Technical Report Series GO-14-2016

Literature Review on Reconciling Data from Agricultural Censuses and Surveys

July 2016

Literature Review on Reconciling Data from Agricultural Censuses and Surveys

Table of Contents

Abstract……… 4
Acronyms and Abbreviations……… 5
1. Introduction……… 6
2. Concepts and definitions……… 8
3. Sources of discrepancy and challenges……… 12
3.1. Sources of discrepancy……… 12
3.2. Difficulties and challenges……… 16
3.3. Conclusion……… 17
4. Methods for reconciling census and survey data……… 18
4.1. Design-based methods……… 19
4.2. Model-based methods……… 20
4.3. Model-assisted weighting methods……… 22
4.4. Growth rate method……… 24
4.5. Handling misclassification……… 29
4.6. Non-response……… 30
4.7. Other data adjustment techniques……… 31
4.8. Country experience: Canada……… 32
4.9. Conclusion……… 35
5. Gap analysis……… 36
5.1. Overall gap analysis……… 37
5.2. Case of livestock data……… 39
5.3. Additional issues……… 39
5.4. Conclusion……… 40
6. Conclusions……… 41
References……… 42

Abstract

This technical paper reviews the literature on the methodologies for reconciling data from agricultural censuses and surveys. The techniques that can be used for data reconciliation are described, and the main advantages and disadvantages of each are assessed. On the basis of the literature, for each relevant methodology, this paper formulates recommendations to be considered in the reconciliation of census and survey data. It also provides a gap analysis that documents and assesses the differences between the various methods.

The Member Nations of the Food and Agriculture Organization of the United Nations (FAO) have requested methodological guidance on reconciling census and survey data. To address this request, the Global Strategy to Improve Agricultural and Rural Statistics (hereinafter, Global Strategy) has prepared this literature review, which can also provide the basis for the development of a handbook.

The authors, Eloi Ouedraogo and Ulrich Eschcol Nyamsi, would like to thank Flavio Bolliger, Naman Keita, Dramane Bako of FAO and Linda J. Young, Director of the USDA/NASS Research and Development Division, for their constructive comments that contributed to improving the final version of this paper.

Acronyms and Abbreviations

BLUP	Best Linear Unbiased Prediction
DSDI	Direction des Statistiques, de la Documentation et de l'Informatique, Ministère de l'Agriculture de Côte d'Ivoire
EA	Enumeration Area
FNRP	Farm Numbers Research Project
FAO	Food and Agriculture Organization of the United Nations
GDP	Gross Domestic Product
GREG	Generalized Regression
JAS	June Agricultural Survey
NASS	National Agricultural Statistics Service (USA)
PSU	Primary Sample Unit
RGAC	Recensement Général de l'Agriculture et du Cheptel au Niger
RNA	Recensement National de l'Agriculture
SSU	Secondary Sample Unit
USDA	United States Department of Agriculture
WCA	World Programme for the Census of Agriculture (FAO)

1

Introduction

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 region, province, district, village, etc.), 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 normally involves collecting key structural data, by means of a complete enumeration of all agricultural holdings, and more detailed structural data, using surveys and sampling methods (FAO, 2010).

Data from agricultural censuses are useful in a variety of economic and social domains, including agricultural- and rural-sector planning and policymaking, as well as monitoring progress towards the Millennium Development Goals and addressing problems relating to poverty, food security and gender. Agricultural census data are also used in the establishment of agricultural indicator benchmarks and tools, to assess and improve current agricultural statistics during inter-census 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. When conducted by means of complete enumeration, agricultural censuses also provide a sampling frame that can be used in designing inter-census 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. These 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. For example, the geographic area covered by one of collections may be incomplete, as urban areas have been excluded. Certain types of holdings, such as small holdings, 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. If sampling is involved, the sample results may suffer from sampling errors. These discrepancies easily arise when the inter-census period is excessively long.

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

This literature review analyses the possible sources of discrepancy between time series from inter-census annual surveys and the results of new censuses. It also reviews the statistical methods that can be applied to address these discrepancies, taking into account countries' experiences. Finally, this technical paper outlines possible strategies and methodological options to implement the systematic reconciliation of intercensal survey data with the results of new censuses.

2

Concepts and definitions

FAO has made robust efforts to harmonize the concepts and definitions relevant to this sector, in collaboration with other international organizations, Member Nations and the scientific community. These concepts and definitions ensure that data are comparable over time and across countries; indeed, even minor variations in definitions increase the risk of inconsistencies arising in data reporting over time. It is thus important to ensure that any revisions made to concepts are well documented, and that the definitions applied in different assessments enable the comparability and consistency of data.

In agricultural statistics, the applicable definitions and classifications are provided by FAO and by other institutions collaborating with FAO. FAO remains the publisher of all terms, and FAO officers maintain and update the definitions falling within their area of expertise, in collaboration with the definitions' originators. Some of the concepts and definitions adopted by FAO and the Global Strategy are provided below. To facilitate reader comprehension, the list also defines other statistical 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 who 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, providing 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).

Area frame: an area frame is a set of land elements, which may be either points or segments of land. The sampling process may involve single or multiple stages. In most agricultural area frame surveys, the sampling unit is associated with a holding (Global Strategy, 2015).

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).

Cluster sampling: term used for sampling plans in which the sampling units are groups (clusters) of population units.

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 WCA, 2020).

Equal Probability Selection Method (EPSEM): 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 cattle, buffalo, horses and other equines, camels, sheep, goats and pigs, as well as poultry, bees, silkworms, etc. 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 (PSU): 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 (PPS) 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 that enables 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 (SSU): 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).

3

Sources of discrepancy and challenges

In general, a new census (with a complete enumeration) provides an opportunity to renew the reference data and the sampling frame. According to the World Programme for the Census of Agriculture 2010 (or WCA 2010; FAO, 2005), the inter-census period is 10 years (FAO, 2005). This chapter discusses the sources of discrepancy that may arise from a variety of data sources and the difficulties that may be encountered when reconciling census and survey data.

3.1. Sources of discrepancy

The sampling frame reflects the structure of agriculture at the time of its construction. Agricultural censuses conducted ten years apart may present inconsistencies in their data, especially if these have not been adjusted during the intercensal period. The sources of data discrepancy are the following:

a) Changes in the sampling frame

Measurements may be sought from agricultural holdings during annual surveys, to take into account any changes in the holdings' practices and therefore any 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, this may lead to inconsistency between data.

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

In Brazil, during the 2006 agricultural census, it was found that 11 per cent of holdings had ceased to provide information on production, while in previous years (specifically, 1980, 1985 and 1996), this rate was only 2 per cent, approximately. Furthermore, the results of the production of certain products 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 was affected by significant underestimation at national level. For soybeans, the underestimation is in the order of 13.6 per cent; for cane sugar, 17.2 per cent; and for orange, 42.9 per cent (Guedes & Oliveira, 2013).

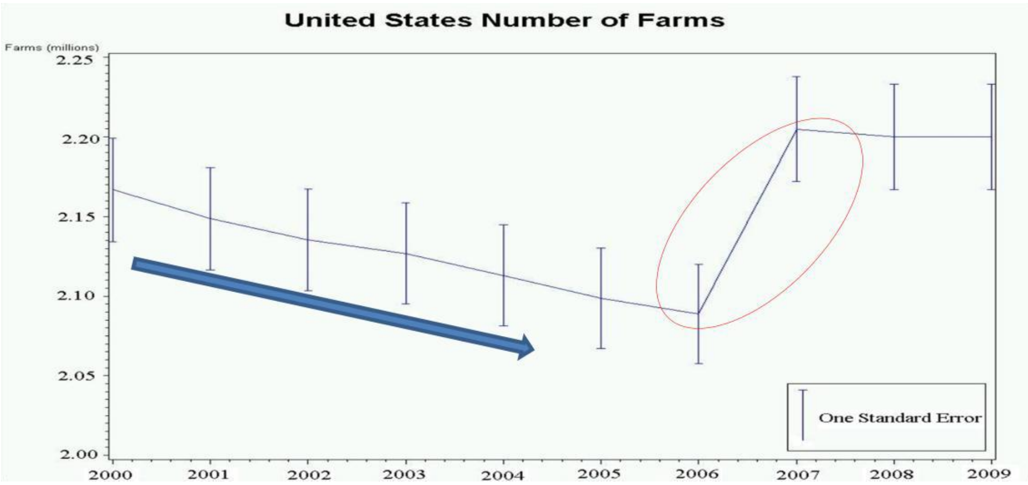
When the surveys are conducted with a panel of agricultural holdings selected from the data of the most recent general agricultural census, the discrepancies between census and survey data could be ascribed to the disappearance, division, or merger of holdings over time due to endogenous or exogenous events. Phenomena occurring in the population may also impair sample quality. These changes adversely affect panel quality because they directly influence sample size and the weight of the statistical units (Global Strategy, 2015).

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 US, the census data consist of responses to a list-based survey, the mailing list for which is created and maintained wholly independently of the JAS area 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 NASS matched each 2012 JAS tract to its 2012 census record. Disagreements in the conferral of farm status between the census and the JAS occurred when (1) tracts identified as non-farms in the JAS were subsequently identified as farms in the census, or (2) tracts identified as farms in the JAS were identified as non-farms in the census. If the tract was identified as a farm in either the JAS 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 JAS. However, in 2009, the Farm Numbers Research Project (FNRP) was conducted. Twenty per cent of the new JAS records were revisited, as these had been added to the sample and that had been 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 (see Figure 3.1).

FIGURE 3.1. Published estimates of the number of US farms from 2000 to 2009.



Source: Abreu et al., 2009.

c) Varying concepts and definitions

In an integrated agricultural statistics system, it is recommended that concepts and definitions be harmonized between agricultural censuses, other censuses (such as population censuses) and agricultural statistical surveys. Inconsistencies in data may be due to changes or variations of 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.

d) Greater reliability of data from 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 gained in previous collection efforts, and thus lead to discrepancies.

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, GPS, tablets, etc.).

e) Non-response

Non-response occurs in all censuses and surveys. To address the problem, several countries estimate the missing data, even though this increases the uncertainty associated with the estimates and may lead to bias. In the US, reporting is mandatory for the census, but is voluntary for surveys. However, legal measures are usually not invoked, to avoid the spectacle of prosecuting farmers. Censuses thus suffer a non-response rate similar to that of surveys. To take into account this non-response, the NASS adjusts the weights for responding records. This also increases uncertainty and may result in bias.

f) Other non-sampling errors

Other non-sampling errors may arise due to inadequate questionnaires or defective methods of data collection, tabulation, coding, etc.

g) Sampling errors

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

Sampling errors arise solely from the drawing of a probability sample, and not from the conduction of a complete enumeration. The methods to address these errors may determine a gap between census and survey data. Sampling errors may be linked to several factors, including a lack of representativeness due to insufficient sample size, errors in the sample selection process or a failure to validate some assumptions made in the sampling theory. For example, in two-stage sampling, the selection probability of an SSU is the product of the selection probability of the corresponding PSU and the conditional selection probability of an SSU for the given PSU. If PPS sampling is applied, this probability is proportional to a measure of size. This measure of size, seen as an auxiliary variable, should at least be positively correlated to the variable of interest, to reflect the correct weights of the sampled units in the population. This means that in repeated PPS sampling, the Horvitz-Thompson estimator usually used to compute estimates during survey operations is an unbiased estimator for the finite population total. However, if the probability of inclusion and the variable of interest are not closely related, this procedure may be rather inefficient due to variation in the selection probabilities. For example, if the measure of size is the number of agricultural households in an Enumeration Area (EA) and the variable of interest is the area harvested, it must be assumed that the number of agricultural households in the EA is at least positively correlated with the area harvested, to ensure that valid sampling weights are obtained[1]. The contrary is also possible, and a sample based on this auxiliary variable should lead to biased estimates of the variable of interest. This generates inconsistency with the data from the new census.

3.2. Difficulties and challenges

To perform data reconciliation, all of the detailed data used to compute estimations during the survey must be available, including all data on estimates, sampling weights, farms sampled, etc. Therefore, institutions should be capable of providing such data for all recommended 10 years. In some countries (mainly developing countries), the intercensal period may be excessively long, reaching even 30 years. In these cases, reconciliation is a very difficult task, as the availability of information (data and metadata) on censuses and subsequent surveys cannot be guaranteed.

[1] The stratification enables a positive correlation to be achieved between the number of holdings in the EA and the area size; however, the stratification may no longer be valid if several years have since passed.

According to FAO recommendations, some data are collected within a thematic module. These data are not collected by means of a complete enumeration, but are, rather, estimates from samples. In these cases (e.g. horticulture), data reconciliation concerns only survey data and can be a difficult task to undertake.

3.3. Conclusion

This chapter has identified situations in which inconsistencies between census and survey data may arise. The discrepancies between the two types of data may be linked to sampling errors or to non-sampling errors. Sampling errors occur only when a probability sample is drawn, and not when a complete enumeration is conducted. Non-sampling errors, on the other hand, arise mainly due to misleading definitions and concepts, inadequate frames, unsatisfactory questionnaires, or defective methods of data collection, tabulation, coding, incomplete coverage of sample units, etc.

4

Methods for reconciling census and survey data

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, these weights play an important role in creating consistent aggregates over time. Surveys select different holdings with different inclusion probabilities due to both intentional design and accidental factors. Some farms are therefore overrepresented compared to others; if the sample estimates are to reflect the population accurately, each farm must be weighted according to its 'true' inclusion probability.

Each farm is weighted by the inverse of its probability of inclusion in the sample (Deaton, 1997). This is reasonable because a household with a low probability of selection represents a large number of households in the population, while a household with a high probability of selection tends to be a minority-type household in the population. These weights are often referred to as "raising" or "inflation" factors, because they inflate the sample to resemble the total population. Divergences in weights across households arise from differences in selection probabilities, which may be ascribed, in turn, to both planned and accidental factors. Accidental differences may arise due to measurement errors and sampling errors, such as use of an obsolete sampling frame or non-response.

Post-stratification adjustment (adjustment to the weights following data collection) seeks to account for these accidental errors by benchmarking the survey data to external aggregate data. However, unlike model-based methods of data reconciliation, the post-stratification adjustment is not well defined, because the relationship between the data to be adjusted (survey data) and external data depends mainly on the nature of the latter. The statisticians appointed must decide which relationship to use, depending on the type of external data available. Therefore, this aspect is open to judgement – and thus to error.

4.1. Design-based methods

4.1.1. Post-stratification method

Post-stratification may be considered as a form of re-weighting. Post-stratification incorporates any form of data adjustment that organizes data into homogenous groups after data collection; however, it is usually performed when external information on these groups is available. Post-stratification adjusts the survey design weight within chosen subgroups, such that the sample reproduces the known population structure.

Post-stratification has three main functions: 1) reducing biases due to coverage and non-response error; 2) constituting a part of the sample design; and 3) potentially, increasing the precision of estimates that present a high correlation with the auxiliary information. As such, post-stratification introduces consistency between the results of surveys and those of other sources. The strata can be rebuilt using information from censuses and the sampling weight can be adjusted accordingly. This adjustment is used to correct any imbalance that may arise between sample design and sample completion – which may occur if the distribution of sample respondents within the external categories differs from that within the population (e.g. if subgroups respond or are covered by the frame at different rates) – and to reduce potential bias in the sample-based estimates.

Post-stratification can be used to adjust the survey sample data to make it more consistent with the population's structural parameters based upon the census data.

4.1.2. The ad hoc trimming method

The ad hoc method establishes an upper cut-off point for large weights, reducing weights larger than the cut-off weight to the cut-off value and then redistributing the weight in excess of the cut-off to the non-trimmed cases. This ensures that the weights before and after trimming add up to the same totals. The specific methods chosen for this process depend on how the cut-off is chosen. The underlying assumption is that a decrease in the variability caused by the outlying weights offsets the increase in bias incurred by the units that absorb the excess weight.

Chowdhury et al. (2007) describe the weight trimming method used to estimate proportions in the US National Immunization Survey (NIS). In 2012, the cut-off used was the median (w_i) +6*IQR (w_i), where median (w_i) and IQR (w_i) are, respectively, the median and the inter-quartile range of the weights.

To reduce extremely large Horvitz-Thompson estimator weights, Potter (1988) proposed trimming all weights $w_i > \sqrt{c\sum_{i\,\epsilon S} w_i^2/n}$ to this cut-off value. This method was used in the 1986 National Association of Educational Progress sample (Johnson et al., 1987). The value of c is "arbitrary and is chosen empirically by looking at values of $nw_i^2/\sum_{i\,\epsilon S} w_i^2/n$" (Potter, 1988). The sum of squared adjusted weights is computed iteratively until no weights exceed the cut-off value; then, the winsorized weights replace the initial weights to estimate the total. Potter (1990) claims that this method outperformed alternatives that minimized the mean square error (MSE), although it does not incorporate the survey response variables of interest.

The ad hoc trimming method is easily applied if it is clear from the census that some holdings were overrepresented in the sample. The problem with the method is the subjectivity necessarily introduced in redistributing the weight exceeding the cut-off value.

4.2. Model-based methods

The methods presented below incorporate models into estimation processes in very different ways. Each method has an implicitly defined formula to compute the sampling weight for the purposes of reconciliation. The following models have been designed to estimate parameters from census data, which are used to build sampling weights applicable in the year of the survey.

4.2.1. Best Linear Unbiased Prediction (BLUP) method

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

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

$$E(Y_i|X_i) = X_i^T\boldsymbol{\beta},\ Var(Y_i|X_i) = \sigma^2 D_i\,,\ \ Cov(Y_i, Y_j) = D_{ij}\sigma^2, i \neq j\,, \qquad (1)$$

where X_i denotes a p-vector of benchmark auxiliary variables for unit i that is known for all population units over the inter-census period. Auxiliary variables could be the labour (workforce), the size of the farm in terms of the number of persons, a dummy for the size of the farm land (from cadastral sources), etc. 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 $\mathrm{X} = [X_s, X_r]^{T}$, 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.

Valliant et al. (2000) show that the optimal estimator of a total is

$$\hat{T}_{BLUP} = 1_s^T Y_s + 1_r^T X_r\, \widehat{\boldsymbol{\beta}}.$$

To reconcile survey and census data, $\widehat{\boldsymbol{\beta}}$ is estimated using the census data and the weights are calculated accordingly. $\hat{T}_{BLUP}$ is the aggregate data.

Since the efficiency of the BLUP method depends on how well the associated model holds, this method may be susceptible to model misspecification. To overcome the potential bias therein, other methods have been developed.

4.2.2. Robust BLUP Method

Chambers et al. (1993) have proposed an alternative to the BLUP approach, in which a model-bias correction factor applies to linear regression case weights. 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, and it is applied to the BLUP estimator.

The model is $Y_i|X_i = m(X_i) + v_i e_i$, with $Var(Y_i|X_i) = \sigma^2 D_i$ as defined above.

Chambers et al. (1993) show that

$$\hat{T}_{RoBLUP} = \hat{T}_{BLUP} + \sum_{i \in s} [X_i^T E(\boldsymbol{\beta}) - m(X_i)],$$

where $\widehat{\boldsymbol{\beta}}$ is estimated using the ridge regression and base on the census data to enable the reconciliation.

The robust BLUP is model-unbiased under the preferred model, whereas the BLUP is not; in addition, the $\boldsymbol{\beta}$ parameter estimates are less susceptible to influence by extreme observations. This may be of assistance when editing errors occur during data collection.

4.2.3. Difference estimator method

Firth and Bennett (1998) have produced a bias-correction factor similar to that formulated by Chambers et al. (1993), for a difference estimator as follows:

$$\hat{T}_{DE} = \hat{T}_{BLUP} + \sum_{i \, \epsilon \, s} [(\pi_i^{-1} - 1)\,(y_i - X_i^T \widehat{\boldsymbol{\beta}})],$$

where $\widehat{\boldsymbol{\beta}}$ is estimated with the BLUP model and π_i is the sampling weight used during the survey.

The estimator is model-unbiased under the BLUP model, but smooths the effects of influential observations; in addition, it is approximately design-unbiased.

4.3. Model-assisted weighting methods

4.3.1. Cross-entropy estimation method

In samples selected from a finite population, auxiliary variables with known population totals may often be observed. The known population totals are usually obtained from external sources, such as administrative data or censuses. Calibration estimation can be described as a method to adjust the original design weights so that the known population totals of the auxiliary variables may be incorporated. Generally, the calibration procedure selects the adjusted weights that minimize distance between the original weights and the adjusted weights, while also satisfying a set of constraints relating to the auxiliary variable information. Kim (2009) has proposed a new type of empirical likelihood calibration estimator that preserves the maximum likelihood interpretation under Poisson sampling.

Entropy estimation, which is a calibration estimation method, uses *all the information available from the data, but nothing more.* The design weights fail to account for accidental changes in sampling probability, and therefore do not inflate the sample on the basis of the population. The cross-entropy approach recalculates the weights to account for these accidental changes; in other words, it makes the sample resemble the population, but at the same time maintains the adjusted weights as similar to the original weights as possible.

The estimation procedure is similar to that presented in Section 4.2 above.

Robilliard and Robinson (2001) present an approach to reconciling household surveys and national accounts data that is based upon the assumption that the macro-data represent control totals to which the household data must be reconciled. Upon this approach, the issue is how to use the additional information provided by the national accounts data to re-estimate the household weights used in the survey, such that the survey results are consistent with the aggregate data and the errors in the aggregates may be simultaneously estimated. The estimation approach is an efficient "information processing rule" that uses an estimation criterion based on an entropy measure of information. The survey household weights are treated as a prior. Using a cross-entropy metric, new weights are estimated to be close to the prior and to be consistent with the additional information. This additional information concerns the probabilities' adding-up normalization constraint and a consistency constraint on the moments. Using this method, information from the census can be capitalized to adjust survey sampling weights.

The challenge lies in identifying the correct moment consistency constraint. For example, with regard to livestock reconciliation data, the intercensal growth rate between two censuses may be used to estimate an aggregate value in the survey year. Therefore, a moment consistency constraint can be determined by means of this aggregate.

4.3.2. Generalized Regression (GREG) method

The Generalized Regression (GREG) method is a calibration approach that requires minimizing a distance function between the base weights and final weights to obtain an optimal set of survey weights. In this context, the term "optimal" is taken to mean that the final weights produce totals that match external population totals for the auxiliary variables *X*, within a certain margin of error.

Specifying alternative calibration distance functions produces alternative estimators. A least-squares distance function produces the following GREG estimator:

$$\hat{T}_{GREG} = \hat{T}_{HT} + \hat{\boldsymbol{\beta}}^T (T_X - \hat{T}_{XHT}),$$

where $\hat{\hat{T}}_{XHT} = \sum_{i \in s} w_i x_i$ is the vector of the Horvitz-Thompson totals for the auxiliary variables, $\hat{T}_X = \sum_{i \in s} x_i$ is the corresponding vector of known totals, and $\hat{T}_{HT}$ is the Horvitz-Thompson estimator used to estimate the total of the variable of interest during the surveys.

$\boldsymbol{\beta}$ is the parameter of a linear regression using census data. The functional relationship between Y and X is assumed to be the same for the census and the survey, for the purposes of data reconciliation.

4.3.4. Spline method (Robust GREG)

The Robust GREG method uses the regression model $Y_i = m(X_i) + e_i,\ e_i \sim N(0, D_i)$, where m is the spline function using a linear combination of truncated polynomials. The degree of the spline is p. Henry and Valliant (2012) show that

$$\hat{T}_{Spline} = \sum_{i \in s} w_i\, Y_i$$

with

$$w_i = \pi_i^{-1} - \left(\frac{\hat{m}_i}{\pi_i} - \frac{\sum_i \hat{m}_i}{n} \right) / Y_i\ ,$$

where π_i is the sampling weight used during the survey.

With this semi-parametric model, units with the same characteristics X will have closed estimates of the variable of interest. In this case, extreme values due to misspecification are reduced.

4.4. Growth rate method

Djety and Akoua (2008) present two approaches for reconciling census and survey data, both of which are based on the growth rate. The first is applied to the area and the yield; then, the production is computed as the product of these variables. The second is directly applied to the production. The two methods are calculated as follows:

First method (*méthode de la DSDI*):

$P_t = (1+ T_S) S_{t-1} \times (1 + T_R) R_{t-1}$

with:

P_t : Production in year *t*.

T_S : Average annual growth of the area.

S_{t-1} : Area size in year *t*-1.

T_R : Average annual growth of the yield.

R_{t-1} : Yield in the year *t*-1

Second method (méthode de l'étude):

$P_t = (1+ T_p) P_{t-1}$

with:

P_t : Production in the year t.

T_p : Average annual growth of the production.

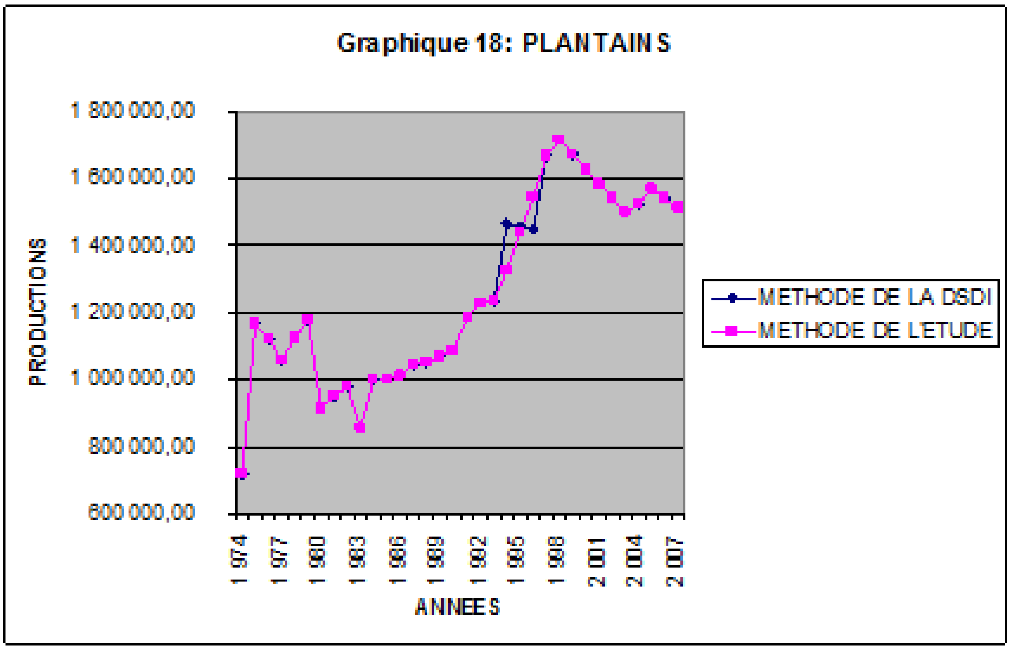
P_{t-1} : Production in the year t-1.

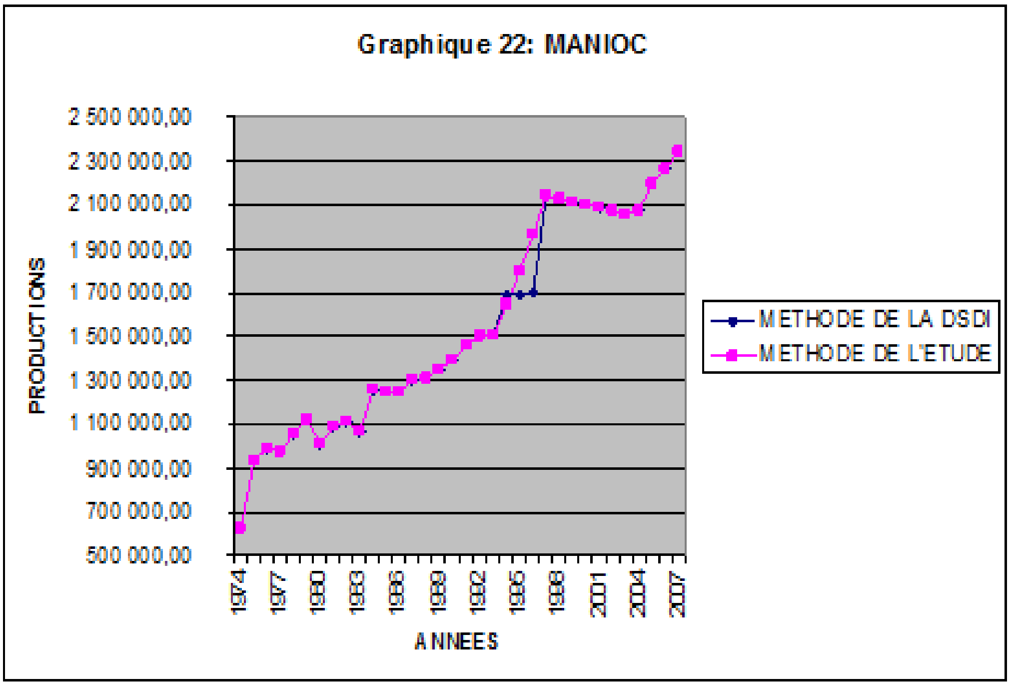
In the cases studied by the authors, the two methods yield similar results for certain crops (such as plantain, groundnut and fonio). For other products, either the first or the second method provides better results (e.g. respectively, millet and cassava). These techniques must be tested by means of simulation before a recommendation can be made.

These methods reconcile data at the national level; therefore, no information is available at subnational level. It may be preferable to construct this estimator at the stratum level, as the data at national level could then be computed as the aggregate of all strata.

Another issue with these methods is that no information is provided on how the average annual growth rate T_R is computed. If it is based on data from the two censuses alone, it fails to incorporate the survey data. This means that all the information from the surveys is lost.

FIGURE 5.1. The two methods, used for plantain and cassava crops





Source: Djety & Akoua, 2008.

After several simulations based on different data sources, the method applying a growth rate to the RGAC data of 1987 was chosen. The growth rate was computed using data from the RGACs of 1987 and 2005. The evolutionary trends from 1970 to 1986 were observed, taking into account the years of pastoral crisis. These were 1973 to 1974 and 1983 to 1984 in particular, when the highest livestock mortality rates were experienced due to drought and lack of pasture (Harouna, 2009).

The results showed that cattle production increased by more than double, compared to the previous estimates (100.3 per cent); the production of sheep rose by 24.3 per cent compared to the previous estimate; goat production by 19.4 per cent; camels by 33 per cent; and asses of 90.3 per cent. The rearing of horses recorded a decrease of 9.4 per cent due to the scarcity or disappearance of the practice of horse breeding (Harouna, 2009).

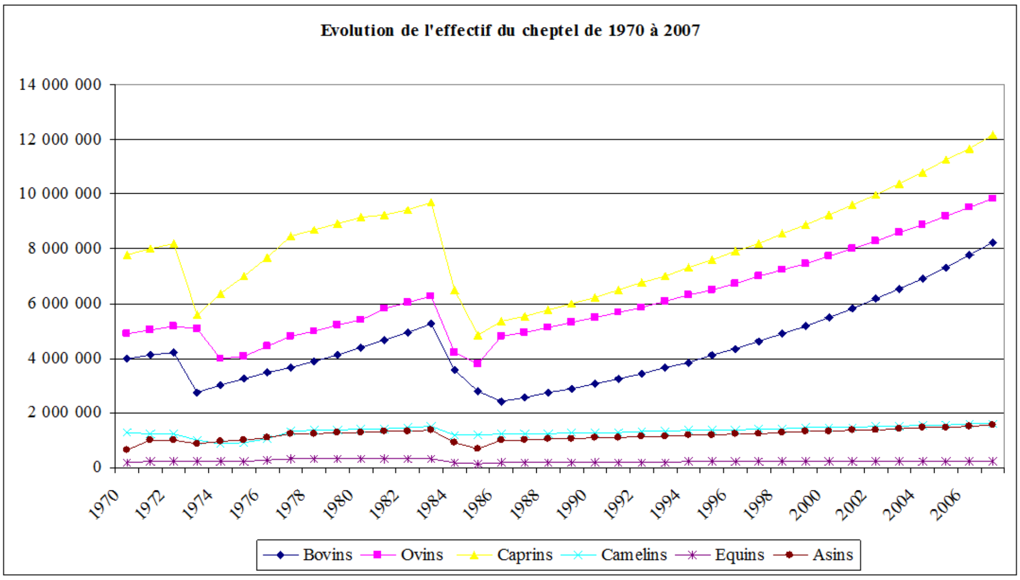
TABLE 5.1. The growth rate used for reconciliation (Niger).

	Growth Rate before 1987	RGAC Growth Rate (1987-2005)
Cattle	Variable rate	6.0 %
Sheep	Variable rate	3.5 %
Goats	Variable rate	4.0 %
Camel	Variable rate	1.3 %
Asses	Variable rate	2.0 %
Horse	Variable rate	1.0 %

Source: Harouna, 2009.

The increase in the level of livestock has also led to a significant improvement in the sector's contribution to national GDP. Thus, the added value of livestock increased from 9.7 per cent of GDP (previous estimate) to approximately 13 per cent, when the data from the RGAC 2005 was taken into consideration (Harouna, 2009).

FIGURE 4. Results after Reconciliation (Niger).



Source: Harouna, 2009.

4.5. Handling misclassification

Lamas *et al.* (2011) identify misclassification as a direct cause of the undercount of the number of farms produced by the JAS in the US. One approach to correct for this undercount is to use the NASS's sampling list frame, which is independent of the area frame. However, the list frame does not present a farm/non-farm status classification. Abreu *et al.* (2011) used matched records from the 2009 JAS, the 2009 list frame, and the 2009 Farm Numbers Research Project (Abreu *et al.*, 2010) to explore the characteristics of the inaccuracies in the list frame farm status. They then developed an estimator of the probability that a 2011 list frame record was a farm using logistic regression, and used this estimator as a foundation for providing an adjusted number of farms for the 2011 JAS. The two estimators were based upon two assumptions: (1) the adjustment was independent of the original JAS estimator of the number of farms; and (2) the previous census farm rates provided a good estimate of the probability of farm status for each list frame record. However, both of these assumptions were questionable.

To address the concerns raised by the previous approach, and to obtain a coherent set of methods for the agricultural census and the JAS, Abreu *et al.* (2014) developed a capture-recapture approach to estimate the number of US farms from the JAS. They proposed the following estimator for the number of farms from the JAS, 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 JAS

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 US level, the estimated misclassification rate for farms was 9.4 per cent.

4.6. Non-response

Generally, in 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 of adjusting for non-response consists in multiplying the respondent's sampling weight by the inverse of the estimated response probability. Kim and Kim (2007) demonstrate that this approach is generally more efficient than relying upon 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 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 NASS 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 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.

4.7. Other data adjustment techniques

The methods presented in this section are techniques of adjustment that may be performed on survey data, as required. However, reconciliation of the survey data with the census data may still be necessary after these techniques are applied. Subsections (a) to (d) below discuss some problems that may affect the statistical unit, together with possible solutions. *These are to be implemented when the survey is being conducted.*

a) Additional samples

Due to 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 these units is available (e.g. 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. These 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 will have to be tracked, especially if its absence is not random. For example, if a portion of a farm changes ownership due to 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 developed by Lavallée (2007). These 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, these 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 to anticipate the loss of statistical units. This helps to maintain sample accuracy, but 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 its implementation, it may still be necessary to proceed to the reconciliation with census data.

4.8. Country experience: Canada

Not all agricultural survey results should be changed when agricultural census estimates are compared. Indeed, the sampled units of some surveys may not be the farm operator (but millers, for example), or some survey variables may not be measured by the census (such as greenhouse area). Consideration is given to historical events that may have introduced a supply or demand shock between census years, to maintain the characteristics of such events during the revision. However, if a shock occurs during the census year, this information will not be used for trend adjustment. In addition, the source of the information will affect decisions on a possible update. For example, administrative data generated from regulatory sources that are widely used across the industry are likely to remain unchanged, unless a clear explanation can be provided.

General considerations

Data from agricultural censuses is used for benchmarking at macro level and for data confrontation and verification. The survey estimates are revised to match the census numbers as closely as possible, adjusted for seasonal variation as appropriate. The revisions made on commodities can be summarized as either a wedge adjustment or a logarithmic adjustment, depending on the characteristics of the data and the commodity. Only the trend is adjusted – not the magnitude of the change from year to year. Variables such as area (and, in some cases, expenses) are first compared between surveys and the agricultural census, to determine the extent of the frame change and the potential intercensal adjustments.

Ratios are also used in various ways for the commodities, to support their analysis: (a) the ratio of published numbers to census numbers; (b) the ratio of census numbers to survey-level estimates; (c) the ratio of average yield (from the survey) and total area (from the census), to adjust production; (d) the census inventory data adjusted for seasonal variation (for e.g. cattle and sheep), etc.

When reconciling, the supply and demand outputs are respected as much as possible. Crop supply and disposition tables can still be revised to maintain balance and validate production, in light of any changes that may have occurred in the relevant area.

The livestock balance sheet follows a similar procedure, examining international and inter-provincial trades, inventory and slaughter. For cattle, adjustments are made to “softer” categories such as calves and heifers. Similarly, in financial terms, the agricultural census may trigger revisions for intercensal years to capital value, farm cash receipts and operating expenses, in light of the new production and inventory values fed from the commodity-adjusted estimates. The intercensal revisions provide an opportunity to include modifications to compilation methods or concepts that have not yet been integrated in published data. Census data is also used to revise the value of a number of commodities for which annual data is not available.

The expense benchmarks established during intercensal revisions are typically within 2 per cent of the census estimates. The trends and levels of tax-based estimates (the source of annual estimates of agricultural expenses) are taken into account when determining the exact level, and indicators, of input price and quantity changes. Information on undercoverage, edit, imputation and validation procedures and the historical relationships between tax and census levels are taken into consideration, as are any changes in the questionnaire (e.g. grouping of expense items). Once the benchmarks have been fixed, a smoothing process is applied which only slightly adjusts the annual changes of the intervening years.

The top contributors are compared, to identify the farms missing from the survey frame. As for the census estimates, this enables any changes in subsectors or emerging agricultural sectors to be better identified. This also provides an opportunity to address these changes in survey questionnaires for future years. For a given commodity or geographic area, in future sample selection, a respondent may be included in a different stratum, in light of its relative importance since the previous census.

Census validation using 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 before being released to the Canadian public. Data validation is a complex process in which human judgement is vital. Validators follow a Data Validation Plan and a Data Validation Checklist as guidelines to the data validation tools available on the Central Processing System (CPS). However, validators will ultimately have to solve problems and make decisions based on the analysis of background information, respondent feedback, expert consultation and common sense.

First, the analysis is focused at the macro level. Aggregate census data are analysed at the provincial and subprovincial levels and compared to the expectations outlined in the senior validator's Data Validation Plan.

The analysis is then directed to the micro level. Changes to individual records must be made when appropriate, to guarantee the quality of provincial and small-area data and the usefulness of the agricultural census data as a sampling frame. Due to resource and time constraints, micro editing is done using a "top-down" approach, in which those records with the largest contribution to a variable estimate are reviewed first.

Finally, the results of analysis for a province – including the final estimates of the variables under study and recommendations for their publication – are presented to a certification committee.

Certification

Revised survey estimates are verified by other members of the team. Provincial experts are also consulted to obtain their views on the possible extent of revision.

Communication plan

A communication plan is established to inform all key users that new intercensal revisions have been made available. Typically, users know that estimates are revised every five years.

Timelines

Intercensal revisions to agricultural commodities are usually completed one to two years after the census data are released. Corresponding revisions to the financial variables (farm cash receipts, operating expenses and net income) are released two to three years after the census data release. Revisions from a new census benchmark normally cover the five-year period back to the previous census. (Statistics Canada, 2011)

Lessons Learnt

Data reconciliation techniques such as ratios and trends may be useful when revising survey data. Furthermore, these 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.

4.9. Conclusion

This chapter examined various methods discussed in the literature that could be used for data reconciliation. The results of the application of some of these methods in some developed and developing countries have also been seen. Three kinds of methods have been distinguished: (i) the design-based methods; (ii) the model-based methods, and (iii) the model-assisted weighting methods. In Chapter 5, a gap analysis will examine the main limitations of these methods and the alternatives that can provide countries with clear guidance on data reconciliation.

5

Gap analysis

Data consistency is an essential requirement of any agricultural statistics system. To achieve data consistency, several countries reconcile census and survey data, although some do not provide documentation on the methodologies applied. The methodologies currently used to reconcile data are rebasing techniques, i.e. the adjustment of sampling weights. In other words, the adjustment of sampling weights is the crucial point of the reconciliation. These methodologies were discussed in Chapter 4.

As mentioned in the literature review, there are different situations in which data reconciliation is necessary, such as:

- **Variation in the number of holdings since the last census:** Changes in the number of holdings since the most recent census may generate a gap in the data trend when a new census is performed.
- **Technological improvements:** these, or changes such as the use of improved seed, could generate changes that cannot be captured during intercensal years. Only a new census would be capable of showing whether that the changes that have occurred are structural. These changes may happen even if the number of holdings remains constant or the sampling frame remains valid.
- **Misclassification:** As discussed above, this is a source of inconsistencies between census and survey data. NASS implements a methodology to estimate the probability that a tract has been misclassified. It is thus possible to estimate the total number of holdings in a given year.
- **When a panel sample is used over the years**, discrepancies may arise due to changes in the statistical units (i.e. their fusion, division or disappearance).

Cases of non-response have been discussed as a possible cause of a gap between census and survey data (Lopiano et al., 2011). Since methodologies are used after a survey to adjust data when non-response occurs, non-response cases should be considered as a problem of estimation. Estimation errors are related to the methodologies used, and may be the reason for a gap between census data and survey data.

These situations may also all occur at the same time. An appropriate methodology should be developed to address these issues.

Another important point is to identify the relationship between the different methodologies of data collection and the presence of a gap in data over a year to determine those that are most likely to generate a gap: whether the area frame or the list frame, etc.

5.1. Overall gap analysis

Developed countries and some developing countries have established systems for the estimation of crop area and production. These systems are based on censuses and sample surveys. Globally, although there may be substantial variation in the methods and practices used for agricultural data reconciliation, these are mainly based on the adjustment of sampling weights. Therefore, it is important to study these related techniques in detail, and to examine the specific method to be applied in each country and under different conditions. Some of the major issues upon which gaps in these methods can be identified, in terms of ascertaining the suitability of each method to be adopted by a particular country are accuracy, cost, complexity, timeliness and availability of existing data.

The differences in the methodologies explained above lie in the quantities of information required for the reconciliation. For some of these, information must be gathered only on one census; for others, two consecutive censuses are required. Additional information from other administrative sources may also be required.

Another major issue is the cost involved in framing the methodology for data reconciliation. It is recommended that further research be conducted upon cost reduction options. When reconciling census data and survey data, the costs relate mainly to the gathering of data from different sources. Of course, the cost of a given methodology depends on the amount of information required.

A system's degree of complexity is also a major factor in determining whether it should be adopted. In terms of data reconciliation procedures, it is always preferable to adopt a simpler system. However, at the same time, efficiency is important.

Timeliness makes a method more desirable for any country. For each type of data, it is important to identify the variables required for reconciliation and to identify an appropriate way to ensure that they are readily available. For example, information gathered during the census on the date of creation of new holdings will be useful during a backward estimation.

Other opportunities for gap analysis arise in the following: (i) identification of the different methods to be applied in each country or region, since countries do not all have the same statistical system; (ii) characterization of variables of interest for each method; (iii) characterization of different sources of information, whether available or required; (vi) design of automated procedures to reconcile data.

Another major issue is to determine which census should be used for data reconciliation. The most reliable census should be chosen. However, upon the assumption that both censuses are reliable, an accurate method should be established to harmonize the estimation obtained using both censuses. For some methods, only data from one census are necessary for the reconciliation. In this case, the most recent census is used. If new estimates can be obtained from both censuses independently, a methodology must be identified to make full use of the information from both censuses.

The level of disaggregation and the coverage of the administrative data are also important factors. A good methodology should take into account the information contained in administrative data.

It is important to know how data have been produced, so that the data can be reconciled efficiently. Statistics based on sample surveys may sometimes be published for population totals for which the true values are known in advance from other sources, such as registers. The methods to calibrate the sampling weights are applied in such a way that the estimates from the sample must necessarily fit the true values exactly. The external information that is thereby incorporated in the weights may also help to improve the estimation of other quantities. Thus, some methods require additional data information sources. This means that it is assumed that these sources are reliable. Therefore, it is necessary to identify clear links between the agricultural census and other data sources (other censuses, administrative data, etc.).

Other possible methods could be based on trend adjustment and yield computation. The trend adjustment method generates estimates following a line, to reconcile inter-census data and the data from the two censuses. As for the yield computation, this may be obtained from the data of the new census. It is used to compute new estimates backwards in time, assuming that there have been no radical changes over the years.

5.2. Case of livestock

For the specific case of livestock data, to estimate livestock in a given year, some countries use the growth rate. This method has two clear disadvantages. First, the samples used to estimate are often rather small, which may lead to imprecise estimations. Second, analysis often requires annual growth rates, rather than an average annual growth rate over a 10-year period.

Bennett and Horiuchi (1984) have shown that it is possible to estimate the number of individuals of age *x* in a population at a given time. Preston & Bennett (1983) proposed a simple method for converting an age distribution of any closed population into the stationary population corresponding to its current mortality conditions. The Preston-Bennett method relates to one another the number of persons in any two age groups at any particular time, in terms of age-specific mortality conditions and growth rates. These components will normally be available from successive censuses. *The reference period need not necessarily be a year*. Preston and Bennett have shown that if errors resulting from differences in the completeness of census coverage or due to migration are present and constant for all ages, then all age-specific intercensal growth rates will present the same amount of error.

A similar model should be established to calculate the total population for each year of the inter-census period. Therefore, sampling weights can be adjusted according to the various techniques mentioned here.

To assess the validity of any given set of livestock statistics, several criteria must be met. First, after adjusting for trade and storage, domestic supply must be equal to demand. Second, the production of meat must be consistent with feed statistics. Third, the numbers must be consistent with trends observed in the economy.

5.3. Additional issues

In addition to the issues mentioned above, other areas must still be investigated. For example, if only one census is available and that the national institutions have previously relied on surveys to produce data. The difference between the methods illustrated above lies in how each estimates sampling weights. These are the most important element in the process of producing estimates from a survey. The appropriate methodology should be based on sampling weight estimation.

To ensure harmonization with census data, survey data must be estimated to reflect the structure of the census data. When auxiliary variables are used, the functional relationship between these variables and the variable of interest must be determined using the census data, and then applied to survey data. Since there are two censuses and it is sought to obtain estimates during the intercensal period, the appropriate methodology should take full advantage of all the information contained in both censuses. For a given year within the intercensal period, a weighting system that gives more weight to the information contained in the closest census could be investigated.

5.4. Conclusion

This gap analysis has allowed us to analyse the main limitations of some methods used for data reconciliation in certain countries. The main challenges and issues to be addressed in the process of reconciling census and survey data have also been outlined. It appears necessary to test some of these models with real data, to identify those that are most suitable to individual countries, and to provide guidance on how to implement them.

6

Conclusion

There is scarce published literature on reconciling census data and survey data in the field of agriculture. However, several techniques applied to produce sampling weight adjustment may be a basis for data reconciliation. This technical paper has reviewed some of these methods, their limits and the challenges to be addressed in their implementation. It has also explored the sources of discrepancy between census data and survey data, and the gap to be addressed to provide countries with guidelines on data reconciliation.

Reconciliation techniques should be applied to data taking into account certain inherent differences in their nature. When reconciling data, all of the issues noted in this document should be considered carefully. Methodologies that are efficient in terms of timeliness and cost must be established.

In some of the examples presented in this literature review, explicit formulas for weights could be obtained. Methods that incorporate realistic models will improve the estimates of totals. By incorporating the relationship between the survey variable and some known auxiliary information, the estimates of the totals may have lower mean square errors. When the model is specified correctly, the associated estimators are optimal. However, when the model does not hold, or if the sample contains outliers, several robust alternative estimators have been developed.

The generalized design-based method smooths weights by modeling them as functions of the observations y. The weight of each unit is then replaced by its regression prediction. Non-response and post-stratification methods are designed to reduce biases or variances.

All of these methodologies should be tested to identify the most suitable ones for individual country situations, and to provide countries with effective and workable guidelines.

References

Abreu, D.A. 2007. *Results from the 2002 Classification Error Study*. Research and Development Division. RDD Research Report No. RDD-07-03. USDA/NASS Publication: Washington, D.C.

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., Sang, H., Lopiano, K.K. & Young, L.J. 2011. *Adjusting an Area Frame Estimate for Misclassification Using a List Frame*. Proceedings of the 2011 Joint Statistical Meetings. Section on Survey Research Methods – JSM 2011. 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.

Abreu, D.A., Lamas, A.C., Young, L.J., Wang, S. & Adrian, D. 2012. *Estimating the Number of U.S. Farms with Adjustment for Misclassification*. Proceedings of the 2012 Joint Statistical Meetings. Section on Survey Research Methods – JSM 2010. American Statistical Association Publication: Alexandria, VA, USA.

Abreu, D.A., Busselberg, S., Lamas, A.C., Barboza, W. & Young, L.J. 2014. *Evaluating a New Approach for Estimating the Number of U.S. Farms with Adjustment for Misclassification.* Proceedings of the 2014 Joint Statistical Meetings. JSM 2014 –Survey Research Methods Section. USDA American Statistical Association Publication: Alexandria, VA, USA.

Bennett, N.G. & Horiuchi, S. 1984. Mortality estimation from registered deaths in less developed countries. *Demography*, 21(2): 217-34.

Berg, E. & Li, J. 2015. *Administrative Data and the Statistical Programmes of Developed Countries*. Technical Report No. 2. Global Strategy to Improve Agricultural and Rural Statistics (GSARS) Publication: Rome.

Chambers, R.L., Dorfman, A.H. & Wehrly, T.E. 1993. Bias Robust Estimation in Finite Populations Using Nonparametric Calibration. *Journal of the American Statistical Association*, 88: 260-269.

Chowdhury, S., Khare, M. & Wolter, K. 2007. *Weight Trimming in the National Immunization Survey*. Proceedings of the 2007 Joint Statistical Meetings. Survey Research Methods Section. American Statistical Association Publication: Alexandria, VA, USA.

Deaton, A. 1997. *The Analysis of Household Surveys: A Microeconometric approach to development policy.* World Bank Publication: Washington, D.C.

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, Ghana.

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

FAO. 2005. *A system of integrated agricultural censuses and surveys. Volume 1: World Programme for the Census of Agriculture (WCA) 2010.* FAO Publication: Rome.

FAO. 2015. *World Programme for the Census of Agriculture (WCA) 2020: VOLUME 1 Programme, concepts and definitions*. FAO Statistical Development Series No. 15. FAO Publication: Rome.

Firth, D. & Bennett, K.E. 1998. Robust models in probability sampling. *Journal of the Royal Statistical Society*, Series B, 60: 3-21.

Garber, S.C. 2009. *Census Mail List Trimming using SAS Data Mining. Research and Development Division*. RDD Report No. RDD-09-02. USDA/NASS Publication: Washington, D.C.

Global Strategy to Improve Agricultural and Rural Statistics (GSARS). 2015. *Integrated Survey Framework*. GSARS 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.

Harouna, I. 2009. *Réconciliation des données existantes sur l'élevage avec les données du Recensement Général de l'Agriculture et du Cheptel (RGAC) au Niger.*
Presentation prepared for the African Commission on Agricultural Statistics (AFCAS) meeting, 28-31 October 2009. Accra.

Johnson, E.G., Hansen, M.H., Tepping, B.J., Lago, J.A. & Burke, J. 1987. "Weighting procedures". In Beaton, A.E. (ed.). *Implementing the new design: The NAEP 1983-1984 technical report* (pp. 493-504). National Assessment of Educational Progress Publication: Princeton, NJ, USA.

Kim, J.K. 2009. Calibration Estimation Using Empirical Likelihood in Survey Sampling. *Statistica Sinica*, 19: 145-157.

Kim, J.K. & Kim, J.J. 2007. Non-response weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*. 35(4): 501-514.

Henry, K. & Valliant, R. 2012. *Comparing Alternative Weight Adjustment Methods*. Proceedings of the 2012 Joint Statistical Meetings. Survey Research Methods Section. American Statistical Association Publication: Alexandria, VA, USA.

Huang, H. 1984. *Obtaining Cross-Sectional Estimates from a Longitudinal Survey: Experiences of the Income Survey Development Program*. Proceedings of the 1984 Joint Statistical Meetings. Section on Survey Research Methods. American Statistical Association Publication: Alexandria, VA, USA.

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

McCarthy, J.S., Jacob, T. & McCracken, A. 2010. *Modelling Non-response in National Agricultural Statistics Service (NASS) Surveys Using Classification Trees*. RDD Research Report No. RDD-10-05. USDA/NASS Publication: Washington, D.C.

Potter, F.A. 1988. *Survey of Procedures to Control Extreme Sampling Weights*. Proceedings of the 1988 Joint Statistical Meetings. Section on Survey Research Methods. American Statistical Association Publication: Alexandria, VA, USA.

Potter, F.A. 1990. *Study of Procedures to Identify and Trim Extreme Sample Weights*. Proceedings of the 1990 Joint Statistical Meetings. Survey Research Methods Section. American Statistical Association Publication: Alexandria, VA, USA.

Preston, S.H. & Bennett, N.G. 1983. *A census-based method for estimating adult mortality*. *Population Studies*, 37(1): 91-104.

Robilliard, A.-S. & Robinson, S. 2001. *Reconciling household surveys and national accounts data using a cross entropy estimation method*. Discussion Paper No. 50. International Food Policy Research Institute Publication: Washington, D.C.

Statistics Canada. 2011. *Guidelines for Data Validation Analysis.* Census of Agriculture (CEAG), 2011. Training manual. Statistics Canada Publication: Ottawa, Ontario, Canada.

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

Valliant, R., Dorfman, A. & Royall, R.M. 2000. *Finite Population Sampling and Inference*. John Wiley & Sons: New York, USA.