Technical Report Series GO-24-2017

Technical Report on Reconciling Data from Agricultural Censuses and Surveys

July 2017

Technical Report on Reconciling Data from Agricultural Censuses and Surveys

Table of Contents

Acronyms and Abbreviations

DGESS	Directorate-General of Studies and Statistics of the MAAH
FAO	Food and Agriculture Organization of the United Nations
GEOSTAT	National Statistics Office of Georgia
MAAH	Ministry for Agriculture, Hydraulics and Halieutics Resources of Burkina Faso
MSF	Master Sampling Frame
PELT	Pruned Exact Linear Time

List of Boxes

List of Figures

List of Tables

Glossary of main technical terms

Administrative data: data holdings containing information that is collected primarily for administrative (not statistical) purposes, by government departments and other organizations, usually during the delivery of a service or for the purpose of registration, record-keeping or documentation of a transaction (Global Strategy 2015).

Agricultural holder: civil person, group of civil persons or legal person who makes the main decisions regarding resource use, and exercises management control over the operation of the agricultural holding. The agricultural holder bears technical and economic responsibility for the holding, and may undertake all responsibilities directly or delegate those relating to day-to-day work management to a hired manager (FAO 2015).

Agricultural holding: economic unit of agricultural production under a single management that comprises all livestock kept and all land used wholly or partly for agricultural production purposes, regardless of title, legal form or size. Single management may be exercised by an individual or household, jointly by two or more individuals or households, by a clan or tribe, or by a legal person, such as a corporation, cooperative or government agency. The holding's land may consist of one or more parcels, which may be located in one or more separate areas or in one or more territorial or administrative divisions, provided that they share the same means of production, such as labour, farm buildings, machinery or draught animals (FAO 2015).

Agricultural sample survey: agricultural survey for which the inference procedure to estimate each survey variable for the total survey area is based on the values of the variable obtained from a sample of reporting units (FAO 1996).

Census of agriculture or agricultural census: statistical operation for collecting, processing and disseminating data on the structure of agriculture, covering the whole or a significant part of a country. Typical structural data collected in a census of agriculture are size of holding, land tenure, land use, crop area, irrigation, livestock numbers, labour and other agricultural inputs. In an agricultural census, data are collected at the holding level, although some community-level data may also be collected (FAO 2015).

Data reconciliation: methodology that uses process information and mathematical methods to correct measurements, focusing on data integrity and quality. The reconciliation between census data and survey data focuses on resolving inconsistencies in the data time series.

Enumeration area: small geographical units defined for the purposes of census enumeration (FAO 2015).

Equal Probability Selection Method: sample selection in which every sampling unit has the same probability of being selected for the sample.

List frame: in agricultural statistics, list frames are lists of farms and/or households obtained from agricultural or population censuses and/or administrative data. The ultimate sampling units are lists of names of holders or households (Global Strategy 2015).

Livestock: all animals, birds and insects kept or reared in captivity mainly for agricultural purposes. This includes, among others, cattle, buffalo, horses and other equines, camels, sheep, goats and pigs, as well as poultry, bees and silkworms. Aquatic animals do not fall under this definition. Domestic animals, such as cats and dogs, are excluded unless they are being raised for food or other agricultural purposes (FAO 2015).

Multiple frame: a combination of the list and area frames.

Non-probability or subjective sample survey: an agricultural sample survey for which the inference procedure to obtain estimates of the desired variables is not based on probability sampling and estimation methods.

Primary Sample Unit: in multiple-stage sampling, a sample unit at the first stage of selection.

Probability sample survey: sample survey for which the inference procedure to obtain estimates of the survey variables is based on probability sampling and estimation methods. In a probability sample survey, it is possible to establish the estimates' statistical precision.

Register: a complete list of objects belonging to a defined object set. The objects in the register are identified by means of identification variables, which make it possible to update the register and link it with other registers (Turtoi *et al*. 2012)

Sample selection with probability-proportional-to-size measure: sampling procedure in which the probability of selection of a sampling unit is proportional to its assigned size, called the measure of size.

Sampling frame: total set of sampling units and their probabilities of selection. More specifically, the list of sampling units from which the sample is selected, together with each of their probabilities of selection. A sample selection method should be adopted allows for the determination of the probability of including each unit. In conducting the survey, the probabilities of selection should be maintained. The inverses of the selection probabilities are then used as weights to form the estimates.

Sampling plan or design: techniques for selecting a probability sample and estimation methods.

Secondary sample unit: in multiple-stage sampling, sample unit at the second stage of selection.

Statistical register: a data set with identifiers in which the object set and variables correspond to the statistical matter (Turtoi *et al.* 2012).

Statistical unit: elements of the population for which data should be collected during a survey; they are subject to inferences.

Stratification: division of the population into subsets, called strata. Within each stratum, an independent sample is selected. In stratified sampling, the survey population is subdivided into non-overlapping sets called strata. Each stratum is treated as a separate population.

System of registers: a number of registers that are linked to one another by means of one or more common identification variables or linkage variables. An efficient system requires good-quality linkage variables and the presence of the same linkage variables in different registers. Furthermore, the definitions of the objects and variables in the system must be harmonized, so that data from different registers can be used together. The reference times must also be consistent (Turtoi *et al.* 2012).

Preface

The Technical Report on Reconciling Data from Agricultural Censuses and Surveys has been prepared within the framework of the Global Strategy to Improve Agricultural and Rural Statistics (Global Strategy). The Global Strategy is an initiative that was endorsed in 2010 by the United Nations Statistical Commission. It provides a framework and a blueprint to meet current and emerging data requirements and the needs of policymakers and other data users. Its goal is to contribute to greater food security, reduced food price volatility, higher incomes and greater well-being for rural populations, through evidence-based policies. The Global Strategy is centred on three pillars: (1) establishing a minimum set of core data; (2) integrating agriculture into national statistical systems; and (3) fostering the sustainability of the statistical system through governance and statistical capacity building.

Methodological guidance in reconciling census and survey data has been requested by a number of member countries of the Food and Agriculture Organization of the United Nations (FAO). To respond to this request, the Global Strategy to Improve Agricultural and Rural Statistics has funded a line of research on this topic.

In the Action Plan to Implement the Global Strategy, the preparation of this technical report on reconciling data from agricultural censuses and surveys was prioritized to provide statisticians in countries with practical guidelines.

Methodologies presented in the present document have been presented and discussed at a dedicated expert meeting organized by the Global Office in Rome in September 2016 with international, regional and national experts. In addition, the members of the Scientific Advisory Committee1 of the Global Strategy reviewed and provided their feedback on the document. Detailed comments were made and have been used to improve the document.

The technical report introduces and discusses the problem of reconciling agricultural census and survey data by exploring methods to recalculate the survey weights, focusing on how to combine information from agricultural censuses. The aim is to offer users methodologies to resolve inconsistencies in the data time series.

[1] Michael Steiner, Cristiano Ferraz, V.K. Bhatia, Seghir Bouzzafour, Jacques Delincé, Eva Laczka, Dalisay "Dax" Maligalig, Backary Sacko, Edwin St. Catherine and Zhengyuan Zhu

Pilot tests have been done in two selected countries using real data from selected countries to assess the methodology and the workability of the methods: Burkina Faso and Georgia. For the desktop test in Burkina Faso, a memorandum has been signed between the FAO regional office in Africa and the Direction Générale des Etudes et des Statistiques Sectorielles (DGESS) du Ministère de l'Agriculture et des Aménagements Hydrauliques (MAAH). The institution, which is the data owner, has provided data for the writing of this report. The National Statistics Office of Georgia (GEOSTAT) has provided some datasets in order to pilot the methods.

This technical report is intended to serve as a reference document that provides technical and operational guidance on various aspects of the process of the reconciliation of data from agricultural censuses and surveys under different country conditions, with an emphasis on developing countries. The publication addresses a significant gap, as there is very little technical guidance on this topic.

The technical report recognizes the diversity of country situations and resources, and consequently, proposes various options. It is conceived as a living document to be subject to periodical review.

When necessary or considered desirable, the technical report refers readers to alternative more detailed methodological documents including the Literature Review on Reconciling Data from Agricultural Censuses and Surveys[2] published by the Global Strategy.

[2] http://gsars.org/wp-content/uploads/2016/07/Literature-Review-on-Reconciling-Data-from-Agricultural-Censuses-and-Surveys-200716.pdf

Acknowledgments

The Technical report on Reconciling Data from Agricultural Censuses and Surveys was prepared by Eloi Ouedraogo, FAO-RAF Regional Statistician, Ulrich Eschcol NYAMSI, Consultant Statistician at FAO-RAF within the Global Strategy, Lassina Pare, Director of DGESS, Audrier Sanou, officer at DGESS and Eric Kabore officer at DGESS. This work has been performed under the supervision of Flavio Bolliger, Research Coordinator of the Global Strategy and Dramane Bako, Consultant Statistician for the Global Strategy.

The authors wish to thank Naman Keita, International Senior Consultant, Linda J. Young, Chief Mathematical Statistician in the National Agricultural Statistics Service (United States) and Dramane Bako for their invaluable contributions.

The authors are also grateful to Vasil Tsakadze, officer at GEOSTAT for his collaboration and data sharing.

Valuable input and comments were provided at different stages by the Scientific Advisory Committee members and by participants of the expert meeting, which was held in Rome on September 2016[3].

[3] Zhengyuan Zhu (Iowa State University-United States), Emily Berg (Iowa State University-United States), Jongho Im (Iowa State University-USA), Leonard Atuhaire (Makerere University-Uganda), Agnes M.N. Ssekiboobo (Makerere University-Uganda), Abraham Yeyo Owino (Makerere University-Uganda), Linda Young (NASS-USA), Cynthia St-Germain (StatCan-Canada), Dalisay Samarita Maligalig (University of the Philippines at Los Banos), Lassina Paré (Agricultural Statistics Department-Burkina Faso), Titus Mwisomba (National Bureau of Statistics-Tanzania), Ndamona C. Kali (Namibian Statistics Agency), Anders Wallgren (International Expert), Moussa Kaboré (International Expert), Flavio Bolliger (FAO HQ, Global strategy), Naman Keita (FAO HQ, Global strategy), Carola Fabi (FAO HQ, Global Strategy), Valerie Bizier (FAO HQ, Global strategy), Dramane Bako (FAO HQ, Global strategy), Ulrich Nyamsi (FAO-RAF), Eloi Ouedraogo (FAO-RAF)

Executive Summary

The purpose of the Technical report on Reconciling Data from Agricultural Censuses and Surveys is to support statisticians in the process to reconcile data. The Action Plan to Implement the Global Strategy establishes a research programme to develop best methods for reconciling data from agricultural census and survey.

For many countries, a main source of agricultural statistics is the agricultural census, which is usually conducted every ten years. Those statistical operations when implemented as a complete enumeration provide a frame for the subsequent annual surveys. However, a problem associated with those censuses is that the data and farms/households listings become obsolete because of the long time span between collection periods. In some cases, several years may elapse before the census data and listings become available, which means that they are obsolete from the very beginning.

One of the main problems affecting the process of data reconciliation is that there is not enough auxiliary data are available. The addition of agricultural modules to population censuses, as recommended by FAO, can yield information that is very useful for reconciliation process.

A review of country practices also indicate that a careful analysis of the country situation is needed, in terms of resources, materials available, institutional support, the scope of the statistical system and objectives of the surveys to ensure that the options selected are suitable and sustainable.

Chapter 1 of the report reviews the main sources of discrepancies in the agricultural time series data identified in the literature review.

Chapter 2 presents the methods presented in the literature review and tested with data from Burkina Faso and Georgia. It also presents a clear explanation on how to implement these methods step by step.

In chapter 3, the main steps to be followed in the data reconciliation process are presented and possible strategies to prevent or minimize discrepancies in future censuses and survey operations are proposed.

1

Introduction

1.1. Background

A census of agriculture (or agricultural census) is a statistical operation aimed at collecting, processing and disseminating data on the structure of agriculture over the whole or a significant part of a country. Typical structural data collected in an agricultural census are the number and size of holdings (broken down by, for example, region, province, district or village), land tenure, land use, crop area harvested, irrigation, livestock numbers, labour and other agricultural inputs. In an agricultural census, data are collected directly from agricultural holdings, although some community-level data may also be collected. A census of agriculture usually entails collecting key structural data, by means of a complete enumeration of all agricultural holdings, and more detailed structural data, using surveys and sampling methods.

Data from agricultural censuses are useful in a variety of economic and social domains, including agricultural and rural-sector planning and policymaking, and for monitoring progress towards achieving the Sustainable Development Goals and dealing with problems relating to poverty, food security and gender. Agricultural census data are also used to establish agricultural indicator benchmarks and tools to assess and improve current agricultural statistics during intercensus periods. In several developing countries, agricultural data are derived mainly from decennial censuses, which provide structural data on agricultural holdings and benchmark data that serve as references for yearly estimates subsequently computed on the basis of sample surveys. Samples for current agricultural surveys are drawn from the sampling frame established for the most recent agricultural census, aiming to provide annual estimates on certain agricultural data items and variables, such as planted or harvested agricultural area, production and yield. Those annual estimates are based on the structure of agriculture identified in the latest census.

When a new census is conducted, discrepancies are often found between its results and the time series derived from the annual sample surveys conducted since the most recent census. Countries tend to encounter difficulties in reconciling crop or livestock data from the most recent agricultural census with the agricultural statistical series obtained from sample survey data. In some cases, there may be valid statistical reasons for these differences. Some of examples of this are: the geographic area covered by one of the collections may

be incomplete, as urban areas may have been excluded; certain types of holdings, such as smallholdings, may have been omitted from one of the collections; different concepts and definitions may have been applied in the treatment of mixed cropping; there may be inconsistencies in the reference periods or in the definition of crop seasons; subnational data may be inconsistent because the agricultural census collects data on the basis of the holder's place of abode, and not the location of the land or livestock; or if sampling is involved, the sample results may be affected by sampling errors. Those discrepancies easily arise when the inter-census period is excessively long.

Although this is a common problem, few studies and methodological guidance systematically address the issues arising after each census, even in countries with more advanced statistical systems.

The present document analyses the possible sources of discrepancy between the time series from intercensus annual surveys and the results of new censuses. It also reviews the statistical methods that can be applied to deal with those discrepancies, taking into account countries' experiences. In addition, the technical paper outlines possible strategies and methodological options to implement the systematic reconciliation of intercensus survey data with the results of new censuses. Simulations have been done in order to assess the methodologies. Furthermore, a pilot test has been carried out in Burkina Faso using real data and data from GEOSTAT.

1.2. Overview of the main sources of discrepancies

When the results of a new census are available, discrepancies may occur in the data, especially when the intercensal period is too long. In this context, the reconciliation of data can be an arduous task for several reasons, the most common of which are mentioned below.

a) Changes in the sampling frame

Measurements may be sought from agricultural holdings during annual surveys to take into account changes in the holdings' practices and therefore changes in the performance of the agricultural holdings sampled. However, if survey weights are not revised to capture the changes in the number of agricultural holdings and their distribution by size or strata, an inconsistency between data may result.

Box.1.1. Changes in the Sampling Frame − Brazil

During the 2006 agricultural census conducted in Brazil, it was found that 11 percent of the holdings had failed to provide information on production, while in previous years (specifically, 1980, 1985 and 1996), this rate was only about 2 per cent. Furthermore, the results of the production of certain produce that could be compared with estimates from other sources – or from the supply balance based on information processing, exports, imports and inventory changes – indicated that the census data were affected by a significant underestimation at the national level. For soybeans, the underestimation was 13.6 percent; for cane sugar, 17.2 percent; and for oranges, 42.9 percent (Guedes & Oliveira 2013).

The changes in the sampling frame can automatically lead to a problem of undercoverage if new farms are created during the intercensus period or to overcoverage if holdings had ceased their activity.

Box.1.2. Case of undercoverage − United States of America

In the United States, the National Agricultural Statistics Service conducts several data collection operations. Two of them are the June Agricultural Survey and the Census of Agriculture. The former is based on an area frame and is conducted annually, whereas the latter is conducted every five years. In 2012, a capture-recapture approach was used to produce estimates for the Census of Agriculture. Two independent surveys are required to use capture-recapture methods: the Census of Agriculture and the June Agriculture Survey were chosen for the operation. Records in which farms were indicated in the census questionnaire are assigned weights that adjust for undercoverage, non-response and misclassification. Generally, follow-up surveys to the Census of Agriculture, conducted during the intercensus years, have been based on the assumption that the National Agricultural Statistics Science list frame, which serves as the foundation for the census mailing list, is complete. Although continual efforts are made to update the list frame, undercoverage persists. Failure of those follow-up surveys to account for such undercoverage has resulted in estimates that are biased downward. In 2016, for its local foods survey, the National Agricultural Statistics Science used a list frame obtained by means of web scraping; capture-recapture methods were used to compute adjusted weights for the list frame records (Global Strategy 2016).

b) Misclassification

Misclassification occurs when an operating arrangement that meets the definition of a farm is incorrectly classified as a non-farm, or when a non-farm

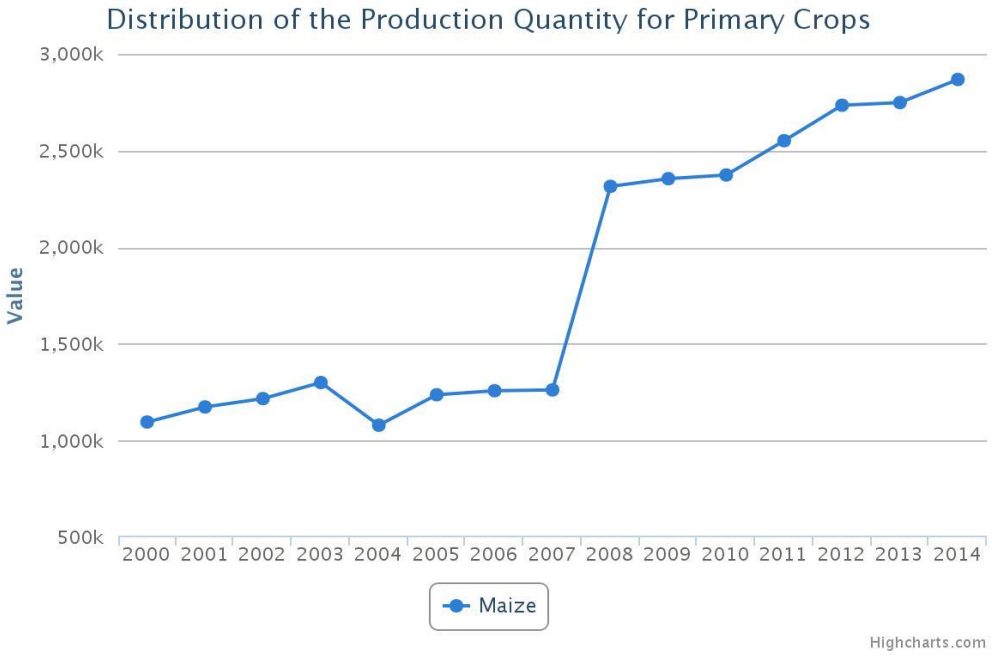
arrangement is incorrectly classified as a farm. In the United States, the census data consist of responses to a list-based survey. The mailing list for this survey is created and maintained wholly independent of the June Agriculture Survey real frame. The census data can be used to assess the degree of misclassification occurring in the survey. For this purpose, when analysing the 2012 Census of Agriculture, the National Agricultural Statistics Science matched each 2012 June Agriculture Survey tract to its 2012 census record. Disagreements in the conferral of farm status between the census and the June Agriculture Survey occurred when (a) tracts identified as non-farms in the June Agriculture Survey were subsequently identified as farms in the census, or (b) tracts identified as farms in the June Agriculture Survey were identified as non-farms in the census. If the tract was identified as a farm in either the June Agriculture Survey or the census, then the tract was considered to be a farm.

For the censuses prior to and including that of 2007, the analysis assumed that there had been no misclassification in the June Agriculture Survey. However, in 2009, the Farm Numbers Research Project was conducted. Twenty percent of the new June Agriculture Survey records were revisited, as these had been added to the sample and estimated to be or designated as non-agricultural during the pre-screening process. This demonstrated that there had been a substantial degree of misclassification; if the rest of the sample was affected by the same rate of misclassification, then the estimate should have included 580,000 more farms (Abreu *et al*. 2010). This was the first indication of an underlying cause that could help to explain the discrepancy in the published estimates.

c) Varying concepts and definitions

In an integrated agricultural statistics system, it is recommended that concepts and definitions be harmonized among agricultural censuses, other censuses, such as population censuses, and agricultural statistical surveys. Inconsistencies in data may be the result of changes or variations in concepts and definitions. Serious changes in concepts and definitions may affect estimates, as the series of data collected in different years do not measure the same variable, or measure the same variable for different survey populations. Either of these variations introduces inconsistencies.

Figure 1.1. Example of discrepancy in time series data



Source: Country STAT – Uganda

d) Greater reliability of data from the latest agricultural census and surveys based on census sampling frame

The most recent agricultural census and surveys based on the census sampling frame may provide more reliable data than those obtained in previous collection efforts. This leads to discrepancies between data from recent census/survey and data from previous surveys.

These may be caused by the following:

- The frame has changed because of changes in the structure and number of holdings and their distribution;
- Improvements in methodology;
- Improvements in the supervision and control system;
- Improvements in the relevant technology (new tools, global positioning systems, tablets, among others).

e) Non-response

Non-response occurs in all censuses and surveys. To overcome this problem, several countries estimate the missing data, even though this increases the

uncertainty associated with the estimates and may lead to bias (Abreu *et al*. 2011).

f) Sampling errors

The sampling errors noted in the literature can clearly be considered sources of discrepancy between the results of surveys and censuses.

g) Other non-sampling errors

Other non-sampling errors may arise because of inadequate questionnaires or defective methods of data collection, tabulation, coding and data entry, among other factors.

1.3. What will the reconciliation between agricultural census and survey data consist of?

The purpose of the reconciliation of census and survey data is to compare the estimates from previous surveys and the censuses results and to correct the discrepancies between them.

It is assumed that the data from the censuses are reference data, namely that they are more reliable compared to data from surveys, so an attempt is made to adjust the intercensal survey data in order to get the results compatible with the two censuses data. The difficulty of this exercise is to reconcile data from different years.

Changes in sample design or in the interview process and shifts in the sampling frame may lead to unrealistic changes in aggregates over a short period of time. The purpose of survey weights is to ensure that the sample represents the population. Therefore, those weights play an important role in creating consistent aggregates over time. One of the goals of the reconciliation process is to adjust the survey weights in order to be more realistic.

It should be noted that **before deciding to start the process of reconciliation, it is necessary to determine that there is no solid reason behind the presence of the gap in the time series.** For instance, a severe crisis or the availability of improved seeds in a particular year can significantly decrease or increase the production and therefore generate a gap in the time series.

In this technical report, it is assumed that the inconsistencies in the data are only because of one of the reasons listed in paragraph 1.2. Its main objective is

to propose methods and techniques to member countries for handling such discrepancies that could have occurred between census and survey data.

To perform data reconciliation, it is necessary to have the required expertise and all the data and metadata used for estimation during the surveys. Throughout the intercensus period, data and metadata should be well archived to allow for the implementation of reconciliation of census and survey data.

2

Addressing the Discrepancies

In this chapter, the common methods discussed in the literature review that can be used to correct discrepancies in the time series data are presented along with a clear explanation on how to implement these methods step by step. Practical examples are given, and the issues associated with the discrepancies are discussed.

2.1. Identifying the discrepancies in the time series data

In order to perform the reconciliation when data from the recent census are available, it is important to determine if there are breaks (important changes) in the time series (estimates from the previous surveys performed since the last census) and levels of confidence of these breaks. Although discrepancies may be the result of legitimate changes in the dynamics of the agriculture structure, common sources of those discrepancies are related to changes in the sampling frame, as well as a number of other factors, including, among them, survey methods, a change in concepts and definitions and methodological improvement. In the absence of a justified gap in the time series caused by, for example, drought, flood and improved seed, it is important to determine if a particular point in the time series – a gap – needs to be reconciled. A change-point analysis can be very useful for solving this issue.

A change-point analysis is capable of detecting multiple changes. For each change, it provides detailed information, including a confidence level indicating the likelihood that a change occurred and a confidence interval indicating when the change occurred. The change-point analysis procedure provided is extremely flexible. It can be performed on all types of time-ordered data, including, among others, attribute data and data from non-normal distributions.

Traditionally, control charts have been used to detect changes. The major difference between a change-point analysis and a control chart is that the latter is intended to be updated following the collection of each data point. A change-point analysis is intended to be performed less frequently to review the performance over a more extended period of time. The two methods can be used in a complementary fashion.

Over the years, several multiple changepoint search algorithms have been proposed to overcome this challenge, most notably the binary segmentation algorithm (Scott & Knott 1974; Sen & Srivastava 1975); the segment neighbourhood algorithm (Auger & Lawrence 1989; Bai & Perron 1998) and more recently the Pruned Exact Linear Time (PELT) algorithm (Killick, Fearnhead & Eckley 2012).

Several methods have been proposed to estimate the point at which the statistical properties of a sequence of observations change. The most common approach to identify multiple changepoints in the literature is to minimize

$$\sum_{i=1}^{m+1} [C(y_{(t_{i-1}+1):t_i})] + \beta f(m) \qquad (a)$$

where C is a cost function for a segment, such as negative log-likelihood and βf(m) is a penalty to guard against over-fitting. The change-point detection can implement three multiple change-point algorithms that minimize (a) binary segmentation (Edwards & Cavalli-Sforza 1965), (b) segment neighbourhoods (Auger & Lawrence 1989) and (c) the recently proposed Pruned Exact Linear Time (PELT) (Killick *et al.* 2012).

The point of change in the mean in a sequence of normal random variables can be estimated from a cumulative sum (CUMSUM) test scheme.

The R packages ***changepoint*** (Killick *et al.* 2016) and ***changepoint.np*** can be used in this regard.

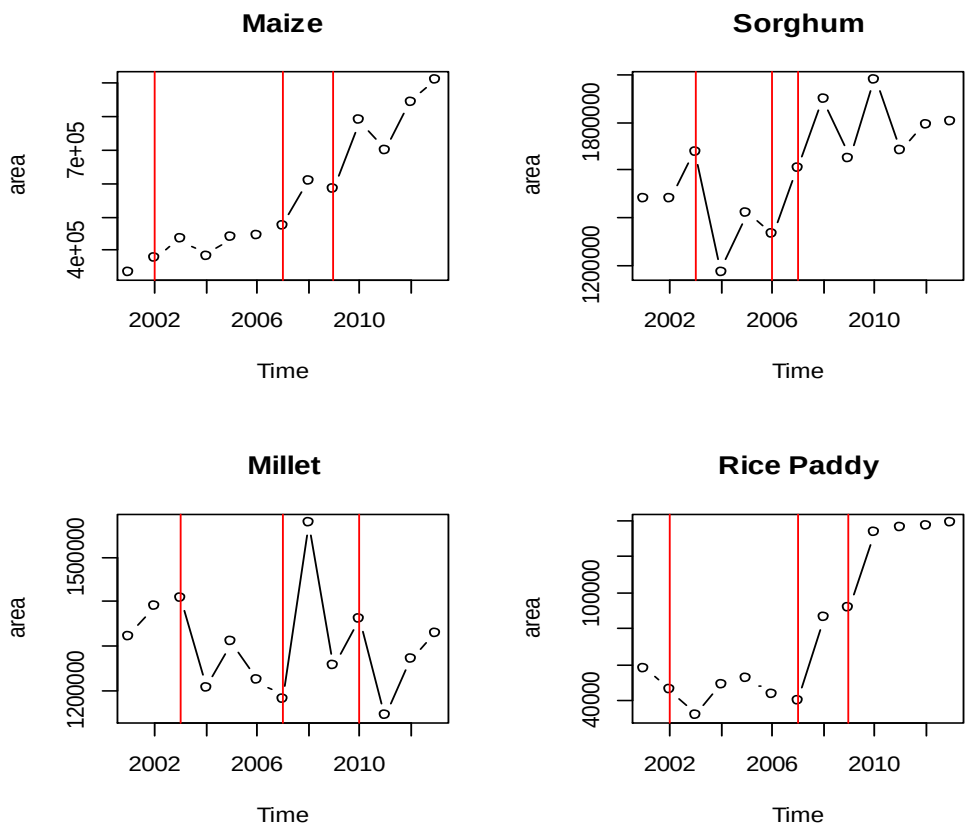
Table 2.1 and figure 2.1 show how to identify the points in the time series where a significant gap is present. The data sources are the Permanent Agricultural Survey and the 2008 census conducted in Burkina Faso. The method used is the Binary Segmentation method. Empirical distribution is used to compute the statistical test and the Modified Bayes Information Criterion has a value of 7.69. For all crops, 2007 is the location of the changepoint. The implementation of the 2008 census could explain the break in the time series data.

Table 2.1. Identification of the change point (number of quantiles=3, binary segmentation)

Crop	Maize	Sorghum	Millet	Rice paddy
Changepoint Locations	2002,2007,2009	2003, 2006, 2007	2003, 2007,2010	2002, 2007,2009

Source: CountrySTAT-Burkina Faso.

Figure 2.1. Identification of the changepoints in the time series data − Burkina Faso



For sorghum, 2003 has been identified as a changepoint. As no census was carried out in that year, the break in the time series cannot be the result of the implementation of a new census.

This method helps in determining if there is a gap between the last survey estimate and the new census result. It also helps in identifying any gap in the year of the census and therefore, if a reconciliation should be implemented.

2.2. The Generalized Cross-Entropy approach

This section describes how difficult statistical estimation problems can often be solved efficiently by means of the Generalized Cross-Entropy approach. This approach can be viewed as an adaptive importance sampling procedure that uses the cross-entropy or Kullback–Leibler divergence as a measure of closeness between two sampling distributions. This method can be used to solve a diverse range of optimization problems.

The approach of reconciliation used in this section entails comparing and analysing two vectors of data: (a) the survey weights derived from the survey; and (b) the weights derived from an extrapolation of the survey weights. The extrapolation is carried out in such a way that the data estimated follow a clear and reasonable pattern throughout the years. This extrapolation can be completed by using growth rate estimation or exponential smoothing.

Djety & Akoua (2008) present an approach for reconciling census and survey data based on the growth rate. The Generalized Cross-Entropy method presented in this document reconciles the original survey weights and the new survey weights obtained based on the growth rate approach.

The model consists of an objective function, which is minimized subject to constraints. The estimation approach uses an estimation criterion based on an entropy measure of information. The survey household weights are treated as a prior. New weights are estimated that are close to the prior using a cross-entropy metric and that are also consistent with the additional information. This additional information is about the adding-up normalization constraint of the probabilities and a moment consistency constraints. Using this method, information from the census can be capitalized to adjust survey weights.

Box 2.1 presents an algorithm useful for the reconciliation process. First, the survey weights for a given year using a simple growth rate method is

estimated. This makes it possible to set a trend to follow a reasonable pattern[4] (in this case linear). Second, the desired weights are estimated by minimizing a distance between original survey weights and the ones derived from the growth rate method. This procedure makes it possible to reduce the gaps in the time series data

The R codes at the annex 3. can be used to perform the algorithm in box 2.1

Box 2.1. Algorithm of data reconciliation using the Generalized Cross-Entropy Method

$\bar{S}_k$is the original sampling survey weight for a given statistical unit, S_k is the new sampling survey weight used for the reconciliation,C_k is the prior obtained from an extrapolation based on census data, $w_l$error weights estimated in the Cross-Entropy procedure, $\bar{w}_{k,l}$ is its prior, $\bar{v}_{k,l}$is the error support set, f_t represents a general aggregator and p_k a probability or a sample weight.

A0) Identify the gap by using change-point/change-point.np R packages;

A1) Estimate the new survey weights C_k using the growth rate method (or exponential smoothing);

$$C_k = \bar{S}_0 * (1 + g)^k$$

Where

$\bar{S}_0$ is the survey weights derived from the previous agricultural census;

g is the yearly relative growth rate, calculated using the censuses data as follow:

$$g = (\frac{P_{n+m}}{P_n})^{1/m} - 1$$

P_{n+m} and P_n are respectively the values of the variable of interest in the year $n+m$ and m (censuses years), and m the intercensal period.

[4] Because of the flexibility of this method, other reasonable patterns can be considered, such as 0arabolic. In this technical report, it is assumed that from one census to another one, the variable is moving around a line.

A2) Set the error support set of $\bar{v}_k$ with five terms equal to (-3σ, -σ ,0,σ, 3σ);

with

$$\sigma = \sum_l \bar{w}_{i,l} \bar{v}_{i,j}{}^2$$

A3) Minimize the equation

$$Min_{S,w} \sum_k S_k \ln\left(\frac{S_k}{\bar{S}_k}\right) + \sum_{k,l} w_{k,l} \ln\left(\frac{w_l}{\bar{w}_l}\right) \quad (1)$$

subject to

$$\sum_k p_k f_t(S_k) = C_t + \sum_l w_{t,l} \bar{v}_{k,l}, \qquad t \in [1, \dots, T], l \in [1, \dots, L] \quad (2)$$

and additional adding-up constraints on the error weights

$$\sum_k p_k = 1, \qquad and \quad \sum_l w_{k,l} = 1 \quad (3)$$

$$Min\{\bar{S}_k, C_k\} \leq S_k \leq Min\{\bar{S}_k, C_k\} \quad (4)$$

where,

the set *l* defines the dimension of the support set for the error distribution and the number of weights that must be estimated for each error. The prior variance of these errors is given by:

$$\sigma = \sum_l \bar{w}_{i,l} \bar{v}_{i,j}{}^2$$

$\bar{w}_{i,l}$ is the prior weights on the error support set.

Box 2.2. Example of data reconciliation using the Generalized Cross-Entropy Method (Burkina Faso-Maize)

Using data from 2001-2007 (surveys) and 2008 (census), the following results are derived:[5]

A0) Identify the gap by using chagepoint/changepoint.np R packages;

Using the package changepoint.np, the location of the change-point in the time series is given. In Burkina Faso, 2008 was the year of the census, and it is identified clearly as a year of change-point, therefore the reconciliation is necessary.

Crop	Maize
Change-point Locations	2008

A1) Estimate the new survey weights C_k using the growth rate method

The new survey weights C_k using the growth rate method is computed using the formula:

$$g = (P_{2008}/P_{2001})^{1/7} - 1$$

$$C_{k=}\ C_{2000} * (1+g)^k$$

where,

P_{2008} is the variable derived from the last census used as the benchmark;

P_{2001} is the variable derived from the 2001 survey used as the benchmark;

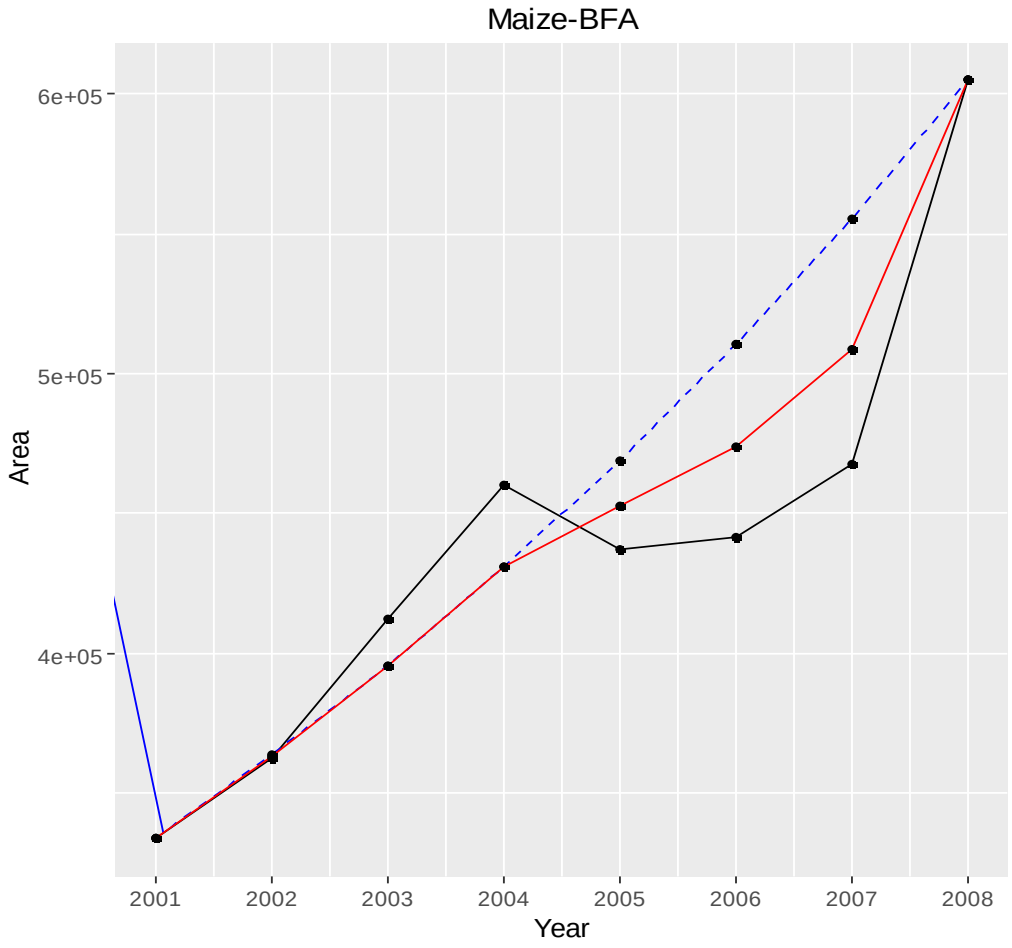
k is a given year in which P_k; is extrapolated.

C) and D) is done using the R code in annex 3

[5] Ae data were available only from 2001, it is assumed that the years 2001 and 2008 (year of the census) are the r reference data, from which all the survey data (2002-2007) should be reconciled.

Figure 2.2 presents the estimates using the growth rate method and the adjusted and original area. The adjustment of the area in 2007 has reduced the gap from 30 percent gap to 19.6 percent. In 2004, the area declined from 460,320 ha to 440,455 ha (4.3 percent) replaced a 4.9 percent increase in 2004 to absorb the decrease line. The area of sorghum increased between 2001 and 2005. In 2005, a decrease of 26 percent was observed. The reconciliation made it possible to reduce this gap and maintain an increasing trend. In fact, for maize, using the test to identify the changepoint, 2007 is no longer a changepoint.

Figure 2.2. Time series plots of the area estimates for maize in Burkina Faso



--- Survey estimates

--- Adjusted data (**Generalized Cross-Entropy Method**)

--- Estimates using the Growth Rate Method

Table. 2.2. Results for Area − Burkina Faso

Year	Area		Adjusted area (CE)		Estimates using Gr. Rate	
	Maize	Sorghum	Maize	Sorghum	Maize	Sorghum
2001	334 682	1 478 359	334 682	1 478 359	334 682	1 478 359
2002	362 565	1 044 212	363 043	1 045 591	364 509	1 532 518
2003	412 547	1 060 596	395 803	1 017 456	396 995	1 588 661
2004	460 320	1 082 747	431 238	1 014 179	432 376	1 646 861
2005	440 455	1 422 272	452 499	1 125 490	470 910	1 707 192
2006	441 745	1 053 717	473 767	1 130 081	512 878	1 769 734
2007	467 988	1 209 955	508 456	1 314 518	558 586	1 834 568
2008	608 368	1 901 776	608 368	1 901 776	608 368	1 901 776

2.3. Regression methods

These methods are based on the estimation of the total numbers of farmers (statistical units) in a survey year using the information from the census. Once it is done, this number is used to adjust the original survey weights.

Generally, a two-stage stratified sample survey is used to estimate the annual production. For instance, in Burkina Faso, the primary sample units are constituted by the administrative villages and sectors of the semi-urban localities of the national territory. The secondary sample units consist of the agricultural households in administrative villages and semi-urban areas. The unit of observation is therefore the farm household, which is defined as a household in which one or more members cultivate plots on behalf of the household. In principle, the primary sample units are first sampled in the

province (domain) in proportion to their size (N_{hi}). The secondary sample units are then sampled simultaneously in the constituted strata. Hence, the survey weight of the secondary sample units:

$$w_i = \frac{N_T}{m_h N_{hi}^*} \times \frac{N_{hi}}{n_h}$$

with

m_h: *number of primary sampled unit sampled from strata h;*

N_T: *number of secondary sampled units (households)of the stata h;*

n_h: *number of households sampled from strata h;*

N_{hi}: *number of households of the primary sampled units which household i belongs to;*

N^*_{hi}: *number average of households of a primary sampled uit.*

Throughout the year N_T varies, as new farms are created or disappear. The problem is that those farms are not included in the sampling frame. The approach proposed in this section allows an update to the value N_T. It makes it possible to estimate the number of farms that are not included in the sampling frame.

3.2.1. Best Linear Unbiased Prediction method

This approach assumes that the population survey response variables Y are a random sample drawn from a larger population, and that they assign a probability distribution $P(Y|\theta)$ with parameters θ.

It is assumed that the population values of Y follow the model

$$E(Y_i|X_i) = X_i^T \boldsymbol{\beta}, Var(Y_i|X_i) = \sigma^2 D_i\,,\ Cov(Y_i, Y_j) = D_{ij}\sigma^2, i \neq j, \qquad \text{(M1)}$$

where X_i denotes a p-vector of benchmark auxiliary variables for unit i that is known for all the population units over the intercensus period. Auxiliary variables could be the labour (workforce), the size of the farm in terms of the number of persons, the machinery, a dummy for the size of the farm land (from cadastral sources) and so on. D_i and $D_{ij}$are constants associated with population unit i and, jointly, with population units i and j, respectively.

The population total can also be written as $T = 1_s^T Y_s + 1_r^T Y_r$, where 1_s^T and 1_r^T are vectors of *n* (sample size) and (*N-n)*. The population matrix of covariates is $X = [X_s, X_r]^T$, where $X_s$is the *n x p* matrix for sample units and X_r is the *(N-n) x p* matrix for non-sampled units.

The result: have $\tilde{Y} = [Y_s, X_r\widehat{\boldsymbol{\beta}}]^T$

$\widehat{\boldsymbol{\beta}}$ is estimated using the census data.

As the survey weights are considered not accurate, $E(\tilde{Y})$ is used in lieu of the Horvitz-Thompson estimates to estimate the population mean in the year of the survey.

N units included in the frame since the last census	n_0 units not included in the frame

Box 3.3 presents an algorithm for the implementation of this method. First, the variable of interest over the entire population is estimated using the Best Linear Unbiased Prediction model and a mean of this variable of interest in a given year is calculated. Second, the total of the variable of interest using the growth rate method for this year is estimated. Third, the difference between the total of the variable of interest using the growth rate method and the Horvitz Thompson estimate using the "old" survey weights is calculated. The ratio between this difference and the mean of the variable of interest is an estimate of the number of statistical units not included in the frame.

The simulation in annex 2 shows that the ratio considered as a variable is normally distributed around its true value.

As the efficiency of the Best Linear Unbiased Prediction method depends on how well the associated model holds, this method may be susceptible to model misspecification. To overcome the potential bias therein, another method has been developed.

2.3.2. Non-parametric Regression

The Robust GREG method uses the regression model

$$Y_i = m(X_i) + e_i, e_i \sim N(0, D_i) \qquad \text{(M2)}$$

This correction factor for bias is produced using a non-parametric smoothing of the linear model residuals against the frame variables known for all population units. The method used in the present document computes a kernel regression estimate of a one (1) dimensional dependent variable on p-variate explanatory data, given a set of evaluation points, training points (consisting of explanatory data and dependent data), and a bandwidth specification using the methods of Racine & Li (2004) and Li & Racine (2004). The difference with the previous methodology is the model specification.

The methodology used is the Non parametric Regression method, which is the same as the Best Linear Unbiased Prediction **model. The unique difference is the model specification** (see M1 and M2).

Box 2.3. Algorithm of data reconciliation using the Non-parametric Regression method

Let's assume that T_k is a linear extrapolation of the total of the variable of interest *k* years after the previous census and T_{HT} the Horvitz Thompson total calculated in the same year using the "old" survey weights. N is the number of the farmers included in the sampling frame and n_0 is the number of farms not present in the sampling frame "created" at a later stage after the census

A) Select the variables X to be included in the model;

B) Check the model specification;

If the Best Linear Unbiased Prediction is valid then do

C1) Using the closest census, perform the Best Linear Unbiased Prediction regression model:

$$E(Y_i|X_i) = X_i^T \boldsymbol{\beta}, Var(Y_i|X_i) = \sigma^2 D_i\ ,\ Cov(Y_i, Y_j) = D_{ij}\sigma^2, i \neq j,$$

with the variables as defined in the section 4.3.2.

C2) Compute $\tilde{Y} = [Y_s, X_r\widehat{\boldsymbol{\beta}}]^T$

C3) Compute $\bar{\tilde{Y}}$, where $\tilde{Y}$ are the estimates of Y and $\bar{\tilde{Y}}$ is the mean of $\tilde{Y}$.

Else

D1) Using the closest census, perform the Best Linear Unbiased Prediction regression model:

$$Y_i = m(X_i) + e_i, e_i \sim N(0, D_i),$$

with the variables as defined in the section 4.3.2.

D2) Compute $\tilde{Y} = [Y_s, X_r\widehat{\boldsymbol{\beta}}]^T$

D3) Compute $\bar{\tilde{Y}}$, where $\tilde{Y}$ are the estimates of Y and $\bar{\tilde{Y}}$ is the mean of $\tilde{Y}$.

E) Compute n_0 as

$$n_0 = \frac{T_k - T_{HT}}{\bar{\tilde{Y}}}$$

Where,

$$T_{k=} T_0 * (1 + g)^k$$

It is assumed that the size/production of the farms remains constant throughout the year.

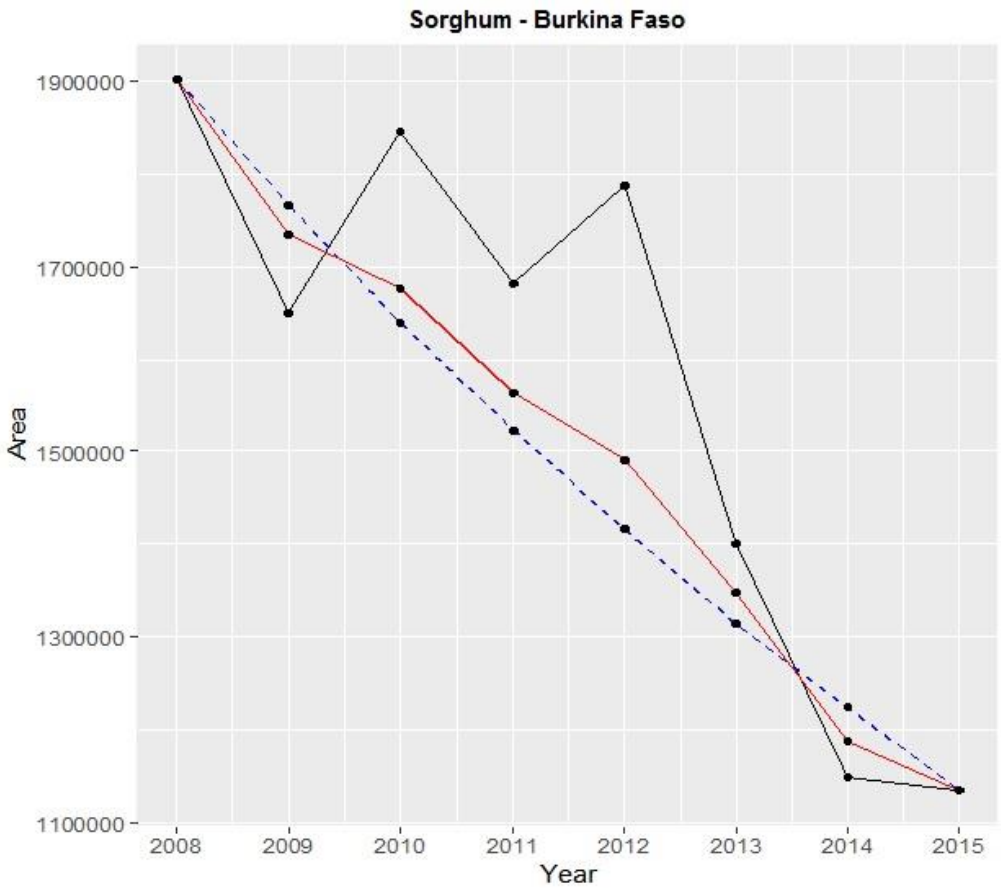
F) Compute the adjusted survey weights as:

$$W_{adj} = \frac{N + n_0}{N} . W_s$$

with W_{adj} and W_s are respectively the adjusted and the original survey weights.

The implementation of this algorithm can be done using the R codes at the annexe 8.3.

Figure 2.3. Time series plots of the area estimates for sorghum − Burkina Faso



--- Survey estimates

--- Adjusted data **(Best Linear Unbiased Prediction)**

--- Estimates using Growth Rate Method

2.4. Others methods to be used for data reconciliation

2.4.1. Handling misclassification

To deal with the concerns raised by the previous approach, and to obtain a coherent set of methods for the agricultural census and the June Agricultural Survey, Abreu *et al.* (2014) developed a ***capture-recapture*** approach to estimate the number of farms in the United States from the June Agricultural Survey. They proposed the following estimator for the number of farms from the June Agricultural Survey, with an adjustment for misclassification:

$$T_2 = \sum_{i \in SARJ} \frac{t_i}{\pi_i} \frac{\hat{p}_i(F \mid SARJ)}{\hat{p}_i(J \mid SARF)\hat{p}_i(R \mid SAF)\hat{p}_i(A \mid SF)},$$

where

i = indexes tract on the June Agricultural Survey

t_i = proportion of a farm represented by tract i

π_i = sample inclusion probability for tract i

S = tract is within the sample

A = tract passes the agricultural screening process

R = tract responds to the survey

F = tract is truly a farm

Logistic regression was used to estimate each of the above probabilities. Based on this estimator, at the United States level, the estimated misclassification rate for farms was 9.4 percent.

(Abreu et al. 2011; Global Strategy 2016)

2.4.2. Handling non-response

Generally, in a case of non-response, the data required are estimated. Therefore, the problem of non-response is related to the estimator error. A vast body of literature exists on how to account for non-response.

To reduce non-response bias in sample surveys, a common method consists of multiplying the respondent's survey weight by the inverse of the estimated response probability. Kim & Kim (2007) demonstrate that this approach is generally more efficient than relying on an estimator that uses the true response probability, provided that the parameters governing this probability are estimated by reference to maximum likelihood. Based on a limited simulation study, they also compare variance estimation methods that account for the effect of using the estimated response probability, and present the extensions to the regression estimator. The authors found that the adjustment using the estimated response probability improves the point estimator's efficiency and

also reduces bias because it incorporates additional information from the auxiliary variables used in the response model. In this case, the variance estimators discussed account for the variance reduction related to the estimation of the response probability.

McCarthy *et al*. (2010) have modelled non-response in the National Agricultural Statistics Survey surveys using classification trees. They describe the use of classification trees to predict survey refusals and inaccessible respondents.

The methods for solving non-response issues may be applied during the reconciliation of census and survey data, if this has not already been done during survey data estimation. Most of these methodologies do not use census data and can thus be applied before the census year. If they have been applied, problems relating to non-response are considered to be estimation problems.

(Global Strategy 2016)

Abreu *et al*. (2011) propose a non-response model to handle non-response. The current estimate for the number of farms based on the June Agricultural Survey can be simplified to the following expression,

$$T = \sum_{i \in R} \pi_i^{-1}.y_i.t_i,$$

where

R denotes the set of respondents, π_i denotes the inclusion probability of respondent *i*, $y_i = 1$

if the tract contains a farm and is 0 otherwise, and t_i= tract-to-farm ratio. If ϕ_i denotes the

probability of response for unit *i*, then the non-response weighted estimate for the total number

of farms would be

$$T_{NR} = \sum_{i \in R} \pi_i^{-1}.\phi_i^{-1}.y_i.t_i,$$

where ϕ_i is the probability the *i-th* tract responds. In practice, ϕ_i is unknown and must be

estimated. It can be estimated in several ways.

(Abreu et al. 2011)

2.4.3. Calibration methods

Information from other sources of data could be used to calibrate the weights to obtain new weights conform to the information presented in the source.

The ideal situation would be that the size of each primary unit in terms of agricultural households is updated every year, which is almost impossible because of the cumbersomeness and cost of the operation. Consequently, outside of the census years, survey weights may be inadequate.

The other sources could be other censuses (population census) and survey, the projection of the agricultural population and administrative data, among others.

As the objective is to reconcile agriculture survey and census data, the agricultural census is an important source of data to consider.

However, before using data from a data source, the validity and the consistency of data should be checked.

Box 2.4. Varying of concepts and definition

When the concepts and definition change, the reconciliation becomes tedious work. Often, the statistical units also change. In those situations, information from the census could be used to calibrate the weights to obtain new weights conform to the information from the censuses.

Let's say from the last survey the statistical unit has changed (for instance, farm to agricultural household). The variable of interest is the area harvested for maize. As the two datasets (survey and the latest census) are from different years, a growth rate is applied to estimate the area harvested in the survey year. The ratio between this estimate (growth rate method) and the survey Horvitz Thompson estimate (using the original or "old" survey weights) is applied to the old survey weights to obtain the new survey weights.

The following is undertaken:

A0) Calculate a yearly growth rate using the two censuses, considered as reference data.
A1) Calculate estimates in the *i* years after the census

$$P_i = P_0 * (1 + g)^i$$

Where
P_0 is the area estimates from the first census
P_i is the area estimates from the second census
g is the yearly growth rate

A2) Calculate the new adjusted weights as follow:

$$W_{adjusted} = \frac{P_i}{P_0} * W_i$$

where,
W_i is the original or "old" survey weights
$W_{adjusted}$ is the new survey weights

2.5. Discussion and recommendations

The advantage of Generalized Cross-Entropy approach is that this method is flexible and can be used in almost all cases. It can even be used in cases in which auxiliary variables are not available. The disadvantage of the method is that compared to the other methods, the implementation of it requires much more work. These methods as presented, reduce significantly the gap in the time series data, but it is possible that a gap persists in the time series data.

By using the growth rate method, it is assumed that the variable of interest is monotonically increasing. This is not always the case. However, due to its flexibility, the Generalized Cross-Entropy approach allows for the consideration of other patterns, such as hyperbolic and exponential.

Changing the survey weights should be done on most variables in a multivariate approach, if not the data become incoherent for further use in economic modelling.

The problem of reconciling data from surveys and census is challenging and necessitates an agreement at the country level, such as through a formal protocol, on the implementation of the process. Discussions on these political decision protocols as adopted by some countries could also help to tackle the problem. Questions, such as when to apply corrections, and to what extent to correct estimates back in time are relevant and should be addressed

Data from agricultural censuses are used for benchmarking at the macrolevel and for data confrontation and verification. The intercensus revisions provide an opportunity to include modifications to compilation methods or concepts that have not yet been integrated in published data. Census data are also used to revise the value of a number of commodities for which annual data are not available.

The validation of the results is also key while reconciling census and survey data. The main objectives of data validation are to guarantee the quality and consistency of the agricultural census data and to make recommendations for their publication. Data validation is a complex process in which human judgement is vital. Validators should follow a data validation plan and a data validation checklist as guidelines to the data validation tools available on a central processing system. However, validators ultimately have to solve problems and make decisions based on the analysis of background information, respondent feedback, expert consultation and common sense.

However, the revised data should be consolidated as much as possible with other data, such as supply and demand outputs. The new estimates should be validated by a pool of experts prior to publication. It is important for personnel who were involved in data collection and estimation to be part of this pool.

The final set of weights may be calibrated to the population distribution based on population data from a statistically superior external source, namely the most recent census or findings from another contemporary national survey with population size estimates of equal or greater quality. Reputable and generally accepted population projections can also be used as the object of calibration.

For the methods presented in the present document, the data of the two censuses are used as reference data. However, when only one census has been implemented, the data for the most recent survey can be considered as reference data. This can be done, especially when this survey was well implemented and the sample ratio was high.

Those methods show that an updated sampling frame is a key input required to avoid discrepancy in data. Therefore, some actions need to be taken into consideration.

a) Additional samples

Because of population movements, over a certain period of time, new statistical units may appear in the population of households or farms. Therefore, discrepancies may arise between the estimates based on survey data and the data from the previous census. If the list frame of those units is available, such as from administrative files, an additional sample of the new units can be drawn. The population of new units may be considered as a stratum, and the new estimates can be obtained (Global Strategy 2015).

b) Tracking

Changes in statistical units adversely affect their representativeness and make estimates less precise, thus generating inconsistencies between census data and survey data. Those changes must be corrected if the integrity of the units is to be maintained. When a part of a unit does not exist at the time of collection, this part needs to be tracked, especially if its absence is not random. For example, if a portion of a farm changes ownership because of a conflict over land, arrangements should be made with the new owner to collect data on this part (Global Strategy 2015).

c) Weight-sharing methods

When the surveys are conducted with a panel of agricultural holdings selected from the data of the most recent general agricultural census, changes in statistical units may also be corrected by means of weight-sharing methods, including the General Weight Share Method, which was developed by Lavallée (2007). Those methods are explored in further detail in another important publication of the Global Strategy: the *Guidelines for the Integrated Survey Framework* (Global Strategy 2015).

If a sample panel is used, those methods of adjustment may be of great assistance to the reconciliation with census data.

d) Oversampling

To cope with the disappearance of statistical units in a region or in a stratum, the size of the sample size may be increased. This helps to maintain sample accuracy, but it does not prevent bias (Global Strategy 2015) This technique is applied when the sample is selected before obtaining the survey results necessary for the reconciliation. Therefore, even after it is implemented, it may still be necessary to proceed to the reconciliation with census data.

3

Operational Strategy

In this chapter, the main steps to be followed in the data reconciliation process are presented.

3.1. Identification of the key variables

Data reconciliation may be a time-consuming task, especially when advanced techniques are required to correct the gaps. In fact, those techniques sometimes require the collection of secondary data for them to be implemented. Accordingly, rigorous data reconciliation may not be possible for all the survey variables. It is therefore important to identify a number of key variables for the reconciliation process.

A census of agriculture is a statistical operation for collecting, processing and disseminating data on the structure of agriculture, covering the whole or a significant part of the country. Typical structural data collected in a census of agriculture are size of holding, land tenure, land use, crop area harvested, irrigation, livestock numbers, labour and other agricultural inputs.

It is necessary to identify variables to implement the reconciliation in the core module. The variable of the production and the labour can be prioritized.

Depending of countries' specificity, the variable could be, for example:

- Distribution of the production quantity for primary crops (ton);
- Distribution of area harvested for primary crop (ha);
- Distribution of area sown for primary crops (ha);
- Distribution of number of live animals (head).

In many developing countries, the production is not directly observed, but it is estimated as the product of the area harvested and the yield. Both variables are observed. In this case, only the area harvested can be adjusted and then the production can be calculated.

3.2. Computing the gaps

This operation is to compute the gap between surveys and the census data regarding the key variables selected. Both absolute and relative growth need to be calculated to assess the gap in the time series. Comparisons of ratios, such as the proportion of maize planted area, may also be useful given that ratios are not likely to change in a short term.

The computation of the gap can help in determining the changepoint locations.

3.3. Diagnosis to identify the unjustified gaps and the sources of discrepancies

For each variable, it is important to analyse if the gap is normal or if there is a discrepancy. Some gaps may be linked to the normal evolution of variables from the year of the survey to the one of the census. In some cases, previous conjectural factors that occurred in the country may explain the differences in data. In this step, the opinions of subject matter experts with deep knowledge of the agricultural economy of the country may be helpful. Secondary data may also help in understanding some gaps. A change-point analysis is very useful to determine if there is a gap in the time series. When discrepancies are identified, their sources should be explored in order to assess the causes of the gaps.

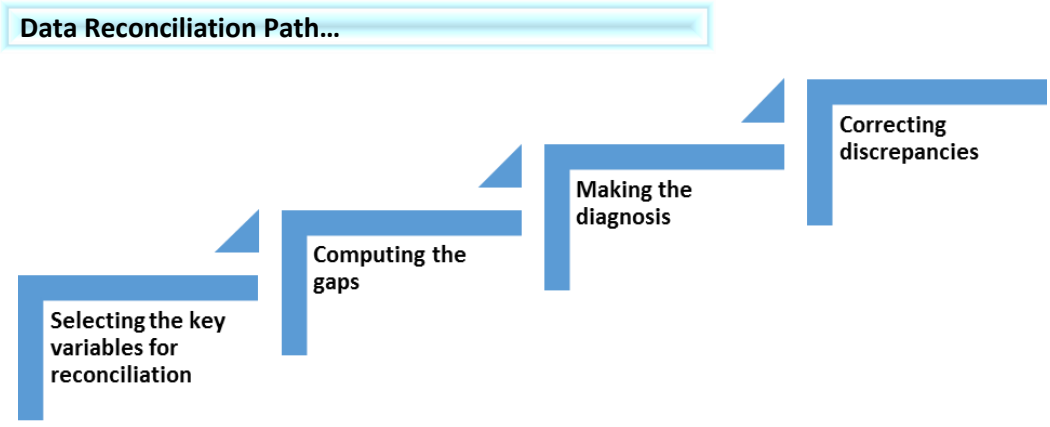
3.4. Applying suitable methods to correct the discrepancies

The methods to apply for correcting discrepancies depend of the availability of auxiliary data. Regressions methods are applied only if correlated variables are available. The Generalized Cross-Entropy method is always applicable, but, on the other hand, it could be time-consuming in terms of computational implementation.

Table 3.1. Application of different methods of data reconciliation

Cases	Methods
- Changes in the sampling frame	- When auxiliary variables are not Available • Generalized Cross-Entropy - When auxiliary variables are available • Generalized Cross-Entropy • Best Linear Unbiased Prediction Method • Non-parametric Method
- Misclassification	- Capture-recapture approach
- Varying concepts and definitions	- Calibrations techniques
- Non-response	- Classification trees,

Figure 3.1. Representation of the operational process of data reconciliation



3.5. Possible strategies to prevent or minimize discrepancies in future census and survey operations

In addition to eventually reducing the discrepancies between the new census results and intercensus agricultural survey results, the analysis of the sources of those discrepancies should be used to develop possible strategies to prevent or minimize such discrepancies in the future. Elements of such a strategy may include:

a) Adoption of an integrated census and survey approach which would facilitate the harmonization of the concepts and definition used across data collection operations;

b) Adoption of a master sampling frame and regularly updating the frame (using, for example, rotating samples may help in reducing the impact of frame obsolescence);

c) Using external information to cross check and validate or adjust annual survey data;

- Use relevant statistical methods to adjust survey data, including
- Adjusting weights when probability proportional to the size is used, taking into account new information available on the size of the sampling, units such as a primary sampled unit;
- Tracking: when a part of a unit does not exist at the time of collection, this part needs to be tracked, especially if its absence is not random;
- Weight-sharing methods: when the surveys are conducted with a panel of agricultural holdings selected from the data of the most recent agricultural census, changes in statistical units may also be corrected by means of weight-sharing methods, including the General Weight Share Method, which was developed by Lavallée (2007);
- Use some other methods described above

Conclusion

The implementation of a new census can reveal gaps in the time series derived from the past surveys. These changes could be caused by: (a) changes in the sampling frame; (b) misclassification; (c) varying concepts and definitions; (d) greater reliability of data from latest agricultural census and surveys based on census sampling frame; (e) non-response; (f) sampling errors; and (g) other non-sampling errors.

As survey weights are used to adequately extrapolate the results of the sample to the population level, they play an important role in creating consistent aggregates over time. Therefore, the process of reconciliation described in the present document is focused on the adjustment of the survey weights to ensure that the correct representativeness of each unit is included in the sample.

This document presents tested methods identified in the literature review to solve the problem of inconsistencies in the data.

The cross-entropy estimation method presented in this report provides an effective and flexible procedure for reconciling survey weights derived from an agricultural survey with those from a census survey. The survey weights are treated as a prior. New weights are estimated that are close to the prior using a cross-entropy metric and are also consistent with the additional information. The results indicate that the approach is powerful and flexible, supporting the efficient use of information from a variety of sources to reconcile data at different levels of aggregation in a consistent framework.

Model regression methods are also presented. These methods make it possible to estimate the total size of the population and to recalculate the survey weights. The limit of these methods is the availability of updated auxiliary variables. Some variables as the size of the holding, the number of holding in an area or the variable to the owning of equipment, among others, are collected during the census. They can be updated using other data sources (population census, administrative file or other survey). Global Strategy (2017) shows these two methods of regression lead to close results, when the Best Linear Unbiased Prediction model is valid.

Others methods of calibration can also be used in other models to make sure that the results of the survey are consistent with information contained in other sources of data, such as a population census or administrative data.

For those methodologies, a very large amount of data is handled, and from a computational point of view, the implementation of the reconciliation can be a heavy task and very time-consuming. Roughly speaking, those methods reduce the gap in the time series. If the gap is very high, the methods reduce the gap; but it may be possible that the gap still exists. This can constitute a limitation of these methods.

It has been observed that some action needs to be undertaken to eventually reduce the discrepancies between the new census results and intercensal agricultural survey results: (a) adoption of an integrated census and survey approach, which would facilitate the harmonization of the concepts and definition used across data collection operations; (b) adoption of a master sampling frame and regularly updating the frame (using for example rotating samples may help in reducing the impact of frame obsolescence); (c) using external information to cross check and validate or adjust annual survey data; and (d) use relevant statistical methods to adjust survey data.

References

Abreu, D.A., Arroway, P., Lamas, A.C., Lopiano, K.K. & Young, L.J. 2010. "Using the census of agriculture list frame to assess misclassification in the June area survey. Proceedings of the 2010 Joint Statistical Meetings. Section on Survey Research Methods – JSM 2010. American Statistical Association Publication: Alexandria, VA, USA.

Abreu, D.A., Arroway, P., Lamas, A.C., Lopiano, K.K. & Young, L.J. 2011. Adjusting the June Area Survey Estimate for the Number of U.S. Farms for Misclassification and Non-response. Research and Development Division. RDD Report No. RDD-11-04. USDA/NASS Publication: Washington, D.C.

Auger, I.E. & Lawrence C.E. 1989. Algorithms for the Optimal Identification of Segment Neighborhoods. *Bulletin of Mathematical Biology*, 51(1), 39-54.

Bai, J. & Perron, P. 1998. Estimating and Testing Linear Models with Multiple Structural Changes. *Econometrica*, 66(1), 47-78.

Djety, G. & Akoua, A.Y. 2008. Rapport sur la réconciliation des résultats du RNA 2001 avec les statistiques courantes. Report of Project TCP/IVC/3101. Ministère de l'agriculture de Côte d'Ivoire & FAO Publication: Accra.

Edwards, A.W. & Cavalli-Sforza, L.L.1965. A Method for Cluster Analysis. *Biometrics*, 21(2), 362-375.

FAO. 1996. Conducting agricultural censuses and surveys. FAO Statistical Development Series No. 6. FAO Publication: Rome.

FAO. 2015. **World Census of Agriculture 2020**. Volume 1: Programme, concepts and definitions. FAO Publication: Rome.

Guedes, C.A.B. & Oliveira, O.C. 2013. The importance of system GCEA to Brazilian agricultural statistics, Paper prepared for the International Conference on Agricultural Statistics VI (IDCB Technical Session 7), 23-25 October 2013. Rio de Janeiro, Brazil.

Global Strategy 2015. *Integrated Survey Framework*. GSARS Publication: Rome.

2016. Literature Review on Reconciling Data from Agricultural Censuses and Surveys. Technical Report Series GO-14-2016. Global Strategy to Improve

Agricultural and Rural Statistics (GSARS). GSARS Publication: Rome. Available at http://gsars.org/wp-content/uploads/2016/07/Literature-Review-on-Reconciling-Data-from-Agricultural-Censuses-and-Surveys-200716.pdf.

2017.Technical Report on Reconciling Data from Agricultural Censuses and Surveys GSARS Publication: Rome.

Killick, R., Fearnhead, P. & Eckley, I. 2012. Optimal Detection of Changepoints with a Linear Computational Cost. *Journal of the American Statistical Association*, 107(500), 1590-1598.

Killick, R, Haynes,K, Eckley, I., Fearnhead, P. & Lee, J. 2016. Package 'Changepoint'. R. Available at: https://cran.r-project.org/web/packages/changepoint/changepoint.pdf.

Kim, J.K. & Kim, J.J. 2007. Non-response Weighting Adjustment Using Estimated Response Probability. The Canadian Journal of Statistics. 35(4): 501-514.

Lavallée, P. 2007. *Indirect Sampling*. Springer: New York, USA.

Li, Q. & Racine, J.2004. Cross-validated Local Linear Nonparametric Regression," *Statistica Sinica*, 14, 485-512.

McCarthy, J.S., Jacob, T. & McCracken, A. 2010. Modelling Non-response in National Agricultural Statistics Service (NASS) Surveys Using Classification Trees (No. 235029).United States Department of Agriculture, National Agricultural Services Publication: Washington, D.C.

Racine, J.& LI, Q. 2004. Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data, *Journal of Econometrics*, 119 (1): 99-130. .

Scott, A.J. & Knott, M. 1974. A Cluster Analysis Method for Grouping Means in the Analysis of Variance. *Biometrics*, 30(3), 507- 512.

Sen A., & Srivastava M.S. 1975. On Tests for Detecting Change in Mean. *The Annals of Statistics*, 3(1), 98-108.

Turtoi, C., Akyildirim, O. & Petkov, P. 2012. Statistical Farm Register in the EU Acceding Countries - A Conceptual Approach. *Economics of Agriculture* 59(1).

Annexes

Annex 1: Generalized Cross-Entropy Model

The Cross-Entropy model is presented as follows:

$\bar{S}_k$is the original sampling survey weight for a given statistical unit, S_k is the new sampling survey weight used for the reconciliation,$C_k$is the prior obtained from an extrapolation based on census data, $w_l$error weights estimated in the Cross-Entropy procedure, $\bar{w}_{k,l}$ is its prior, $\bar{v}_{k,l}$is the error support set, f_t represents a general aggregator and p_k a probability or a sample weight.

$$Min_{S,w} \sum_k S_k \ln\left(\frac{S_k}{\bar{S}_k}\right) + \sum_{k,l} w_{k,l} \ln\left(\frac{w_l}{\bar{w}_l}\right) \quad (1)$$

subject to

$$\sum_k p_k f_t(S_k) = C_t + \sum_l w_{t,l}\bar{v}_{k,l}, \qquad t \in [1,\dots,T], l \in [1,\dots,L] \quad (2)$$

and additional adding-up constraints on the error weights

$$\sum_k p_k = 1, \qquad and \quad \sum_l w_{k,l} = 1 \quad (3)$$

$$Min\{\bar{S}_k, C_k\} \leq S_k \leq Min\{\bar{S}_k, C_k\} \quad (4)$$

The set l defines the dimension of the support set for the error distribution and the number of weights that must be estimated for each error. The prior variance of these errors is given by:

$$\sigma = \sum_l \bar{w}_{i,l}\bar{v}_{i,j}{}^2$$

$\overline{w}_{i,l}$ is the prior weights on the error support set.

Assuming a prior distribution with zero mean and a standard error equal to σ , a support set with five terms equal to (-3σ , -σ ,0,σ, 3σ) is uses. Assuming normality of the prior distribution, the prior values of the weights can be computed given only knowledge of the prior mean and standard error. The constraint (2) is stochastic, where C_t is assumed to have a measurement error. The minimization is performed by a non-linear optimization algorithm. Constraint (4) makes sure that S_k lies between the original $\bar{S}_k$ andC_k. The minimization is performed by a non-linear optimization algorithm.

(Global Strategy, 2017)

Annex 2: Simulation of the distribution of n_0

In this simulation test, there is a randomly generated a dataset containing three variables, of which two of them are correlated to the third one. The size of the dataset is 12,000 units, of which 10,000 of them are considered as part of the data frame. The remaining n=2000 are considered out of the data frame.

1,000 units have been selected randomly from the 10,000 units.

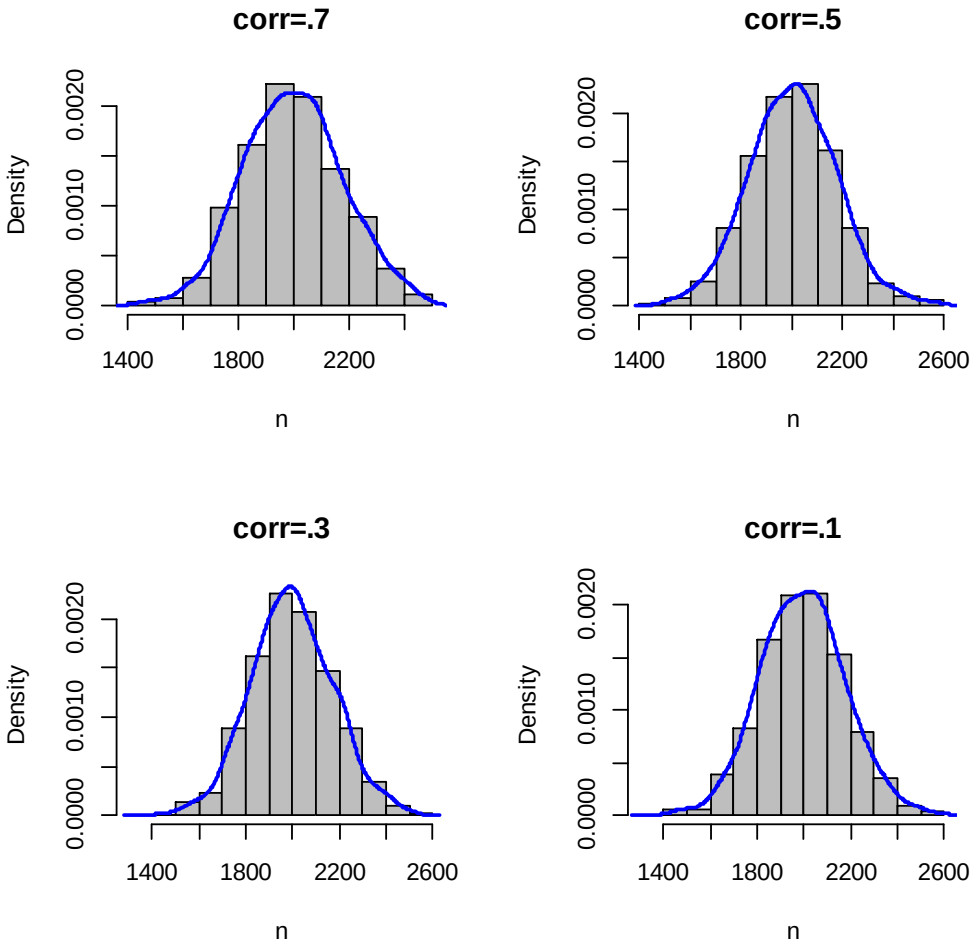
To estimate the number of units out of the sample, it is assumed that the total of the variable of interest on the whole population (12,000) is known. This number can be extrapolated by using the reference data from the censuses.

1000 simulations have been done and the distribution is presented in figure 8.1.

10,000 units (1,000 selected)	2,000 units

Figure 5.7 shows that the estimated number of the units out of the data frame is distributed normally around its true value: 2000. Even, when the correlation between the dependant variable and the independent variables is lower the estimation of n is good.

Figure 5.1. Distribution of the number of units out of the data frame (n)



Annex 3: R codes for Generalized Cross-Entropy Procedure

```
library(nloptr)

CE.fn=function(weights.vc,weightsNEW.vc){


# # # R E C O N C I L A T I O N P R O C E D U R E # # #

# S E T V A L U E S

# Original Survey Weights


F <- round(weights.vc,3)

# Weights derived from the Growth Rate Method


H <- round(weightsNEW.vc,3)


if (min(F)==mean(F)) {

return((H+F)/2)

}

else if (all(F==H)) {

return(H)

}

 else {


F=1/F

H=1/H


# Number of item groups

K <- length(F)

# Size of the support set

L <- 5
```

```
# Define the domain of the error distributions

# It is a K x (K*L) matrix

vb <- matrix(rep(c(-3*var(F),-1*var(F),0,1*var(F),3*var(F)),K),L,K)


# Define the error weights of the support set

# It is a K x (K*L) matrix

wb <- matrix(rep(c(1/72,27/72,16/72,27/72,1/72),K),L,K)l=K

# D E F I N E T H E M O D E L

# All updated parameters are in vector u:

# u[1], ..., u[K] are the updated FBS shares

# u[K+1], ..., u[2*K] are the updated weights of error support set
forH[1]

# u[2*K+1], ..., u[3*K] are the updated weights of error support set
for H[2]

# u[3*K+1], ..., u[4*K] are the updated weights of error support set
for H[3]

# u[4*K+1], ..., u[5*K] are the updated weights of error support set
for H[4]

# u[5*K+1], ..., u[6*K] are the updated weights of error support set
for H[5]


# OBJECTIVE FUNCTION

eval_f0 <- function( u , F, H, wb, vb ) {

return(

sum(u[1:K]*log(u[1:K]/F))+

sum(u[(K+1):(K+K*L)]*log(u[(K+1):(K+K*L)]/wb))

)

}

# CONSTRAINTS

#Create a multiplicator matrix with the form:

# 1 1 1 1 1 0 0 0 0 0 ....

# 0 0 0 0 0 1 1 1 1 1 ....

# . . .
```

```
# (it sums the error terms in constraint (2))

sumMatrix <- matrix(c( rep( c(rep(1, L), rep(0,(L*K))),(K-1)),

rep(1, L)), K, byrow = T)

# Define the constraint function

# (Equality constraints are built by combining 2 inequality
constraints).

eval_g_ineq0 <- function( u, F, H, wb, vb ) {

return(c(

# Constraint (2): WeightsNew.vc = H + error term


u[1:K] - H - sumMatrix %*% matrix(t(u[(K+1):(l+K*L)]*vb[1:(K*L)]))

,-u[1:K] + H + sumMatrix %*% matrix(t(u[(K+1):(l+K*L)]*vb[1:(K*L)]))

# Constraint (3): Updated weights to sum up to 1

, sum(u[1:K])-1

,-sum(u[1:K])+1

# Constraint (4): Updated weights to sum up to 1

, sumMatrix %*% matrix(t(u[(K+1):(l+K*L)])) -1

,-sumMatrix %*% matrix(t(u[(K+1):(l+K*L)])) +1

) )

}

# Constraint (5): Define lower and upper bounds for the algorithm.

LB = c(pmin(F[1:l],H[1:K]),rep(0,(K*L)))

UB = c(pmax(F[1:l],H[1:K]),rep(1,(K*L)))

# DERIVATIVES

# Gradient of the objective function (1)

eval_grad_f0 <- function( u, F, H, wb, vb ){

return(

log(u/c(F,wb))+1

)

}

# Jacobian of constraints (2)-(4)
```

```
# (Jacobian matrices of constraints (2)-(4) have to be
# collected in one function.)
# Predefine Jacobian matrices of constraint (2) and (3)
stochConstraint<-c()
weightSumConstraint <-c()
for (i in 0:(K-1)) {
# Jacobian of constraint (2)
stochConstraint= rbind(stochConstraint,
c(rep(0,i),1,rep(0,K-1-i) ,
rep(0,i*L),-vb[(L*i+1):(L*i+L)],
rep(0,(K-1-i)*L )
)
)
# Jacobian of constraint (3)
weightSumConstraint= rbind(weightSumConstraint,
c(rep(0,K) ,
rep(0,i*L),rep(1,L),rep(0,(K-1-i)*L )
)
)
}
# Jacobian of constraint (4)
uFBSSumConstraint= c(rep(1,K),rep(0,K*L))
# Collect Jacobian matrices in a function
eval_jac_g0 <- function(u, F, H, wb, vb){
return( rbind(stochConstraint,-stochConstraint,
uFBSSumConstraint,-uFBSSumConstraint,
weightSumConstraint,-weightSumConstraint
)
)
}#(K+1):(K*L+K)
# S O L V E
```

```
# First calculate a rough global optimum using ISRES algorithm.
# Take the mean of the HBS and FBS shares as starting values.
starting.shares=(H+F)/2
results.isres <- nloptr(x0=c(starting.shares,wb)
,eval_f=eval_f0
,lb = LB
,ub = UB
,eval_g_ineq = eval_g_ineq0
,opts = list("algorithm"="NLOPT_GN_ISRES",
maxeval=1.0e+3,
"xtol_rel"=1.0e-5,
"ftol_rel"=1.0e-10,
"stopval" = Inf)
,F = F
,H = H
,wb = wb
,vb = vb
)

# Save the global optima.
new.starting.values=results.isres$solution
# Take the results of the global optimization as starting values
# for a more precise local optimization.
# Local optima are calculated by the SQP algorithm.
results <- nloptr( x0=c(new.starting.values)
,eval_f=eval_f0
,eval_g_eq=NULL
,eval_g_ineq = eval_g_ineq0
,eval_grad_f=eval_grad_f0
,eval_jac_g_eq = NULL
,eval_jac_g_ineq = eval_jac_g0
```

```
,lb = LB
,ub = UB
,opts = list("algorithm"="NLOPT_LD_SLSQP",
maxeval=1.0e+5,
"xtol_rel"=1.0e-5,
"ftol_rel"=1.0e-10)
,F = F
,H = H
,wb = wb
,vb = vb
)
##
HF=results$solution[1:K]


return(1/HF)
}
}
```