\hat{\theta}_R^{proj}) = var_2(\hat{e}_j)\,. \tag{2.20}$$

In this case too, a specific domain projection estimator can be obtained by introducing a domain indicator variable:

$$\hat{\theta}_{Rd}^{proj} = \sum_{j \in R}\delta_j(d)\tilde{Y}_j\,; \tag{2.21}$$

this is asymptotically unbiased if Condition 2.17 holds. Finally, the variance for the domain estimator is

$$var(\hat{\theta}_{Rd}^{proj}) = var_2(\delta_j(d)\hat{e}_j)\,. \tag{2.22}$$

3

Model-based methods for integrating information

3.1 INTRODUCTION OF THE PROBLEM

Suppose that a situation of multiple sources of information has arisen, as described in the previous sections, to which the projection estimator (Kim and Rao, 2012) may be applied. In this context, the information on target variables is observed from a given sample, the size of which cannot guarantee reliable estimates, and the auxiliary information available in the survey can also be observed from a larger survey or in a register. Recalling the notation adopted in the previous pages, we denote as Survey S_1 and Survey S_2 the larger and the smaller survey, respectively, while R indicates the register.

Kim and Rao (2012) proposed an asymptotically unbiased model-assisted estimator, using common unit-level auxiliary information from different sources. To this aim, a working model is fitted, using data from the smaller survey, and then predicted values are projected onto the larger survey, thus obtaining proxy or synthetic values of the target variable that were not observed in this survey.

In the following Sections, we will refer to a population J of generic units j and the following sets of data: a register R which collects values observed for a set of data over the entire population J, a sample S_1 for which only auxiliary information is available, and a smaller sample S_2 for which both target and auxiliary information is observed. The variable of interest is denoted as Y, and the set of auxiliary variables as $\mathsf{X} = (\mathsf{X}_1, \ldots, \mathsf{X}_p)$. The size of the generic set of data b in the area d will be denoted with ${}_bN_d$, $b = \{R, S_1, S_2\}$. We will indicate with ${}_bT_{y,d}$ and ${}_bT_{\mathbf{x},d}$ the totals of the target variable Y, and of the vector of auxiliary variables X over the area d in the set of data b. Analogously, ${}_bT_{\bar{y},d}$ and ${}_bT_{\bar{\mathbf{x}},d}$ denote the mean values of y and x over the area d. The subscript d will be omitted when the aggregation is performed over the entire set of data, instead of a specific area d.

Section 3.2 describes working models based on unit-level small area estimation models, while Section 3.3 analyzes the corresponding estimator when projecting the model-based predicted values on the larger sample, as in Kim and Rao (2012).

3.2 SMALL AREA UNIT-LEVEL ESTIMATION METHODS

In this Section, we give brief descriptions of the unweighted predictor EBLUP (Empirical Best Linear Unbiased Predictor) and the weighted predictor pseudo–EBLUP when a standard unit-level linear mixed model is adopted. Using the notation established previously, we refer to a sample S_2, for which both target and auxiliary information is available, and a population J, for which auxiliary variables' total or mean values should be known. Consider the nested error regression model given by

$$Y_{jd} = \mathbf{X}_{jd}^t \beta + \nu_d + \varepsilon_{jd}, \quad j = 1,\ldots,{}_J N_d, d = 1,\ldots,D, \qquad (3.1)$$

where Y_{jd} and $\mathbf{X}_{jd}$ are the values observed for the unit i in the domain d for the target variable Y, and the vector of auxiliary variables $\mathsf{X} = (\mathsf{X}_1,\ldots,\mathsf{X}_p)$, ν_d and ε_{jd} are i.i.d. respectively $\mathrm{N}(0,\sigma_\nu^2)$ and $\mathrm{N}(0,\sigma_\varepsilon^2)$. Estimates $\hat{\sigma}_\nu^2$ and $\hat{\sigma}_\varepsilon^2$ of the variance components σ_ν^2 and σ_ε^2 can be obtained by using the methods of moments, maximum likelihood (ML) or restricted maximum likelihood (REML) (see Rao, 2003, pp. 100–102).

Under the Model described in Formula 3.1, the *d*-th area total ${}_J T_{y,d} = \sum_{j=1}^{J N_d} Y_{jd}$ can be approximated by $\theta_d = {}_J \mathbf{T}_{\mathbf{x},d}^t \beta + {}_J N_d \nu_d$, where ${}_J \mathbf{T}_{\mathbf{x},d} = \sum_{j=1}^{J N_d} \mathbf{X}_{jd}$. The parameter θ_d can also be expressed in terms of mean values, i.e. $\bar{\theta}_d = \theta_d / {}_J N_d$, where $\bar{\theta}_d$ can be rewritten as $\bar{\theta}_d = {}_J \mathbf{T}_{\bar{\mathbf{x}},d}^t \beta + \nu_d$, with ${}_J \mathbf{T}_{\bar{\mathbf{x}},d} = \sum_{j=1}^{J N_d} \mathbf{X}_{jd} / {}_J N_d$. Then, an estimator of the d–th area mean value $\bar{\theta}_d$ can be computed as $\hat{\bar{\theta}}_d = {}_J \mathbf{T}_{\bar{\mathbf{x}},d}^t \hat{\beta} + \hat{\nu}_d$.

The unweighted predictor EBLUP (Battese, Harter and Fuller, 1988) of $\bar{\theta}_d$ is given by

$$\hat{\bar{\theta}}_d^u = {}_J \mathbf{T}_{\bar{\mathbf{x}},d}^t \hat{\beta}_u + \hat{\gamma}_d \left({}_{S_2} T_{\bar{y},d} - {}_{S_2} \mathbf{T}_{\bar{\mathbf{x}},d}^t \hat{\beta}_u \right), \qquad (3.2)$$

where $\hat{\gamma}_d = \hat{\sigma}_\nu^2 / (\hat{\sigma}_\nu^2 + \hat{\sigma}_\varepsilon^2 / {}_{S_2} N_d)$, ${}_{S_2} T_{\bar{y},d} = \sum_{j=1}^{S_2 N_d} Y_{jd} / {}_{S_2} N_d$

and ${}_{S_2}\mathbf{T}_{\bar{\mathbf{x}},d} = \sum_{j=1}^{S_2 N_d} \mathbf{X}_{jd} / {}_{S_2}N_d$ are the unweighted sample mean of, respectively, the target variable Y and the vector of variables X, and the estimated unweighted regression vector $\hat{\beta}_u$ is given by

$$\hat{\beta}_u = \left[\sum_{d=1}^{D} \sum_{j=1}^{S_2 N_d} \mathbf{X}_{jd} (\mathbf{X}_{jd} - \hat{\gamma}_d {}_{S_2}\mathbf{T}_{\bar{\mathbf{x}},d})^t \right]^{-1} \left[\sum_{d=1}^{D} \sum_{j=1}^{S_2 N_d} (\mathbf{X}_{jd} - \hat{\gamma}_d {}_{S_2}\mathbf{T}_{\bar{\mathbf{x}},d}) Y_{jd} \right].$$

Denoting with $\tilde{w}_{jd} = w_{jd} / \sum_{j=1}^{S_2 N_d} w_{jd}$ the normalized sampling weights for the area d, the weighted predictor pseudo–EBLUP (You and Rao, 2002) is given by

$$\hat{\bar{\theta}}_d^w = {}_J\mathbf{T}_{\bar{\mathbf{x}},d}{}^t \hat{\beta}_w + \hat{\gamma}_{dw} \left({}_{S_2}T_{y_{\tilde{w}},d} - {}_{S_2}\mathbf{T}^t_{\mathbf{x}_{\tilde{w}},d} \hat{\beta}_w \right), \qquad (3.3)$$

where $\hat{\gamma}_{dw} = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \delta_{dw}\hat{\sigma}_\varepsilon^2)$, with $\delta_{dw} = \sum_{j=1}^{S_2 N_d} \tilde{w}_{jd}^2$, ${}_{S_2}T_{y_{\tilde{w}},d} = \sum_{j=1}^{S_2 N_d} \tilde{w}_{jd} Y_{jd}$, ${}_{S_2}\mathbf{T}_{\mathbf{x}_{\tilde{w}},d} = \sum_{j=1}^{S_2 N_d} \tilde{w}_{jd} \mathbf{X}_{jd}$ are the weighted sample mean of y and $\mathbf{x}$ and the estimated weighted regression vector $\hat{\beta}_w$ is

$$\hat{\beta}_w = \left[\sum_{d=1}^{D} \sum_{j=1}^{S_2 N_d} \tilde{w}_{jd} \mathbf{X}_{jd} (\mathbf{X}_{jd} - \hat{\gamma}_{wd} {}_{S_2}\mathbf{T}_{\mathbf{x}_{\tilde{w}},d})^t \right]^{-1} \cdot \cdot \left[\sum_{d=1}^{D} \sum_{j=1}^{S_2 N_d} \tilde{w}_{jd} (\mathbf{X}_{jd} - \hat{\gamma}_{wd} {}_{S_2}\mathbf{T}_{\mathbf{x}_{\tilde{w}},d}) Y_{jd} \right].$$

The pseudo–EBLUP estimator (Formula 3.3) depends on the design weights, and is design-consistent. Furthermore, assuming that the weights are calibrated on the population sizes ${}_J N_d$ and that the unit-level model includes the intercept term, this estimator satisfies the benchmarking property, in the sense that it adds up to the direct survey regression estimator when aggregated over the areas (You and Rao, 2002). That is, $\sum_{d=1}^{D} \hat{\theta}_d^w$ equals to $\hat{T}_{y,d} + (\mathbf{T}_{\mathbf{x},d} - \hat{\mathbf{T}}_{\mathbf{x},d})^t \hat{\beta}_w$, where $\hat{T}_{y,d} = {}_{S_2}T_{y_{w,d}} = \sum_{d=1}^{D} \sum_{j=1}^{S_2 N_d} w_{jd} Y_{jd}$ and $\hat{\mathbf{T}}_{\mathbf{x},d} = {}_{S_2}\mathbf{T}_{\mathbf{x}_{w,d}} = \sum_{d=1}^{D} \sum_{j=1}^{S_2 N_d} w_{jd} \mathbf{X}_{jd}$ are the direct estimators of the overall totals ${}_J T_y = \sum_{d=1}^{D} \sum_{j=1}^{J N_d} Y_{jd}$ and ${}_J\mathbf{T}_{\mathbf{x}} = \sum_{d=1}^{D} \sum_{j=1}^{J N_d} \mathbf{X}_{jd}$ respectively.

The mean squared error (MSE) of the EBLUP estimator $\hat{\theta}_d^u$ is estimated by

$$mse(\hat{\bar{\theta}}_d^u) = g_{1d}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2) + g_{2d}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2) + 2g_{3d}(\hat{\sigma}_v^2, \hat{\sigma}_\varepsilon^2), \qquad (3.4)$$

where

$$g_{1d}(\hat{\sigma}_\nu^2, \hat{\sigma}_\varepsilon^2) = (1-\hat{\gamma}_d)\hat{\sigma}_\nu^2,$$
$$g_{2d}(\hat{\sigma}_\nu^2, \hat{\sigma}_\varepsilon^2) = \left({}_J\mathbf{T}_{\bar{\mathbf{x}},d} - \hat{\gamma}_{d\,S_2}\mathbf{T}_{\bar{\mathbf{x}},d}\right)^t \left(\sum_{d=1}^{D} \mathbf{x}_d^t \hat{\mathbf{V}}_d^{-1} \mathbf{x}_d\right) \left({}_J\mathbf{T}_{\bar{\mathbf{x}},d} - \hat{\gamma}_{d\,S_2}\mathbf{T}_{\bar{\mathbf{x}},d}\right),$$
$$g_{3d}(\hat{\sigma}_\nu^2, \hat{\sigma}_\varepsilon^2) = {}_{S_2}\, N_d^{-2}\left(\hat{\sigma}_\nu^2 + \hat{\sigma}_{\varepsilon\,S_2}^2 N_d^{-1}\right)^{-3} h(\hat{\sigma}_\nu^2, \hat{\sigma}_\varepsilon^2),$$

with

$$\hat{\mathbf{V}}_d^{-1} = \hat{\sigma}_\varepsilon^2 \mathbf{I}_{S_2 N_d} + \hat{\sigma}_\nu^2 \mathbf{1}_{S_2 N_d} \mathbf{1}_{S_2 N_d}^t,$$

and

$$h(\hat{\sigma}_\nu^2, \hat{\sigma}_\varepsilon^2) = \hat{\sigma}_\varepsilon^4 var(\sigma_\nu^2) - 2\hat{\sigma}_\nu^2 \hat{\sigma}_\varepsilon^2 cov(\hat{\sigma}_\nu^2, \hat{\sigma}_\varepsilon^2) + \hat{\sigma}_\nu^4 var(\sigma_\varepsilon^2).$$

The MSE estimator of Formula 3.3 is

$$mse(\hat{\bar{\theta}}_d^w) = g_{1dw}(\hat{\sigma}_\nu^2, \hat{\sigma}_\varepsilon^2) + g_{2dw}(\hat{\sigma}_\nu^2, \hat{\sigma}_\varepsilon^2) + 2g_{3dw}(\hat{\sigma}_\nu^2, \hat{\sigma}_\varepsilon^2), \quad (3.5)$$

where

$$g_{1dw}(\hat{\sigma}_\nu^2, \hat{\sigma}_\varepsilon^2) = (1-\hat{\gamma}_{dw})\hat{\sigma}_\nu^2,$$
$$g_{2dw}(\hat{\sigma}_\nu^2, \hat{\sigma}_\varepsilon^2) = \left({}_J\mathbf{T}_{\bar{\mathbf{x}},d} - \hat{\gamma}_{dw\,S_2}\mathbf{T}_{\mathbf{x}_{\tilde{w}},d}\right)^t \boldsymbol{\Psi}_w \left({}_J\mathbf{T}_{\bar{\mathbf{x}},d} - \hat{\gamma}_{dw\,S_2}\mathbf{T}_{\mathbf{x}_{\tilde{w}},d}\right),$$
$$g_{3dw}(\hat{\sigma}_\nu^2, \hat{\sigma}_\varepsilon^2) = \hat{\gamma}_{dw}(1-\hat{\gamma}_{dw})^2 \hat{\sigma}_\varepsilon^{-4} \hat{\sigma}_\nu^{-2} h(\hat{\sigma}_\nu^2, \hat{\sigma}_\varepsilon^2),$$

with

$$\boldsymbol{\Psi}_w = \sigma_\varepsilon^2 \left(\sum_{d=1}^{D}\sum_{j=1}^{S_2 N_d} \mathbf{X}_{jd}\mathbf{Z}_{jd}^t\right)^{-1} \left(\sum_{d=1}^{D}\sum_{j=1}^{S_2 N_d} \mathbf{Z}_{jd}\mathbf{Z}_{jd}^t\right) \cdot$$
$$\cdot \left[\left(\sum_{d=1}^{D}\sum_{j=1}^{S_2 N_d} \mathbf{X}_{jd}\mathbf{Z}_{jd}^t\right)^{-1}\right]^t + \sigma_\nu^2 \left(\sum_{d=1}^{D}\sum_{j=1}^{S_2 N_d} \mathbf{X}_{jd}\mathbf{Z}_{jd}^t\right)^{-1} \cdot$$
$$\cdot \left[\sum_{d=1}^{D}\left(\sum_{j=1}^{S_2 N_d} \mathbf{Z}_{jd}\right)\left(\sum_{j=1}^{S_2 N_d} \mathbf{Z}_{jd}\right)^t\right]\left[\left(\sum_{d=1}^{D}\sum_{j=1}^{S_2 N_d} \mathbf{X}_{jd}\mathbf{Z}_{jd}^t\right)^{-1}\right]^t,$$

and $\mathbf{Z}_{jd} = w_{jd}\left(\mathbf{X}_{jd} - \hat{\gamma}_{dw\,S_2}\mathbf{T}_{\mathbf{x}_{\tilde{w}},d}\right)$. The MSE estimator in Formula 3.5 is almost unbiased under non–informative sampling (see Rao, 2003).

3.3 UNIT-LEVEL MODEL-BASED PROJECTION

The projection estimator is based on the availability of a set of common auxiliary variables, between two different sets of data arising from two blocks of data in the IDS or, alternatively, two sets of data arising from two separate sources of data. The core of the method is to project, onto a block of data, synthetic values $\widetilde{Y}$, computed by means of a model defined on a set of covariates $\mathbf{x}$ and fitted onto a smaller block data, for which both the target variables Y and X are observed. In this section, instead of applying design-based methods as in Kim and Rao (2012), we apply model-based estimation methods, similarly to what Elbers *et al*. (2003) operated for poverty mapping (ELL method). Elbers *et al*. (2003) use the standard unit-level linear mixed model to project synthetic values on census or larger survey data. We will use model-based methods to create the synthetic values that are different from those proposed by Elbers *et al*. (2003).

The idea is to use the Model described in Formula 3.1 and the unweighted or weighted model parameters in the EBLUP (Formula 3.2) or the pseudo–EBLUP (Formula 3.3) to compute synthetic values $\widetilde{Y}_{jd}$ for each unit i in the area d, for all i in the larger set of data. This means that $\widetilde{Y}_{jd}$ can be derived, as

$$\widetilde{Y}_{jd} = \mathbf{X}_{jd}^{t}\hat{\beta} + \hat{v}_{d}.$$

Denote with J the target population, and let ${}_{J}N_{d}$ indicate the size of J in the area *d*. Recall the expression of the target parameter $\overline{\theta}_{d}$ given in the previous section; we can write

$$\overline{\theta}_{d} = {}_{J}\mathbf{T}_{\bar{\mathbf{x}},d}^{t}\beta + \nu_{d} = \frac{1}{{}_{J}N_{d}}\sum_{j=1}^{{}_{J}N_{d}}\mathbf{X}_{jd}^{t}\beta + \nu_{d}. \qquad (3.6)$$

An estimator $\hat{\overline{\theta}}_{d}$ of the d –th area means that $\overline{\theta}_{d}$ can be computed as

$$\hat{\overline{\theta}}_{d} = {}_{J}\mathbf{T}_{\bar{\mathbf{x}},d}\hat{\beta} + \hat{v}_{d} = \frac{1}{{}_{J}N_{d}}\sum_{j=1}^{{}_{J}N_{d}}\left(\mathbf{X}_{jd}^{t}\hat{\beta} + \hat{v}_{d}\right) = \frac{1}{{}_{J}N_{d}}\sum_{j=1}^{{}_{J}N_{d}}\widetilde{Y}_{jd}. \quad (3.7)$$

Here, $\hat{\beta}$ and $\hat{v}_{d}$ are estimated using the sample data of a survey S_{2}, for which the target variable Y and the set of auxiliary variables X are observed.

Suppose that the mean values ${}_{J}\mathbf{T}_{\bar{\mathbf{x}},d}$ are unknown. Of course, also values $\mathbf{X}_{jd}$ are unknown, and Equation 3.7 cannot be computed. The common practice is to estimate the unknown variable ${}_{J}\mathbf{T}_{\bar{\mathbf{x}},d}$ from external sources of data, either registers or larger and

reliable samples. As we will see, this is equivalent to projecting synthetic values $\widetilde{Y}_{jd}$ to the external source of data, i.e. a register or a larger survey sample data set. Expressions of Equation 3.7 will be given when the synthetic values $\widetilde{Y}_{jd}$ are derived from both the EBLUP (Formula 3.2) and the pseudo–EBLUP (Formula 3.3) predictors.

3.3.1 PROJECTION ON REGISTER DATA SETS

Suppose that the target variable and a set of auxiliary variables are observed on a small sample data set, from a survey S_2. Furthermore, the same set of auxiliary variables are available in a register R. Synthetic values $\widetilde{Y}_{jd}$ for each unit i in the area $d, j = 1,\ldots,{}_{R}N_d$ can be computed, using Model 3.1 and the unweighted or weighted estimates. This means that $\widetilde{Y}_{jd}$ can be derived as

$$\widetilde{Y}_{jd} = \mathbf{X}_{jd}^{t}\hat{\beta} + \hat{v}_d, \quad j = 1,\ldots,{}_{R}N_d, d = 1,\ldots,D.$$

Assuming that the register R is built upon the same units of J, the means ${}_{J}\mathbf{T}_{\bar{\mathbf{x}},d} = \sum_{j=1}^{J N_d}\mathbf{X}_{jd}/{}_{J}N_d$ can be computed using R data by ${}_{R}\mathbf{T}_{\bar{\mathbf{x}},d} = \sum_{j=1}^{R N_d}\mathbf{X}_{jd}/{}_{R}N_d$.

Therefore, Formula 3.7 can be rewritten as

$$\hat{\bar{\theta}}_d = {}_{R}\mathbf{T}_{\bar{\mathbf{x}},d}^{t}\hat{\beta} + \hat{v}_d = \frac{1}{{}_{R}N_d}\sum_{j=1}^{R N_d}\left(\mathbf{X}_{jd}^{t}\hat{\beta} + \hat{v}_d\right) = \frac{1}{{}_{R}N_d}\sum_{j=1}^{R N_d}\widetilde{Y}_{jd}. \qquad (3.8)$$

Then, Formula 3.8 is the EBLUP predictor (Formula 3.2) or the pseudo–EBLUP predictor (3.3), depending upon whether unweighted or weighted model parameters $\hat{\beta}$ and $\hat{v}_d$ are computed using the smaller block of data S_2. Consequently, the expressions of the MSE estimators in Formulas 3.4 and 3.5 are still valid.

3.3.2 PROJECTION ON SAMPLE DATA SETS

Once again, suppose that the target variable is observed only on a small-sized sample data set, from survey S_2. Furthermore, the same set of auxiliary variables is observed for S_2 and for a larger survey S_1. As for register R, synthetic values $\widetilde{Y}_{jd}$ for each unit i in the area $d, j = 1,\ldots,{}_{S_2}N_d$ may be computed from Model 3.1, adopting either the unweighted or the weighted estimation method relating to the EBLUP and the pseudo–EBLUP respectively. Then, it results that $\widetilde{Y}_{jd}$ can be derived as

$$\widetilde{Y}_{jd} = \mathbf{X}_{jd}^{t}\hat{\beta} + \hat{v}_d, \quad j = 1,\ldots,{}_{S_1}N_d, d = 1,\ldots,D.$$

Suppose that from S_1 it is possible to obtain sound estimates of the mean values ${}_J\mathbf{T}_{\bar{\mathbf{x}}},d = \sum_{j=1}^{J\,N_d}\mathbf{X}_{jd}/{}_JN_d$, i.e. ${}_J\hat{\mathbf{T}}_{\bar{\mathbf{x}},d} = {}_{S_1}\mathbf{T}_{\mathbf{x}_{\tilde{\mathbf{w}}},d} = \sum_{j=1}^{S_1\,N_d}\mathbf{X}_{jd}\,{}_{S_1}\tilde{w}_{jd}$, where ${}_{S_1}\tilde{w}_{jd}$ are the normalized sampling weights of S_1, and is a good approximation of the unknown variable ${}_J\mathbf{T}_{\bar{\mathbf{x}}},d$. In this case, the EBLUP and pseudo–EBLUP (Formulas 3.2 and 3.3) assume the following form:

$$\hat{\bar{\theta}}_d = {}_{S_1}\mathbf{T}^t_{\mathbf{x}_{\tilde{\mathbf{w}}},d}\hat{\beta} + \hat{v}_d. \qquad (3.9)$$

Therefore, Formula 3.9 can be rewritten as:

$$\hat{\bar{\theta}}_d = {}_{S_1}\mathbf{T}^t_{\bar{\mathbf{x}},d}\hat{\beta} + \hat{v}_d = \left(\sum_{j=1}^{S_1\,N_d}\mathbf{X}^t_{jd}\,{}_{S_1}\tilde{w}_{jd}\right)\hat{\beta} + \hat{v}_d = \sum_{j=1}^{S_1\,N_d}\tilde{Y}_{jd}\,{}_{S_1}\tilde{w}_{jd}. \quad (3.10)$$

Therefore, similarly to Formula 3.8, the Predictor 3.10 or, equivalently, 3.9, is the EBLUP (Formula 3.2) or the pseudo–EBLUP (Formula 3.3) respectively, when unweighted or weighted model parameters are attained with the smaller sample data S_2. Therefore, the corresponding expressions of the MSE estimators of Formula 3.9 are given by Formulas 3.4 and 3.5.

Highlights

It is well-known that when estimates for several domains must be produced, and the information available in the sample, denoted with S_2, is insufficient to produce reliable direct estimates, it is possible to "borrow strength" from other areas, implicitly or explicitly using a statistical model. In this case, it is crucial to use good auxiliary information.

The inputs required for applying models are the auxiliary variables' total or mean values, for all domains. This information is usually taken from other sources of data, i.e. may be computed from a register R; alternatively, it is possible to use estimates from other survey samples, say S_1, which are considered reliable because of their large sample size.

The target variable is, naturally, not observed in the external source of data; otherwise, we would compute the target variable domain estimates using these larger sets of data, i.e. R or S_1. A model-based projection estimator is obtained by computing synthetic values of the target variable on a larger set of data, that has common covariates with the sample data set S_2, i.e. a register R or a larger sample S_1, and then computing the estimates of the larger data set by means of the target variable's synthetic values.

This is equivalent to computing the estimates by applying the optimal predictors deriving from the models mentioned, to "borrow strength" from the other areas in favour of the original values of the target variable in S_2. In detail, model-based projection estimators are equivalent to EBLUP or pseudo–EBLUP predictors, depending on whether unweighted or weighted models are used to

compute the synthetic values of the target variable.

This equivalence is essential, since, as already described in literature with regard to the EBLUP or pseudo–EBLUP predictors, in terms of estimate computation and quality assessment holds, of course, true for the model-based projection estimator.

3.4 MODEL-BASED PROJECTION WITH RECORD LINKAGE ERRORS

Record linkage can be used to join two sets of data that contain information on the same individuals, when there are no unique personal identification codes. The errors that can affect linkage may cause problems for estimating the relationships between variables in the two sets of data. But how do errors in record linkage affect subsequent analyses?

Recent research focuses on the reduction of record linkage error rates in the probabilistic record linkage process. However, it is not possible to provide error-free record linkage data and, as indicated by Neter *et al*. (1965), a small amount of linkage error in a linked data set could lead to significant error when ignoring the linkage errors in the linked data. Following Neter *et al*. (1965), Scheuren and Winkler (1993, 1997) and Lahiri and Larsen (2005) sought to adjust the bias due to linkage errors in the linear regression models, applying the weights used in the probabilistic record-matching process.

Here, we suppose that the sample S_2 does not contain all relevant auxiliary information that must be used in the model specification for overall or domain estimation. In this case, the missing auxiliary information can be retrieved from other sources of data, such as a register R. Perfect linkage between S_2 and the register R does not, of course, affect the estimation.

Let $\mathbf{P}$ denote the ${}_{S_2}N \times_R N$ dimensional matrix of linkage probabilities between the sample data set S_2 and the register R. The generic element $p_{jj'}$ of $\mathbf{P}$ indicates the probability that the pair $\{j, j'\}$, $j \in S_2$ and $j' \in R$, is a match. Of course, ideally, the record linkage would identify only perfect matches. This means that for each row of $\mathbf{P}$, only one element is equal to 1 and all the others are equal to 0, and that the values in each column can be either all 0's, or one value is equal to 1 and all others equal to 0. Unfortunately, this is not always the case. In this situation, it is common practice for each unit $j \in S_2$ to define the unit $j' \in R$ as a matching unit in register R, maximizing the linkage probability $p_{jj'}$, and then to treat all matches as perfect matches in subsequent computations. However, this results in underestimating the errors associated to the final estimates, and therefore an erroneous evaluation. Therefore, the more $\mathbf{P}$ departs from the ideal situation of perfect linkage, the more consideration should be given to the uncertainty introduced into the estimation process by record linkage.

It is possible to take into account the uncertainty deriving from the record linkage operations by means of the linkage probability matrix $\mathbf{P}$, and the values of the auxiliary

variables observed for all the records in the register R. For the sake of simplicity, suppose that in the model specification (Formula 3.1), all the auxiliary variables cannot be observed in Sample S_2 (this assumption will be removed later). In particular, the model parameters are estimated from the following model specification:

$$Y_{jd} = \mathbf{P}_j \mathbf{X}\beta + \nu_d + \varepsilon_{jd}, \quad j = 1,\ldots,_{S_2} N_d, d = 1,\ldots,D, \tag{3.11}$$

where $\mathbf{P}_j$ is the j–th row of the linkage probability matrix $\mathbf{P}$ related to the record linkage between S_2 and the R, and $\mathbf{X}$ is the ${}_R N \times p$ matrix with the values observed for the auxiliary variables in register R. It is expected that the values in the columns $\widetilde{\mathbf{X}} = \mathbf{PX}$ will be more concentrated than the values of the auxiliary variables associated to the sample S_2 after the record linkage. Furthermore, the shrinkage effect upon the auxiliary variables' mean values is expected to be greater as the departure of $\mathbf{P}$ from $\mathbf{P}^*$ increases, i.e. when the variables used in the record linkage process have a low discriminating power to detect matches.

This will result in a lower correlation between the values observed for the target variable Y and values in $\widetilde{\mathbf{X}}$, with respect to the correlation computed using the original values in $\mathbf{X}$. Therefore, it is expected that larger errors of model parameter estimates in Formula 3.11 and, consequently, a larger MSE, will be experienced when the linkage probability matrix $\mathbf{P}$ is considered in the model specification. In the extreme case that the discriminating power is null, the linkage probability matrix is given by $\mathbf{P} = \{1/_b N\}$. Therefore, the rows of $\widetilde{\mathbf{X}}$ would be equal to $\widetilde{\mathbf{X}}_j = \left(\overline{X}_1,\ldots,\overline{X}_p\right)$, and the introduction of the auxiliary variables in Model 3.11 would be ineffective.

Previously, we supposed that all variables X are not observed in S_2, but in other data sets, and that they are included in S_2 by means of record linkage. Naturally, it is possible that some auxiliary variables may be observed in the sample S_2, and that some other variables must be included in S_2 by using record linkage with other sources of data. In this case, the matrix $\mathbf{P}$ of linkage probabilities will multiply only submatrix $\mathbf{X}$, related to the auxiliary variables included in S_2 after the record linkage.

Let us denote with $\mathsf{X}^1 = (\mathsf{X}^1_1,\ldots,\mathsf{X}^1_{p_1})$ the set of auxiliary variables observed in S_2, and with $\mathsf{X}^2 = (\mathsf{X}^2_1,\ldots,\mathsf{X}^2_{p_2})$ the auxiliary variables added in S_2 after the record linkage operations. Then, Model 3.11 can be rewritten as

$$Y_{jd} = \{\mathbf{X}^1_{jd}\}^t \beta_1 + \mathbf{P}_j \mathbf{X}^2 \beta_2 + \nu_d + \varepsilon_{jd}, \quad j = 1,\ldots,_{S_2} N_d, d = 1,\ldots,D, \tag{3.12}$$

where $\mathbf{P}_j$ is the j –th row of the linkage probability matrix $\mathbf{P}$ of the record linkage between the sample S_2 and the register R, $\mathbf{X}^1_{jd}$ is a p_1 –dimensional vector containing the values observed for the generic unit j in area d in S_2 for variables X^1, and X^2 is the ${}_R N \times p_2$ matrix of the values observed for the auxiliary variables X^2 in R.

Again, the introduction of uncertainty related to the record linkage by means of the linkage probability matrix $\mathbf{P}$, will result in the model having lesser predictive power and, consequently, in a larger MSE of the EBLUP and pseudo–EBLUP predictors.

Highlights

Suppose that one or more relevant auxiliary variables are missing in the sample data set. If this set of variables are also included in a register, the missing variables may be added to the sample by means of a record linkage operation with the same register. Once the record linkage is performed, one of the outputs is the linkage probability matrix $\mathbf{P}$. The generic element $p_{jj'}$ of $\mathbf{P}$ is defined as the probability that the pair $\{j, j'\}$, of unit j in the sample and unit j' in the register, identifies a match.

Usually, each unit in the sample is linked to the values of the unit in the register, maximizing the corresponding linkage probability. Thus, the uncertainty introduced by the record linkage operations is not considered in the estimation process and in quality assessment.

One way to take into account the record linkage uncertainty is to introduce the linkage probability matrix in the model specifications.

In this way, the MSE estimates of EBLUP and pseudo–EBLUP predictors, and, consequently, the MSE estimates of the corresponding model-based projection estimators, automatically take into account the uncertainty deriving from possible wrong matches.

Part 5B: Application

4

Case study

4.1 DESCRIPTION OF THE APPLICATION

In this Section, we illustrate the study executed to compare the performances of the estimation methods described in Part 5 of this Technical Report. The case study was based on the data deriving from Mozambique. In particular, since two sets of files were available, a preliminary step of matching of the two types of information was performed. Indeed, the first data set consists of a subset of the 2007 Population Census, limited to Districts 7, 8 and 9 of Gaza province and a second, named Section G, which concerns the agricultural and livestock activities of households. To set up the application of estimation methods, the two data sets were matched, to achieve a single register containing both demographic and livestock variables for each farmholder. The overall size of this data set is equal to 53,679. More detailed information on Mozambique's data is available in Part 2B of this Report.

As described in the theoretical discussion of the estimation method, generally, the design- and model-based techniques proposed can be applied when the multiple sources of information can be distinguished as a register R, a large sample S_1, and a small sample S_2. The assumption is that information on target variables is available only for S_2, while auxiliary information is observed in all data sources but with differing levels of detail, which depend on the type and size of the information sources.

Since the data available for the case study are not structured as described above, we will mimic that arrangement. We will consider, as target variables, livestock variables associated with each farmholder (or, equivalently, each household), and as auxiliary information, the related demographic variables available from the Population Census. Then, two stratified samples of different sizes were drawn from the Census data, with sample sizes equal to 2,148 and 13,419, corresponding to a sample fraction equal to 4% and 25%. In this context, the register coincides with the Census, while the two selected samples coincide with S_1 and S_2. In this way, the situation with the three different data sources as displayed in Figure 2.1 above was recreated, supposing that target variables are unknown in both R and S_1. Moreover, among the several livestock variables, we chose the number of goats owned by each farmholder as the target variable, while the auxiliary information is given by the farmholder's gender, age, educational level and marital status.

To summarize the above considerations, the data set-up for the application is the following:

- register ($N = 53{,}679$): census data set, containing only demographic information;
- large sample S_1 ($n_1 = 13{,}419$): sample data set, drawn from census data containing only demographic information;
- small sample S_2 ($n_2 = 2{,}148$): sample data set, drawn from census data containing both demographic and livestock information.

In the application, the estimates are computed at national and sub-national domain levels. Since there are only three areas, i.e. districts, represented in the data, we defined as estimation domains the cross-classification of district, type of local authority, type of city and type of neighbourhood. The population domain size ranges from 83 to 4,228 households, while the average and the median sizes are, respectively, equal to 1,285.9 and 1,084 (while Q1 and Q3 are 515 and 1,697 respectively).

Since the administrative variables are not strongly correlated with the target variable, for the experimental study, artificial variables that were more correlated with the target variable were generated, to verify what would happen if more predictive auxiliary information were available. More specifically, two auxiliary variables x_1 and x_2 were created, with correlation coefficients having target variables equal to 0.7 and 0.5 respectively.

4.1.1 RESULTS OF THE PROJECTION FROM THE SMALL SAMPLE TO THE LARGER SAMPLE

In this Section, the results of the projection estimator applied to a larger sample S_1, as described in Section 5.1, are illustrated. The model used to predict the synthetic values is chosen by means of a reverse selection procedure, on the basis of the *Akaike Information Criteria* (*AIC*). Three different models were selected. The first model, MOD1, is best when only the administrative information is considered. It presents a Multiple R-squared coefficient of $R^2 = 0.3365$. The second model MOD2 is best when, as auxiliary information, we consider the original administrative covariates and the artificial variable x_1, the correlation coefficient of which has a target variable equal to 0.7. It features an $R^2 = 0.957$. Similarly, the third model MOD3 is best when the artificial variable x_2, with a correlation coefficient equal to 0.5, is added to the available auxiliary information. It presents an $R^2 = 0.7408$.

Finally, the performance of the estimators were compared to the true value of the target variable, since the samples were drawn from the Mozambique Population Census available. The comparison was performed by means of an *Averaged Absolute Relative Error* (*AARE*) and an *Averaged Squared Error* (*ASE*). These evaluation criteria are given, respectively, by:

$$AARE\left(\hat{\theta}\right)=\frac{1}{D}\sum_{d=1}^{D}\left|\frac{\hat{\theta}_d}{\theta_d}-1\right| \tag{4.1}$$

$$ASE\left(\hat{\theta}\right)=\frac{1}{D}\sum_{d=1}^{D}\left(\hat{\theta}_d-\theta_d\right)^2 \tag{4.2}$$

,

where $d=1,\cdots,D$ is the domain of interest, $\hat{\theta}_d$ is the *d*-th estimate of interest, computed with the different estimator, and θ_d is the true parameter of interest for the domain d.

The results in terms of the $AARE$ and ASE are shown in Tables 4.1 and 4.2 respectively, below. In these Tables, we denoted with Proj.HT and Proj.GREG the estimates computed by applying the H-T estimator and the GREG estimator to the synthetic $\tilde{y}$ values, projected upon large sample S_1. The final weights were obtained by calibrating, with respect to the known population, the administrative auxiliary information. We denoted with HT the H-T estimates based on the real values y, collected with the small sample S_2, while with GREG.area and GREG, the corresponding regression estimates computed by calibrating the auxiliary information on the known population. The difference between the latter two estimators is that the GREG.area estimator is based on a working model in which an intercept term for each domain of interest was added to the auxiliary information.

The GREG.area estimator was considered, to evaluate the estimates computed with respect to the same working model used to produce the synthetic values projected upon the sample S_1. The output shows that more precise estimates can be computed through the synthetic values associated with the larger sample, rather than the estimates computed with the original values observed in the smaller sample. In addition to general improvements, it is possible to note that the projection method is always more reliable than other methods, because of the great predictive capability of the model used to compute the target variable's synthetic values.

Model	Proj.HT	Proj.GREG	HT	GREG.area	GREG
MOD 1	0.2702342	0.2511199	0.3124775	0.2818680	0.3089311
MOD 2	0.1187851	0.1326969	0.3124775	0.2663313	0.2869637
MOD 3	0.1402675	0.1360465	0.3124775	0.2699044	0.2961277

TABLE 4.1. Averaged Absolute Relative Error (AARE)

Model	Proj.HT	Proj.GREG	HT	GREG.area	GREG
MOD 1	425,914	374,912	863,364	428,648	837,136
MOD 2	125,441	148,628	863,364	442,863	773,640
MOD 3	179,668	192,272	863,364	437,891	770,262

TABLE 4.2. Averaged squared error (ASE)

Figures 4.1, 4.2, and 4.3 below show the distribution of the estimates computed with the different estimators, with respect to the known true values of the target variable and the distribution of the respective coefficients of variation. The three figures concern MOD1, MOD2 and MOD3 respectively, and show that the projection method enables significant improvements in the precision of the estimates. Moreover, these improvements are related to the model under consideration, and then to the correlation between the target variable and the auxiliary information. This is even clearer from Figure 2.4, which displays the performance of the projection estimators based on the three different models; it can be seen that the projection estimator based on MOD2 enables greater improvements.

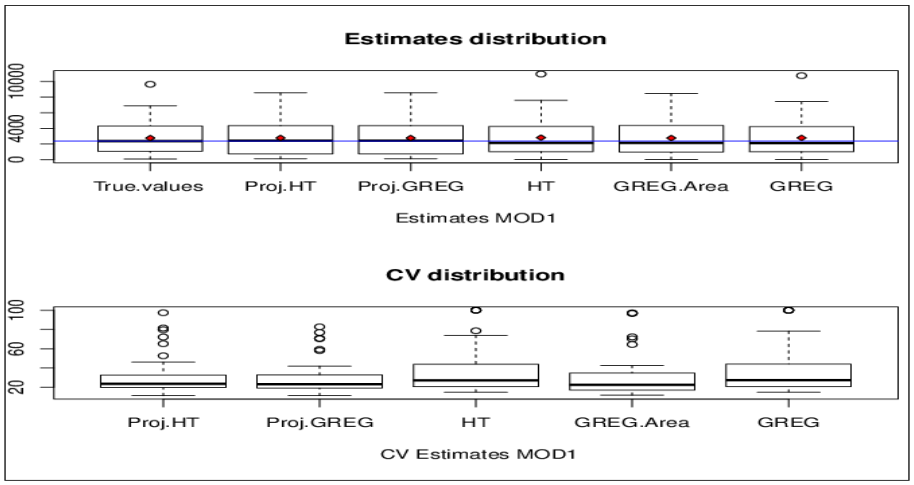

FIGURE 4.1. Estimator based on MOD1

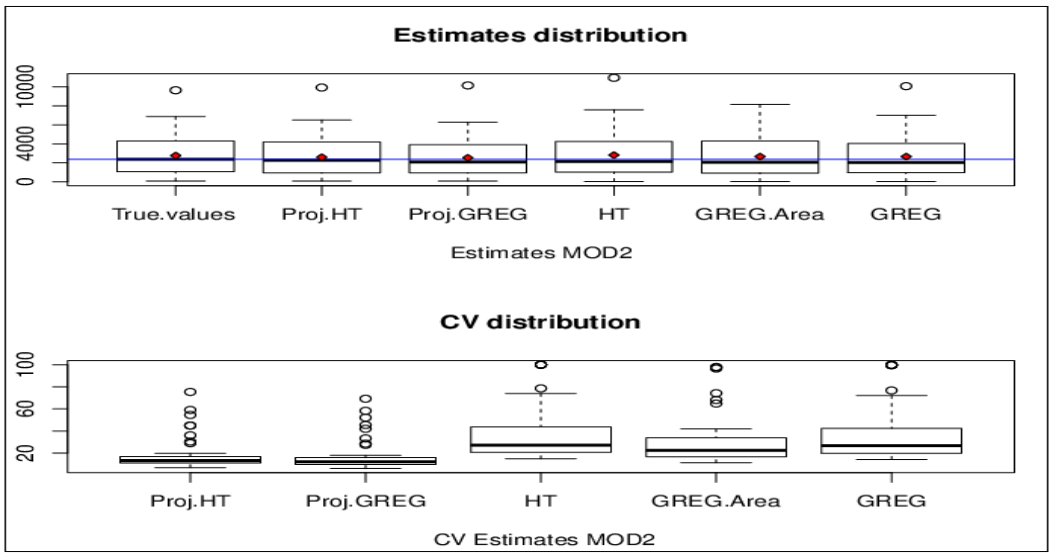

FIGURE 4.2. Estimator based on MOD2

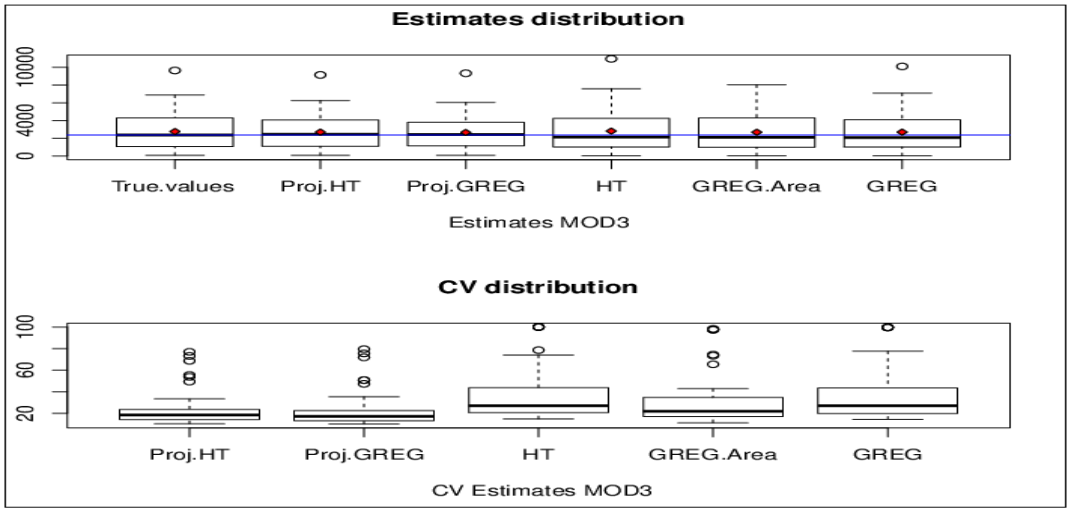

FIGURE 4.3. Estimator based on MOD3

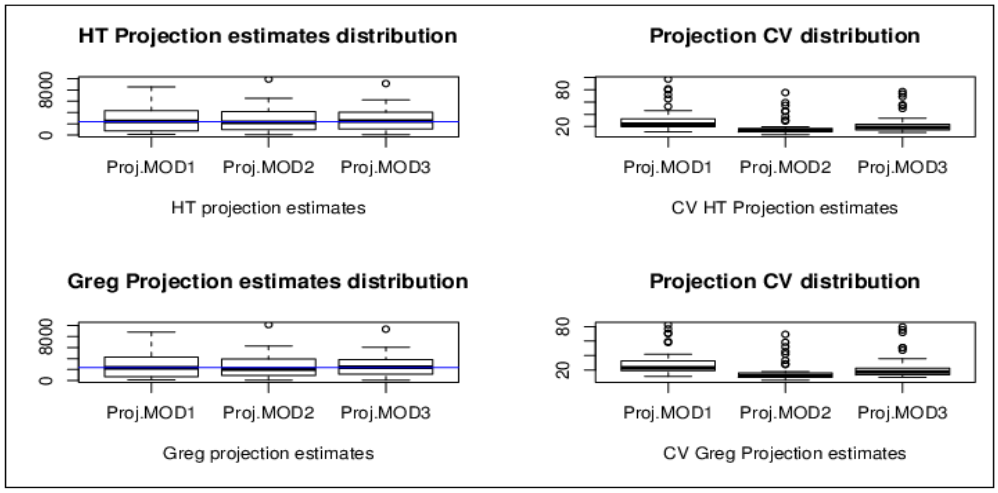

FIGURE 4.4. Projection estimators

4.1.2 RESULTS OF THE PROJECTION METHOD FROM THE SMALL SAMPLE TO THE REGISTER

In this Section, the results of the projection estimator applied to the register R as described in Section 5.2 are shown. The application is based on the same three models considered in the previous Section. In Tables 3 and 4, we present the performances of the estimators in terms of the *AARE* and *ASE*, respectively. As in the case where the synthetic values are projected onto a larger sample, the results show that gains in terms of estimate accuracy can be obtained if the projection method is applied. In this case too, the level of improvement strongly depends upon the predictive power of the model used to determine synthetic values. Indeed, MOD2 is the model that enables greater

improvements. The good performance of the projection method is also highlighted in Figures 4.5, 4.6, 4.7 and 4.8 below.

Model	*Proj*	*HT*	*GREG.area*	*GREG*
MOD1	0.261	0.312	0.282	0.309
MOD2	0.091	0.312	0.267	0.287
MOD3	0.152	0.312	0.270	0.296

TABLE 4.3. Average Absolute Relative Error (AARE)

Model	*Proj*	*HT*	*GREG.area*	*GREG*
MOD1	376.999	863.364	428.648	837.136
MOD2	109.133	863.364	442.863	773.640
MOD3	178.975	863.364	437.891	770.262

TABLE 4.4 Average squared error (ASE)

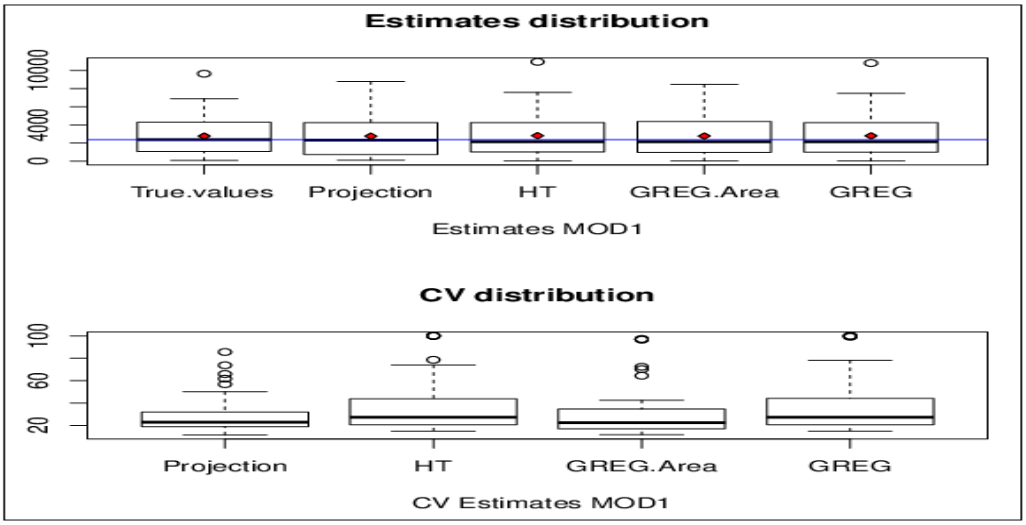

FIGURE 4.5. Estimator based on MOD1

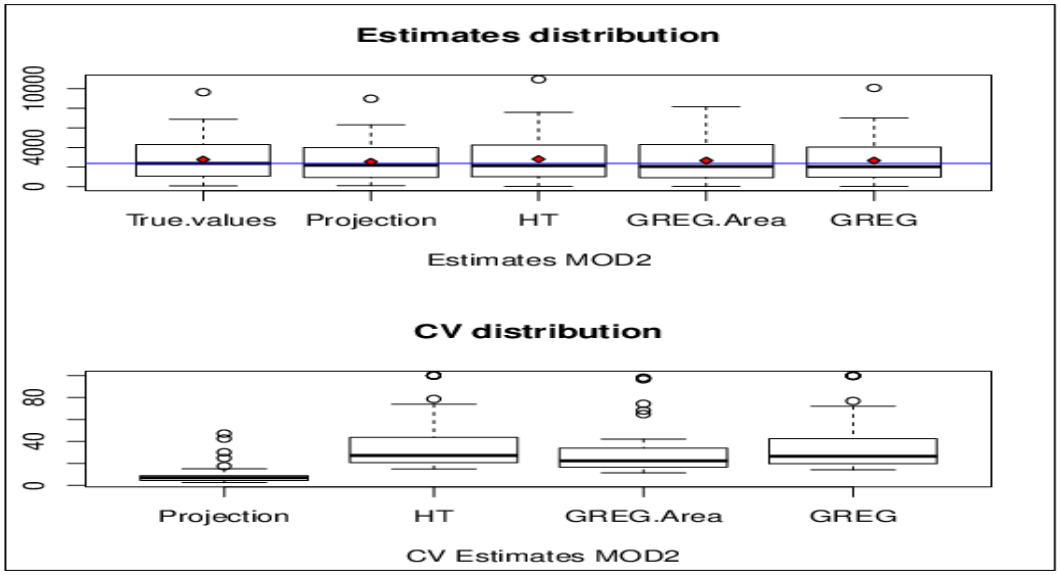

FIGURE 4.6. Estimator based on MOD2

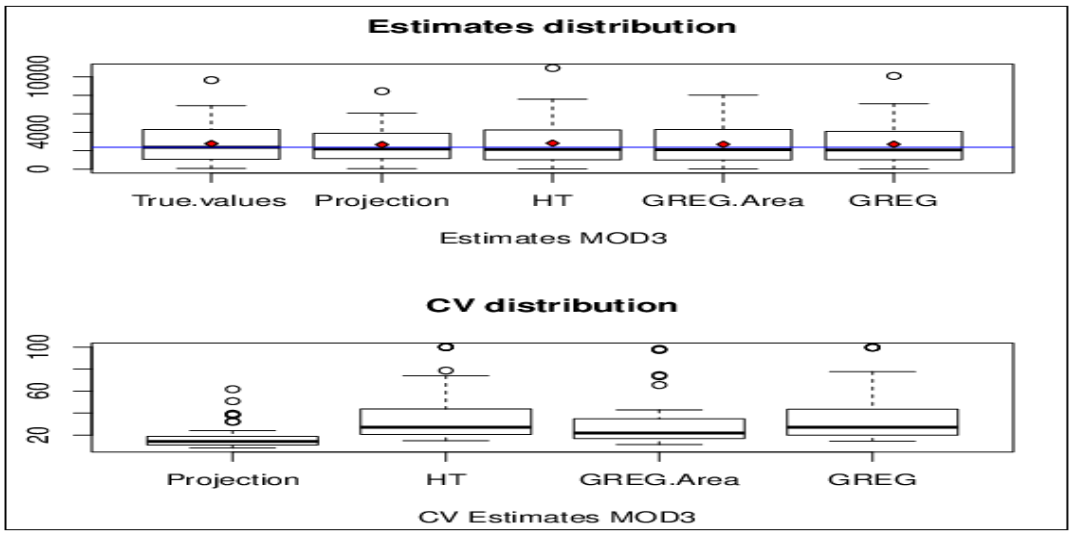

FIGURE 4.7. Estimator based on MOD3

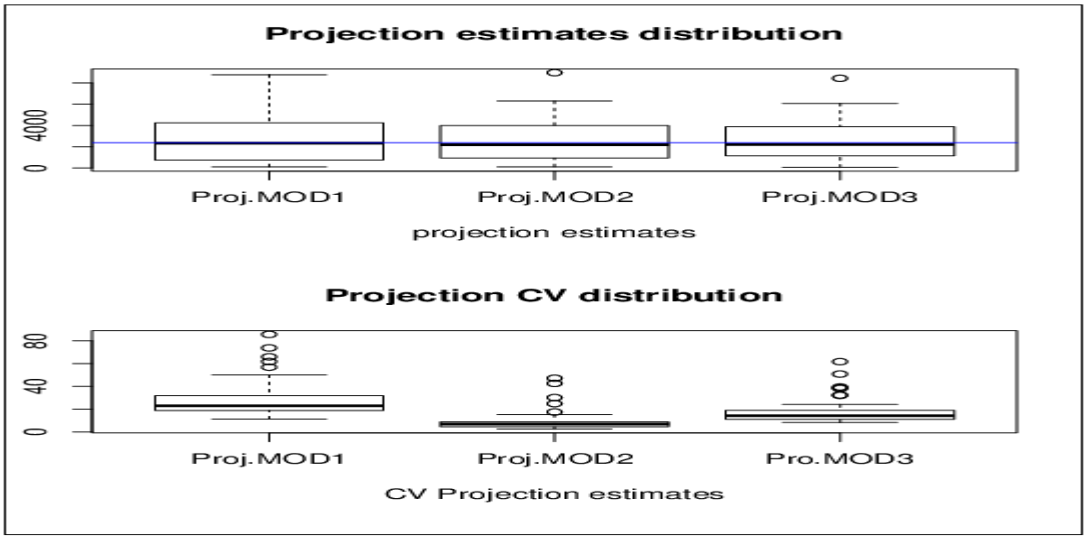

FIGURE 4.8. Projection estimators

4.1.3 RESULTS OF THE REPEATED WEIGHTING METHOD, IN TERMS OF ESTIMATE CONSISTENCY

This Section describes the performance of the repeated weighting method. This method aims to obtain consistency among estimates of the same parameters, computed by means of different surveys. To test this method, we suppose that the target variable is the domain estimation of the number of goats, classified according to farmholder gender. The application of this method ultimately seeks to obtain final weights, that enable efficient estimates of the target variable to be computed, with the constraint that the estimates of the gender counts computed with these final weights must be equal to estimates of gender counts computed by applying the GREG estimator to the larger sample data set S_1. The results in terms of the *AARE* and *ASE* are showed in Table 5 below, in which GREG denotes the single-step calibrated estimator, while REW denotes the repeated weighting estimator. Sex1 and Sex2 stand, respectively, for Male and Female.

Estimator	*AARE*	*ASE*
GREG.Sex1	0.3445638	744,667
REW.Sex1	0.2752959	164,561
GREG.Sex2	0.4494405	218,384
REW.Sex2	0.3640834	184,021

TABLE 4.5. Estimator performance

The results refer to the estimation of the number of goats (our target variable) classified according to the gender of the farmholders; the evaluation criteria are computed by averaging the estimation domains. The estimates computed with the final weights obtained by repeated weighting (the REW estimator), even if applied to ensure coherence among estimates, outperform the results obtained with the single step calibration procedure (GREG estimator), in terms of the ARE and ASE. Therefore, as the first result, we obtain that the repeated weighting method enables estimate efficiency gains.

The same conclusion may be drawn from Figure 9. Indeed, the box-plots of the CVs of the estimates of the number of goats, for male and female farmholders respectively, show that the domain estimates for REW estimators are lower than the CVs of the corresponding GREG estimates. In addition, it may be noted that the distribution of the REW estimates is also closer to the true values than the corresponding distribution of estimates computed with single-step calibration. This holds true for estimates of the number of goats for male and female farmholders. Figure 10 illustrates the coherence of the gender counts computed by applying repeated weighting on a smaller sample (REW.Male and REW.Female), with respect to the same estimates computed through the GREG estimator applied to the larger sample S_1 (GREG.Male, GREG.Female). The REW estimates are the margins of the table under consideration, that are also compared with the corresponding GREG estimates (GREG.Male.step1, GREG.Female.step1), obtained by applying a single step of calibration on sample data S_2. The true values available for both gender counts are also displayed. As may be noted by examining the figure, the REW estimates based on the small block of data are consistent with the GREG estimates computed with a bigger block of data. Instead, the estimates computed with single-step calibration, computed on the smaller block of data, are not consistent.

Therefore, the repeated method enables both coherence of estimates, and improvement of estimate efficiency. The method's good performance is also highlighted in Figure 4.9, while in Figure 4.10 it may be seen that the repeated weighting method enables consistency among estimates to be obtained.

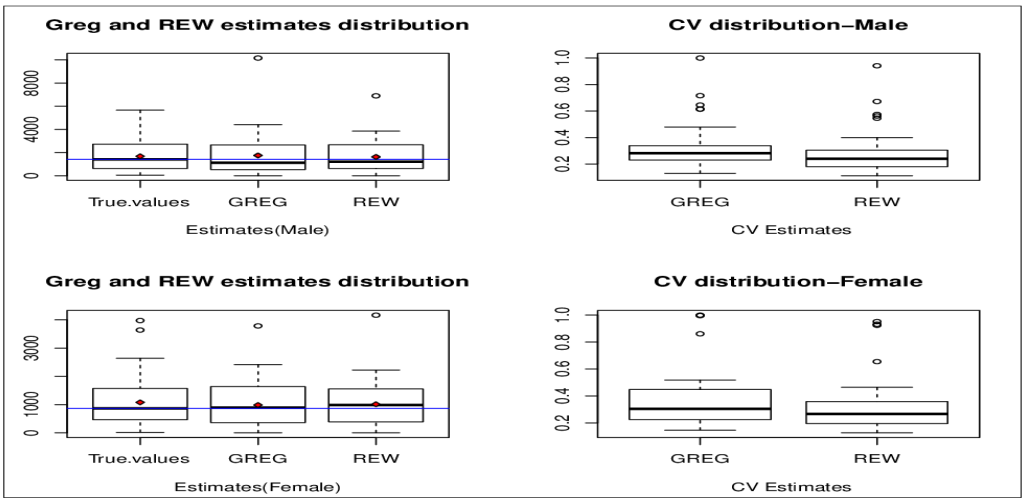

FIGURE 4.9. Comparison of estimates

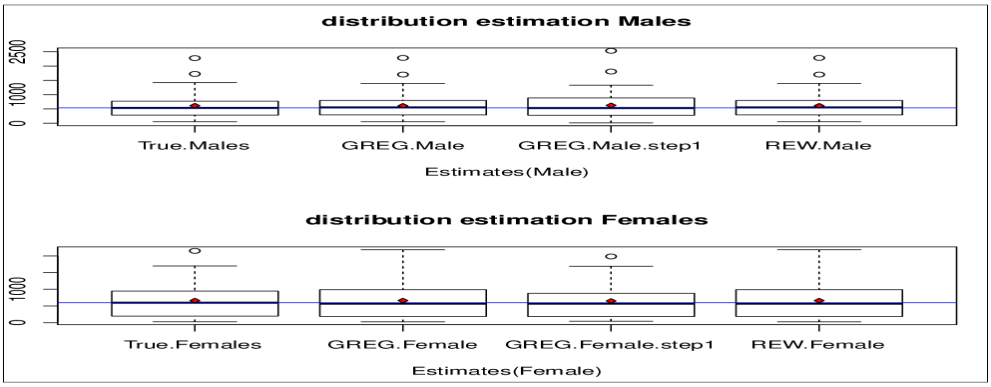

FIGURE 4.10. Consistency of the estimates

4.2 RESULTS OF MODEL-BASED PROJECTION ONTO THE REGISTER

In this Section, we report the results of the model-based projection estimator from the small sample, containing both target and auxiliary information, onto the register, i.e. the data set from the Population Census, which contains only observed values of the auxiliary variables.

Synthetic values were projected from the sample S_2 to the register R, using both unweighted and weighted model parameter estimates. As seen in the Sections above, this results in computing EBLUP and pseudo–EBLUP predictors respectively, using, as the known population means, the corresponding values evaluated on the register R. Therefore, the analysis compares the performances of the EBLUP and the pseudo–EBLUP on the three model specifications, MOD1, MOD2 and MOD3. Furthermore, the estimates deriving from the EBLUP and pseudo–EBLUP predictors, and the corresponding synthetic predictors, obtained without considering the estimates of the area random effect $\hat{v}_d$, were also considered.

Table 4.6 below reports the performances of the estimation methods under comparison in the three model settings, in terms of $AARE$, ASE and RE.

Model	Synthetic	EBLUP	pseudo–Synthetic	pseudo–EBLUP
MOD1	1.125	0.686	1.125	0.747
MOD2	0.081	0.080	0.081	0.080
MOD3	0.344	0.175	0.344	0.174

TABLE 4.6: Averaged Absolute Relative Error (AARE)

Model	Synthetic	EBLUP	pseudo–Synthetic	pseudo–EBLUP
MOD1	5312366.22542	3.969974e+13	5312354.41170	4411359.41765
MOD2	68848.63275	66890.07000	68864.82389	66816.12809
MOD3	594370.60921	238516.90000	594651.50062	237822.44903

TABLE 4.7. Averaged Squared Error (ASE)

Model	Synthetic	EBLUP	pseudo–Synthetic	pseudo–EBLUP
MOD1	1.015	0.788	1.015	0.845
MOD2	0.877	0.880	0.877	0.880
MOD3	0.825	0.800	0.825	0.806

TABLE 4.8. Relative Error (RE)

The results displayed in the tables above show that increases in the predictive power of the model are matched by improvements in the performance of the estimation methods. Furthermore, the comparison of the performances under the models MOD2 and MOD3 highlights that the differences between the performance of the synthetic and the composite predictors tend to disappear when the quality of the model improves. In this case, the information provided by the auxiliary variables can give good results, even without introducing extra variability between areas.

Plots 4.11, 4.12, 4.13 and 4.14 below display the distribution of the four estimation methods under the three model settings, with respect to the distribution of true area mean values.

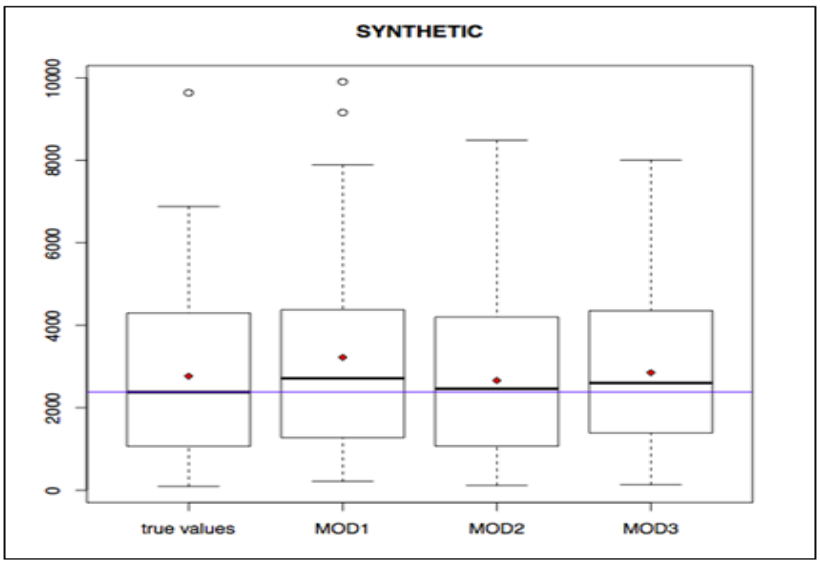

FIGURE 4.11. Distribution of synthetic estimates

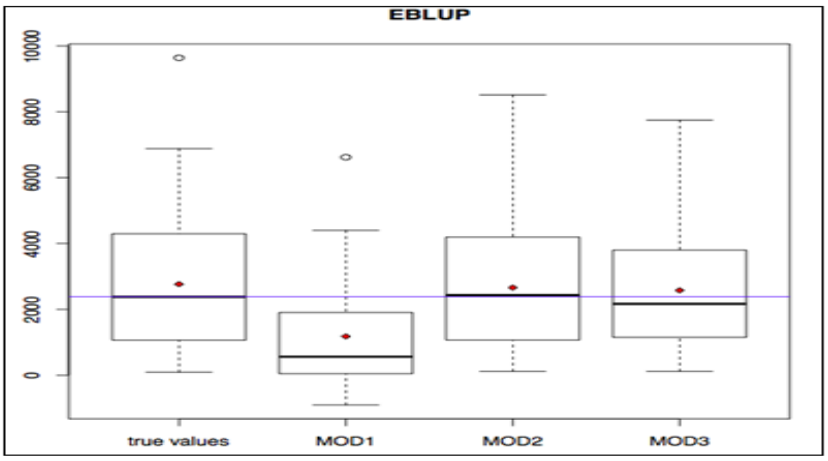

FIGURE 4.12. Distribution of EBLUP estimates

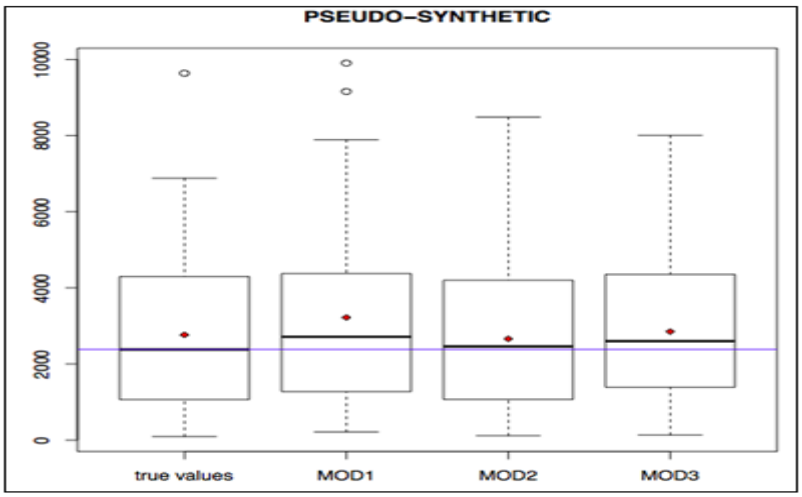

FIGURE 4.13. Distribution of pseudo–synthetic estimates

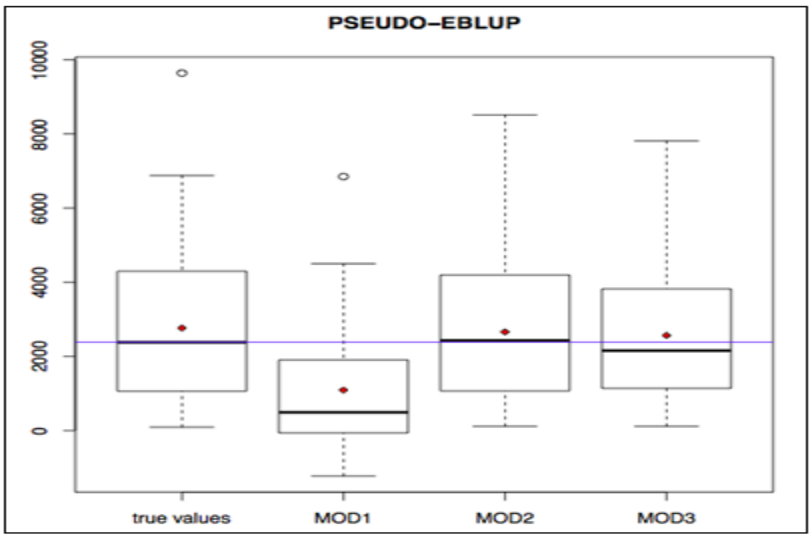

FIGURE 4.14. Distribution of pseudo–EBLUP estimates

Figures 4.15, 4.16, 4.17 and 4.18 below show the estimated coefficient of the variations for all estimation methods. The estimates of the coefficients of variation related to Model MOD1 are not included in the plots; due to their high values, they would make it difficult to see the distribution of the coefficient of variations for the other two models.

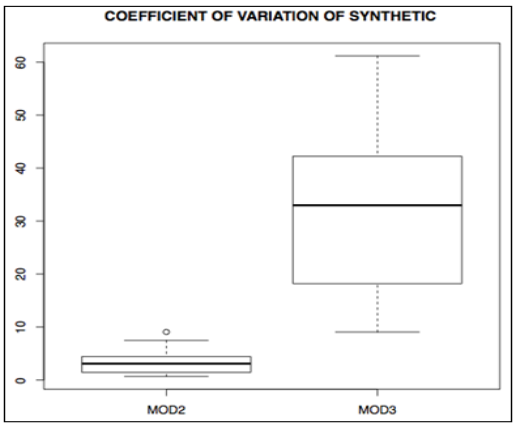

FIGURE 4.15. Distribution of the coefficient of variations of the synthetic estimates

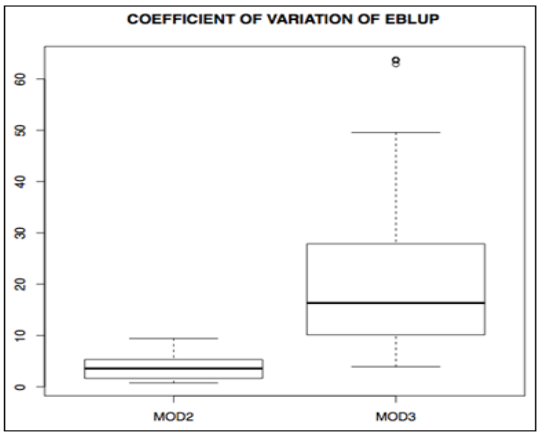

FIGURE 4.16. Distribution of the coefficient of variations of EBLUP estimates

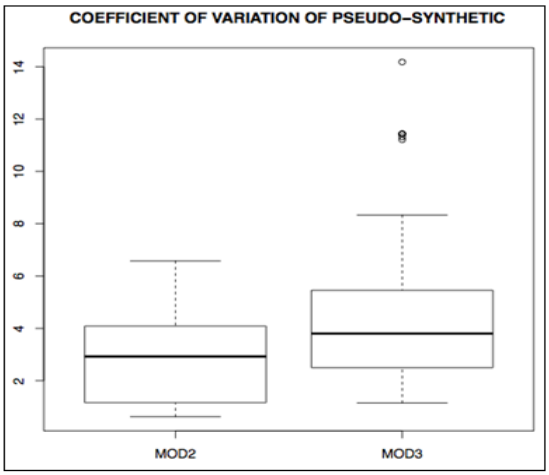

FIGURE 4.17. Distribution of the coefficient of variations of pseudo–synthetic estimates

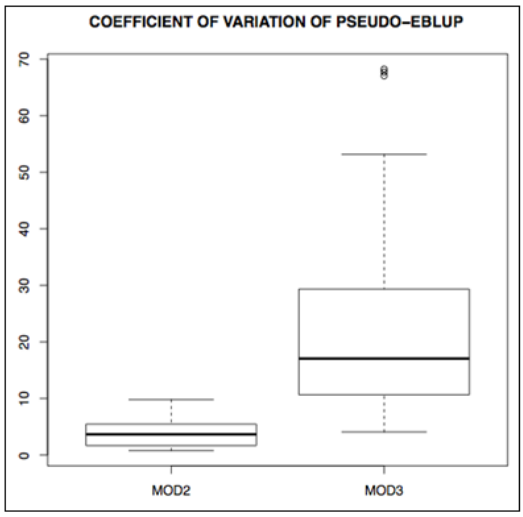

FIGURE 4.18. Distribution of the coefficient of variations of pseudo–EBLUP estimates

As would be expected, the errors associated with the more predictive Model *MOD2* are much smaller than the errors associated to MOD3.

Figures 4.19, 4.20 and 4.21 below provide an evaluation of the three models that display the distribution of the estimation methods under models MOD1, MOD2 and MOD3.

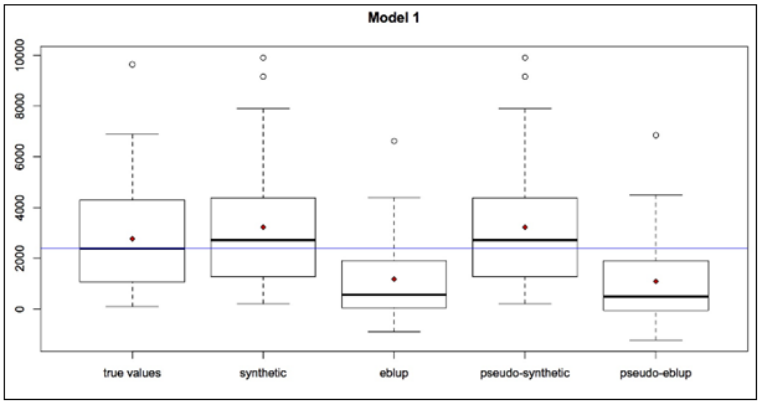

FIGURE 4.19. Distribution of the estimates under Model MOD1

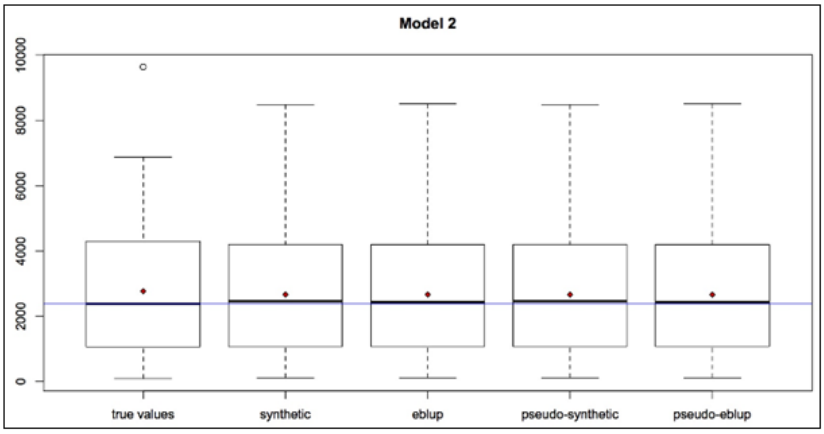

FIGURE 4.20. Distribution of the estimates under Model MOD2

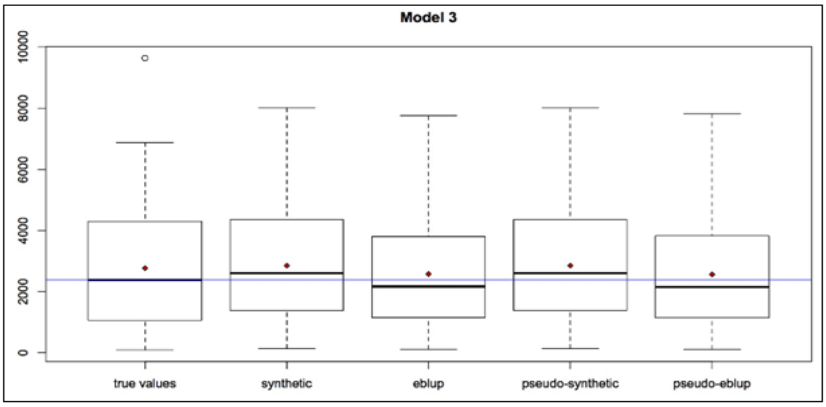

FIGURE 4.21. Distribution of the estimates under Model MOD3

Highlights

This case study illustrates the performance of design- and model-based projection estimators, with respect to standard direct estimates. The tables and the plots display how the three classes of estimators perform under the three model specifications defined in the case study. Model MOD1 is a model specification with poor predictive power, while the predictive power of models MOD2 and MOD3 can be classified as high and medium, respectively.

The results show that both design- and model-based projection estimators outperform the direct estimator. As for the comparison between design- and model-based projection estimators, the former gives better results with the worse-performing Model MOD1, while the latter outperforms the competitor estimator when applied to the best-performing Model MOD2. The performance of the two estimators appears to be similar in Model MOD3; therefore, in this case, it is not possible to draw any definitive conclusions from this case study.

References

Alleva, G. 2012. *La qualità dell'informazione statistica nell'era del digitale*, Paper prepared for the *Seconda Giornata Nazionale della Statistica*, 23 October 2012. Rome, Istituto nazionale di statistica (Istat).

Ardilly, P. & Le Blanc, D. 2001. Sampling and Weighting a Survey of Homeless Persons: A French Example. *Survey Methodology*, 27(1): 109-118.

ARF. 2003. *ARF guidelines for data integration*, ARF, New York.

Bankier, M. 1988. Power allocation: determining sample sizes for sub-national areas, *The American Statistician*, 42: 174–177.

Battese, G. E., Harter, R. M., & Fuller, W. A. 1988. An error-components model for prediction of county crop areas using survey and satellite data, *Journal of the American Statistical Association*, 83: 2836.

Belin, T. R. & Rubin, D. B. 1995. A method for calibrating false-match rates in record linkage, *Journal of the American Statistical Association*, 90: 694-707.

Bethel, J. 1989. Sample allocation in multivariate surveys, *Survey Methodology*, 15: 47–57.

Bryant, E., Hartley, H. & Jessen, R. 1960. Design and estimation in two-way stratification, *Journal of the American Statistical Association*, 55: 105–124.

Burkard, R. & Derigs, U. 1981. Assignments and matching problems: solution methods with fortran programs. In *Lecture Notes in Economics and Mathematical Systems*, No. 184 (pp. 1-11). New York: Springer Verlag.

Canada. Statistics Canada. 2008. *Policy on Record Linkage*. Statistics Canada Policy Manual. Available at: http://icn-rci.statcan.ca/10/10c/10c_025_e.htm.

_____ 2009a. *Quality Guidelines*, 5th ed., Statistics Canada.

_____ 2009b. *Use of Administrative Data portal*. Available at http://www.statcan.gc.ca/pub/12-539-x/2009001/administrative-administratives-eng.htm..

Carletto, C., Jolliffe, D., & Banerjee, R. 2013. *African economic development: Measuring success and failure*, Vancouver, Canada, 18-20 April 2013. (Conference paper)

Causey, B., Cox, L. & Ernst, L. 1985. Applications of transportation theory to statistical problems, *Journal of the American Statistical Association*, 80: 903–909.

Chambers, R. 2009. *Regression Analysis of Probability-Linked Data*, Official Statistics Research Series (Statistics New Zealand), 4. Available at http://www.statisphere.govt.nz/official-statistics-research/series/default.htm. Accessed on 19 June 2014.

Chatterjee S. 1967. A note on optimum allocation, *Skandinavisk Aktuarietidskrift*, 50: 40–44.

Chatterjee S. 1968. Multivariate stratified surveys, *Journal of the American Statistical Association*, 63: 530–534.

Chaudhuri A. 2010. Estimation with inadequate frames. *In* R. Benedetti, M. Bee, G. Espa,

Chesher, A., Nesheim, L. 2004. *Review of the Literature on the Statistical Properties of Linked Datasets*, London, Report to the Department of Trade and Industry.

Choudhry, G.H., Rao, J.N.K. & Hidiroglou M.A. 2012. On sample allocation for efficient domain estimation, *Survey Methodology*, 18: 23–29.

Chromy, J. 1987. Design optimization with multiple objectives. In *Proceedings of the American Statistical Association, Section on Survey Methods Research* (pp. 194–199).

Cochran, W. 1977. *Sampling Techniques*. Wiley: New York, USA.

Colledge, M.J. 1999. Statistical Integration through Metadata Management, *International Statistical Review / Revue Internationale de Statistique*, 67(1): 79-98.

Cook, R.D. 1977. Detection of Influential Observations in Linear Regression. Technometrics, *American Statistical Association* 19(1): 15–18.

Copas, J. & Hilton, F. 1990. Record linkage: statistical models for matching computer records, *Journal of the Royal Statistical Society*, Series A, 153: 287–320.

Costa, A., Satorra, A. & Ventura, E. 2004. Using composite estimators to improve both domain and total area estimation, *SORT*, 28: 69–86.

Cox, L.H., Boruch, R.F. 1988. Record linkage, privacy and statistical policy. *Journal of Official Statistics*, 4(1): 3-16.

Dalenius, T. 1953. The multivariate sampling problem, *Skandinavisk Aktuarietidskrift*, 36: 92–102.

Dalenius, T. 1957. Sampling in Sweden: Contributions to the Methods and Theories of Sample Survey Practice, Stockholm, *Almqvist och Wiksel*.

Data Documentation Initiative (DDI). 2008. *Technical Specification, Part I: Overview*, Version 3.0, DD Alliance.

Dempster, A., Laird, N. & Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, Series B, 39: 2–38.

Deville, J.C. & Lavallée, P. 2006. Indirect sampling: The foundations of the generalized weight share method, *Survey Methodology*, 32: 165–176.

Deville, J.C. & Tillé, Y. 2004. Efficient balanced sampling: the cube method, *Biometrika*, 91: 893-912.

Deville, J.C. & Tillé, Y. 2005. Variance approximation under balanced sampling, *Journal of Statistical Planning and Inference*, 128: 569–591.

Deville, J.C. & Maumy-Bertrand, M. 2006. Extension of the Indirect Sampling Method and its Application to Tourism. *Survey Methodology*. Statistics Canada. 32(2): 177-185.

Elbers, C., Lanjouw, J. & Lanjouw, P. 2003. Micro-Level Estimation of Poverty and Inequality, *Econometrica,* 71(1): 355-364.

Eurostat. 2003. Working Group on "Assessment of quality in statistics", Proceedings of Sixth Meeting, Luxembourg.

_____ 2003. *Quality assessment of administrative data for statistical purposes*, Eurostat Publication, Luxembourg.

_____ 2005. *Standard Quality Indicators, Working Group "Quality in statistics"*, Eurostat Publication, Luxembourg.

_____ 2007. *Handbook on Data Quality Assessment Methods and Tools*, Eurostat Publiacation, Wiesbaden, Germany. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/HANDBOOK%20ON%20DATA%20QUALITY%20ASSESSMENT%20METHODS%20AND%20TOOLS%20%20I.pdf.

_____ 2007. Handbook on Data Quality, Assessment Methods and Tools, Eurostat Publication, Luxembourg.

_____ 2011. The European Statistics Code of Practice For the National and Community Statistical Authorities, Adopted by the ESS Committee on 28 September 2011.

Falorsi, P.D., Pallara, A. & Russo, A. 2005. *L'integrazione di dati di fonti diverse, Tecniche e applicazioni del Record Linkage e Metodi di stima basati sull'uso congiunto di fonti statistiche e amministrative*, Franco Angeli, Milan, Italy.

Falorsi, P.D. & Righi, P. 2008. A balanced sampling approach for multi-way stratification designs for small area estimation, *Survey Methodology*, 34: 223–234.

_____ 2012. A Unified Approach for Defining Optimal Multivariate and Multi-Domains Sampling Designs. In: *Electronic proceedings of the 46th Scientific Meeting of the Italian Statistical Society*, Meeting held in Rome by the *Società Italiana di Statistica*, on 20 June 2012. Document available at: http://meetings.sis-statistica.org/index.php/sm/sm2012/schedConf/presentations.

Falorsi, P.D. 2013. *Research Plan – Integrated Survey Framework for Agricultural Statistics*. Unpublished internal document, FAO.

FAO. 1983. *Use of household surveys for collection of food and agricultural statistics*. FAO Economic and Social Development Paper, FAO Publication, Rome.

_____ 1986. *Food and Agricultural Statistics in the context in the Context of a National Information System*. FAO Publication, Rome.

_____ 2003. *Agri-environmental information and decision support tools for sustainable development*. Paper prepared for the 17th Session of the Committee on Agriculture, FAO, Rome. Available at: www.fao.org/DOCREP/MEETING/005/Y8343e.html..

_____ 2010. *Economic and Sector Work. Global Strategy to improve Agricultural and Rural Statistics*, Report No. 56719-GLB. FAO Publication, Rome.

_____ 2012. *Action Plan of the Global Strategy to improve Agricultural and Rural Statistics. For Food Security, Sustainable Agriculture and Rural Development*. FAO Publication, Rome.

_____ 2012. *Guidelines for Linking Population and Housing Censuses with Agricultural Censuses with selected country practices*. Special Issue of the FAO Statistical Development Series. FAO Publication, Rome.

_____ 2013. *FAO Statistics Quality Assurance Framework*. FAO Publication, Rome. Available at: http://www.fao.org/docrep/019/i3664e/i3664e.pdf..

_____ Various years. *A system of integrated agricultural censuses and surveys*. FAO Publication, Rome.

_____ Various years. *Global strategy to improve agricultural and rural statistics, Technical report, United Nations Economic and Social Council Statistical Commission*. FAO Publication, Rome.

Faulkenberry G.D. & Garoui, A. 1991. Estimating a Population Total Using an Area Frame, *Journal of the American Statistical Association*, 86: 414.

Fellegi, I. & Sunter, A. 1969. A theory of record linkage, *Journal of the American Statistical Association*, 64: 1183–1210.

Fortini, M., Liseo, B., Nuccitelli, A. & Scanu, M. 2001. On Bayesian record linkage, *Research in Official Statistics*, 4: 185–198.

Freund, R. & Wilson, W. 1998. *Statistical Methods IM*, Academic Press Inc., San Diego.

Gartner Inc. 2013. *Data integration*, Gartner Publication.

Gill, L. 2001. Methods for Automatic Record Matching and Linkage and their use in National Statistics, *National Statistics Methodological Series* No. 25, National Statistics, London.

Goldstein, H., Harron, K. & Wade A. 2012. The analysis of record-linked data using multiple imputation with data value priors, *Statistics in Medicine*, 31(28): 3481–3493.

Goodman, R. & Kish, L. 1950. Controlled selection - a technique in probability sampling, *Journal of the American Statistical Association*, 45: 350–372.

Green, P.J. & Mardia, K.V. 2006. Bayesian alignment using hierarchical models, with application in protein bioinformatics, *Biometrika*, 93: 235–254.

Guptill, S.C. 1999. Metadata and data catalogues. In Longley P.A., Goodchild M.F., Rhind D.W., & Maguire D.J. (eds) *Geographical Information Systems: Principles and Applications* (677-692). John Wiley and Sons, New York.

Gutman, R., Afendulis, C.C. & Zaslavsky, A.M. 2013. A Bayesian procedure for file linking to analyze end-of-life medical costs, *Journal of the American Statistical Association*, 108(501): 34–47.

Gutu, S.Z. 2001. Developing agricultural statistics within the overall national statistical systems. Paper presented at the *Workshop on Strengthening Food and Agricultural Statistics in Africa*, 22-26 November 2001, South Africa.

Hall, R., Steorts, R.C. & Fienberg, S.E. 2013. Bayesian parametric and nonparametric inference for multiple record linkage, *Working paper*, Carnegie Mellon University.

Hartley, H.O. 1959. Analytic Studies of Survey Data, in *A Volume in Honor of Corrado Gini* (page numbers). Rome: Istituto nazionale di statistica.

Hof, M. & Zwinderman, A. 2012. Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables, *Statistics in Medicine*, 31(30): 4231– 4242.

Hung, T. & Yasuoka, Y. 2000. Integration and application of socio-economic and environmental data within GIS for development study in Thailand. *AARS*. Available at: www.GISdevelopment.net

Hwang, D., Rust, A.G., Ramsey, S., Smith, J.J, Leslie, D.M., Weston, A.D., Pedro, A. et al. 2004. Data integration issues for a farm decision support system, *Transactions in GIS*, 8(4): 459-477.

IMF. 2012. The generic IMF Data Quality Assessment Framework (DQAF), Dissemination Standards Bulletin Board, *Data Quality Reference Site* (DQRS).

Ireland. Central Statistics Office (CSO). 2006. Standards and Guidelines, Vol. 1, Quality in Statistics, Quality Assurance & Audit Section, Version 1.0. *Research Series*, Vol. 4. Available from the CSO website: http://www.cso.ie .

Jaro, M. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, *Journal of the American Statistical Association*, 84: 414– 420.

_____ 1995. "Probabilistic Linkage of Large Public Health Data Files", *Statistics In Medicine*, 14: 491–498.

Kasnakoglu, H. & Mayo, R. 2004. *The main FAO Statistical Data Quality Framework, A multi-layered approach to monitoring and assessment*, Paper prepared for the Conference on Data Quality for International Organizations, 27-28 May 2004. Wiesbaden, Germany.

Keita, N. 2004. Improving Cost-Effectiveness and Relevance of Agricultural Censuses in Africa: Linking Population and Agricultural Censuses. FAO Publication: Rome. Available at: http://www.siea.sagarpa.gob.mx/mexsai/trabajos/t32.pdf..

Keita, N. 2007. International Conference on Agricultural Statistics IV (ICAS-IV). Agricultural and Rural Statistical Development – Capacity Building Session 3, Asia and Pacific Commission on Agricultural Statistics Twenty-second session. Beijing, China

Khan, M.G.M., Mati, T. & Ahsan, M.J. 2010. An optimal multivariate stratified sampling design using auxiliary information: An integer solution using goal programming approach, *Journal of Official Statistics*, 26: 695–708.

Kim, J.K. & Rao, J.N.K. 2012. Combining data from two independent surveys: a model assisted approach, *Biometrika*, 99(1): 85-100

Kim, G. & Chambers, R. 2009. *Regression analysis under incomplete linkage*, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper No. 17-09. Available at: http://ro.uow.edu.au/cssmwp/37,.

Klosterman, R.E. 1995. The appropriateness of geographic information systems for regional planning in the developing world. *Computer, Environment and Urban Systems*, 19(1): 1-13.

Knotterus, P. & van Duin, C. 2006. Variances in repeated Weighting with an Application to the Dutch Labour Force Survey. *Journal of Official Statistics*, 22(3): 565-584.

Koch, C. 2001. Data Integration against Multiple Evolving Autonomous Schemata. University of Vienna, Austria. (Ph.D. Thesis)

Kokan, A. & Khan, S. 1967. Optimum allocation in multivariate surveys: An analytical solution, *Journal of the Royal Statistical Society*, Series B, 29: 115–125.

Kovacevic, M. 1999. Record linkage and statistical matching - they aren't the same! *SSC Liaison*. 13(3): 24-29.

Kyeyago Ouma, F., Muwanga Zake, E. & Mayinza, S. 2010. *In the Construction of an International Agricultural Data Quality Assessment Framework (ADQAF)*, Paper prepared for the *The Fifth International Conference On Agricultural Statistics* (ICAS V), 13 October 2010. Uganda Bureau of Statistics? , Kampala.

Lahiri, P. & Larsen, M.D. 2005. Regression analysis with linked data. *Journal of the American Statistical Association*, 100: 222-230.

Larsen, M. 1999. Multiple imputation analysis of records linked using mixture model. In: *Proceedings of Survey Methods Section* (65-71). Regina, Canada, Statistical Society of Canada.

Larsen, M.D. & Rubin, D. 2001. Iterative automated record linkage using mixture models, *Journal of the American Statistical Association*, 96: 32–41.

Larsen, M. 2005. Advances in record linkage theory: Hierarchical Bayesian record linkage theory, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3277–3283.

Lavallée, P. & Caron, P. 2001. Estimation Using the Generalized Weight Share Method: The Case of Record Linkage. *Survey Methodology*. Statistics Canada, Catalogue n. 12-001, 27(2) : 155-169.

Lavallée, P. & Deville, J.C. 2006. Indirect Sampling: The Foundation of Generalized Weight Share Method. *Survey Methodology*. Statistics Canada, Catalogue n. 12-001-XPB, 32(2): 165-176.

Lavallée, P. 2007. *Indirect Sampling*. Springer: Ottawa.

Lavallée, P. & Rivest, L.P. 2012. Capture-recapture sampling and indirect sampling. *Journal of Official Statistics*, 28: 1-27.

Lavallée, P. & Labelle-Blanchet, S. 2013. Indirect Sampling applied to Skewed Populations. *Survey Methodology* 39: 183-215.

Lenzerini, M. 2002. Data Integration: A Theoretical Perspective. *PODS* 2002: 243-246.

Lindley D. 1977. A problem in forensic science, *Biometrika*, 64: 207–213.

Liseo, B., Montanari, G.E. & Torelli, N. 2006. *Metodi statistici per l'integrazione di dati da fonti diverse*. Franco Angeli: Milan.

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

_____ 2011b. Some advances on Bayesian record linkage and inference for linked data. In: *Proceedings of the ESSnet Data Integration Workshop*, Madrid, 24-25 November 2011.

Lu, W. & Sitter, R.R. 2002. Multi-way stratification by linear programming made practical, *Survey Methodology*, 2: 223–234.

Meta Group. 2004. *The future of data integration technologies. A Meta Group White Paper*. Available at: www.metagroup.com/www.synopsis.com..

Mecatti, F. 2007. A single frame multiplicity estimator for multiple frame surveys. *Survey Methodology*, 33(2): 151-157.

Mogoa, E.G.M. & Nyangito, M.M. 1999. Constraints to the delivery of animal health services in pastoral areas of Kenya: A review. *The African Pastoral Forum,* Working Paper No. 20.

Mireku Kwakye, M. 2011. A Practical Approach To Merging Multidimensional Data Models. University of Ottawa. (Ph.D. Thesis).

Nedyalkova, D. & Tillé, Y. 2008. Optimal sampling and estimation strategies under the linear model, *Biometrika*, 95: 521–537.

Neter, J., Maynes, E.S. & Ramanathan, R. 1965. The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60(312): 1005-1027.

Newcombe, H., Kennedy, J., Axford, S., & James, A. 1959. Automatic Linkage of Vital Records. *Science*, 130: 954–959.

Neyman, J. 1934. On the two different aspects of the representative methods: The method stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society*. 97: 558–606.

Oluoch-Kosura, W. 2007. What does integration imply in choosing a unit of enumeration: enterprise, holding or individual? Does it matter? Perspectives from Africa. Paper prepared for the *Fourth International Conference on Agriculture Statistics* (ICAS-4), October 2007, Beijing, China.

Openshaw, S. 1983. The modifiable areal unit problem. *Concepts and Techniques in Modern Geography*, 38.

Piersimeoni, F., Benedetti, R., Bee, M. & Espa, G. 2010. *Agricultural Survey Methods.*, Wiley, New York.

Polach, R. & Rodgers, M. 2006. *The importance of data integration*. IIM National. Available at: www.iim.org.au/national/htm/default.cfm..

Priest, G. 2010. The Struggle for Integration and Harmonization of Social Statistics in a Statistical Agency: A Case Study of Statistics Canada, *IHSN Working Paper* No. 004.

Rao, J.N.K. 2003. *Small Area Estimation*. Hoboken, USA: John Wiley & Sons.

Ray, S.S. et al. 2009. Combining Multi-Source Information through Functional Annotation based Weighting: Gene Function Prediction in Yeast. *IEEE Transactions on Biomedical Engineering* 56(2): 229–236.

Rodgers, D., Emwanu, T. & Robinson, T. 2006. Mapping poverty in Uganda using socioeconomic, environmental and satellite data. FAO Publication, Rome. Available at: http://www.fao.org/ag/pplpi.html.

Saei, A. & Chambers, R. 2003. Small area estimation under linear and generalized linear model with time and area effects, *Working Paper* M03/15, Southampton Statistical Sciences Research Institute, University of Southampton.

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

Scanu, M. (ed). 2008. *Metodi statistici per il record linkage*, Istat: Rome.

Scheuren, F. & Winkler, W.E. 1993. Regression analysis of data files that are computer matched. *Survey Methodology*, 19: 39–58.

Scheuren, F. & Winkler, W.E. 1997. Regression analysis of data files that are computer matched, Part II. *Survey Methodology* 23: 157–165.

Schuldt, R. & Shauger, R.L. 2011. *UDEF – Six Steps to Cost Effective Data Integration*. CreateSpace Independent Publishing Platform.

Singh, A.C. & Mecatti, F. 2011. Generalized Multiplicity-Adjusted Horvitz-Thompson Estimation as a Unified Approach to Multiple Frame Surveys. *Journal of Official Statistics*, 27(4): 633–650.

Sitter, R.R. & Skinner, C.J. 1994. Multi-way stratification by linear programming. *Survey Methodology*, 20: 65–73.

Statistics New Zealand. 2005a. *Data Integration Policy Guidelines*, Statistics New Zealand.

_____ 2006. *Data Integration Manual*, Statistics New Zealand.

Sverdrup, U. 2005. *Administering information: Eurostat and statistical integration*. Working Paper No. 27, Centre for European Studies, University of Oslo, Norway.

Tancredi, A., Guagnano, G. & Liseo, B. 2005. Inferenza statistica basata su dati prodotti mediante procedure di record linkage. In: *L'integrazione di dati di fonti diverse: tecniche e applicazioni del Record Linkage e metodi di stima basati sull'uso congiunto di fonti statistiche e amministrative* (41-59). P.D. Falorsi, A. Pallara & A. Russo eds. Franco Angeli, Milan, Italy.

Tancredi, A. & Liseo, B. 2011. A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems. *The Annals of Applied Statistics*, 5(2B): 1553-1585.

Tadingar, T. 1994. Pastoral development in Sub-Saharan Africa: An integration of modern and indigenous technical knowledge. *The African Pastoral Forum, Working Paper* No. 2.

Ullman, J.D. 1997. Information Integration Using Logical Views. *ICDT* 1997: 19–40.

United Kingdom, National Statistics. 2006. *Guidelines for measuring statistical quality*, Office for National Statistics, UK.

_____ 2008. *Code of Practice, Protocol on Statistical Integration and Classification*, 2nd ed., National Statistics, UK.

_____ 2009. Code of Practice 2009. Available at http://www.statisticsauthority.gov.uk/assessment/code-of-practice/.

UN. 1994. Fundamental Principles of Official Statistics, Adopted by The United Nations Statistical Commission, in its Special Session of 11-15 April 1994.

____ 1999. *Information Systems Architecture for National and International Statistical Offices – Guidelines and Recommendations*, United Nations, 1999UNECE (2000), Assessment of the quality in statistics, Definition of quality in statistics, Luxembourg - 4-5 April 2000.

UNECE. 2009. Generic Statistical Business Process Model (Version 4.0 – April 2009), Joint UNECE/Eurostat/OECD *Work Session on Statistical Metadata* (METIS), Prepared by the UNECE Secretariat.

_____ 2010. Statistical Data Quality in the UNECE, 2010 Version, Steven Vale, Statistical Division.

United Nations Statistics Division. 2008. Principles and recommendations for

population and housing censuses, Revision 2. *Statistical Papers* Series M, No. 67/Rev.2.

United Nations Statistics Division. 2013. Guidelines on Integrated Economic Statistics, New York, USA, http://unstats.un.org/unsd/nationalaccount/docs/IES-Guidelines-e.pdf

United States of America. U.S. Office of Management and Budget. 2002. Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies: 8452-8460.

United States of America. Census Bureau, Office of Management and Budget. 2006. *Standards and Guidelines for Statistical Surveys, September 2006*.

United States of America, U.S. Census Bureau. 2009. Policy Statement On Record Linkage. Available at https://www.census.gov/foia/pdf/ds014.pdf. Accessed on day month year.

_____ 2009. Statistical Quality Standard C4: Linking Data Records.

White, C. 2005. Data integration: Using ETL, EAI, And EII tools to create an integrated enterprise, *BI Research*, http://www.tdwi.org.

Winkler, W.E. 1988. Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. In *Proceedings of Survey Research Methods Section, American Statistical Association*, 667–671.

_____ 1993. Improved decision rules in the Fellegi-Sunter model of record linkage, in: *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 274–279.

_____ 1995. Matching and record linkage, in: *Buisness Survey Methods*, Wiley.

_____ 2001. Multi-way survey stratification and sampling, *Technical report*, U.S. Bureau of the Census, Statistical Research Division.

_____ 2009. Sample allocation and stratification, *Technical report*, U.S. Bureau of the Census, Statistical Research Division.

Wolter, K.M. 1986. Some coverage error models for census data, *Journal of the American Statistical Association*, 81: 338–345.

World Bank, FAO & UN. 2010. Global Strategy to improve agricultural and rural statistics, Report No. 56719-GLB.

You, Y. & Rao, J.N.K. 2002. A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights, *The Canadian Journal of Statistics*, 30(3), pp. 565-584.