Reconciliation of censuses and survey data during the next round of agricultural census

Naman KEITA, Senior Statistician, Food and Agriculture Organisation of the United Nations
Viale delle Terme di Caracalla
00153 Rome, Italy
E-mail: Naman.Keita@fao.org;

Eloi OUEDRAOGO, Senior Statistician, Food and Agriculture Organisation of the United Nations
2, Gamal Abdul Nasser Road
Accra, Ghana
E-mail: Eloi.Ouedraogo@fao.org;

Ulrich E. NYAMSI. Statistician Consultant, Food and Agriculture Organisation of the United Nations
2, Gamal Abdul Nasser Road
Accra, Ghana
E-mail: Ulrich.Nyamsi@fao.org.

DOI: 10.1481/icasVII.2016.f35d

ABSTRACT

In many developing countries, Agricultural data are mainly derived from decennial censuses providing structural data on agricultural holdings and benchmark data serving as reference for furthers yearly estimates from sample surveys. When conducted through complete enumeration, agricultural censuses also provide a sampling frame for intercensus sample surveys. Samples for current agricultural surveys are drawn from the sampling frame of the last agricultural census to provide annual updates on some of the agricultural data items and variables such as agricultural area, production etc. These annual estimates are based on the agriculture structure during the last census.

Very often, when a new census is conducted, there are discrepancies between the time series coming from the annual surveys conducted since the last census and the results of the new census. Although discrepancies may be due to legitimate changes in the already dynamic agriculture structure, common sources of these discrepancies are related to changes in the sampling frame, survey methods, concepts and definitions. These especially occur when the intercensal period is too long (more than the ten-year period that is usually recommended by FAO).

Even though this is a very common problem, even in countries with more advanced statisticalsystems, there are very limited studies and methodological guidance to address the issuessystematically after each census.

This paper analyses the possible sources of the discrepancies between time series fromintercensus annual surveys and the results of the new census. It reviews the statistical methodsthat can be used to address them; taking into account country experiences, and results of asimulation conducted on real data form a pilot country. It will present strategies and methodologicaloptions that can be considered on the ways to systematically reconcile inter-censuses surveysdata and results of a new census. The paper builds on the content of a technical report prepared inthe framework of the Research Programme of the Global Strategy to Improve Agricultural andRural Statistics.

Keywords: Reconciliation, Regression, Sampling.

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.

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.

2. Objectives

The main objective of the study is to deeply explore the methodologies displayed in the literature review and to develop an appropriated statistical methodology for reconciling agricultural census and survey data. Simulated data are used to assess the proposed method.

3. Source of Discrepancies

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 forundercoverage, 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 farmsin 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.

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.

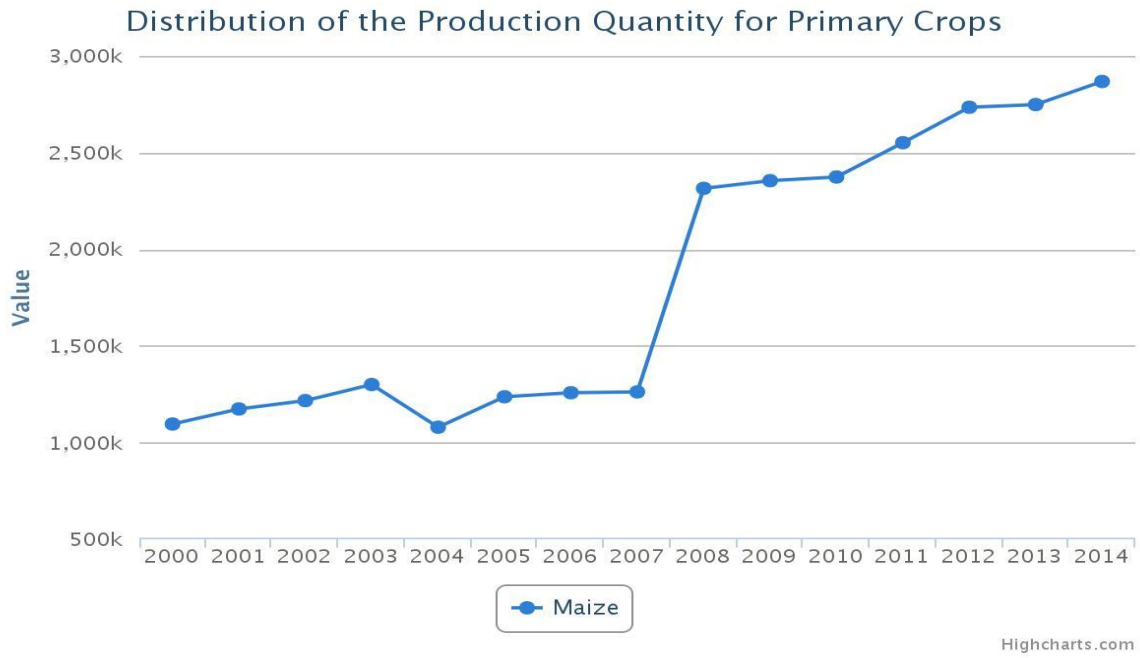

Figure 1. Example of discrepancy in time series data. Source: CountrySTAT-Uganda

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 theseerrors 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. 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.

4. Methodological Approach

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.

4.1. ChangepointDetection

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 minimise

$$\sum_{=1}^{+1}[\ (\ (\ _{-1}+1):\ )] + \quad (\ ) \qquad\qquad (\ )$$

where *C* is a cost function for a segment e.g., negative log-likelihood and *βf(m)* is a penalty to guard against over fitting. The changepointdetection could be implements three multiple changepoint algorithms that minimise (*a*); Binary Segmentation (Edwardsand Cavalli-Sforza, 1965), Segment Neighbourhoods (Auger and Lawrence, 1989) and therecently proposed Pruned Exact Linear Time

(PELT) (Killick et al., 2012). The R packages*changepoint*and *changepoint.np*could be used at this regard.

The figure 1 shows the plots of the time series for 4 crops. The point is to identified whether the year of the census is a changepoint. The table 1 presents the results for the selected crops. The method used is the PELT method, the empirical distribution is used to compute the statistical test and the Modified Bayes Information Criterion (MBIC) has value 8.124151. For all crops, 2007 is the location of the changepoint. The implementation of the 2008 censuscould explain the break in the time series data. For Rice Paddy, the method identifies 2007 as the changepoint, while on the plot is not clear. For sorghum, 2004 has been identified as a changepoint. Since any census have been done in this year, the break in the time series cannot be the result of the implementation of a new census. The method allows to identify a changepoint, but since we intend to reconcile census and survey data, only changepoint in the year of the census could be taken into consideration.

Crop	Maize	Sorghum	Cassava	Rice Paddy
Changepoint Locations	2007	2004, 2007	2007	2007

Table 1. Identification of the change point (number of quantiles=3)

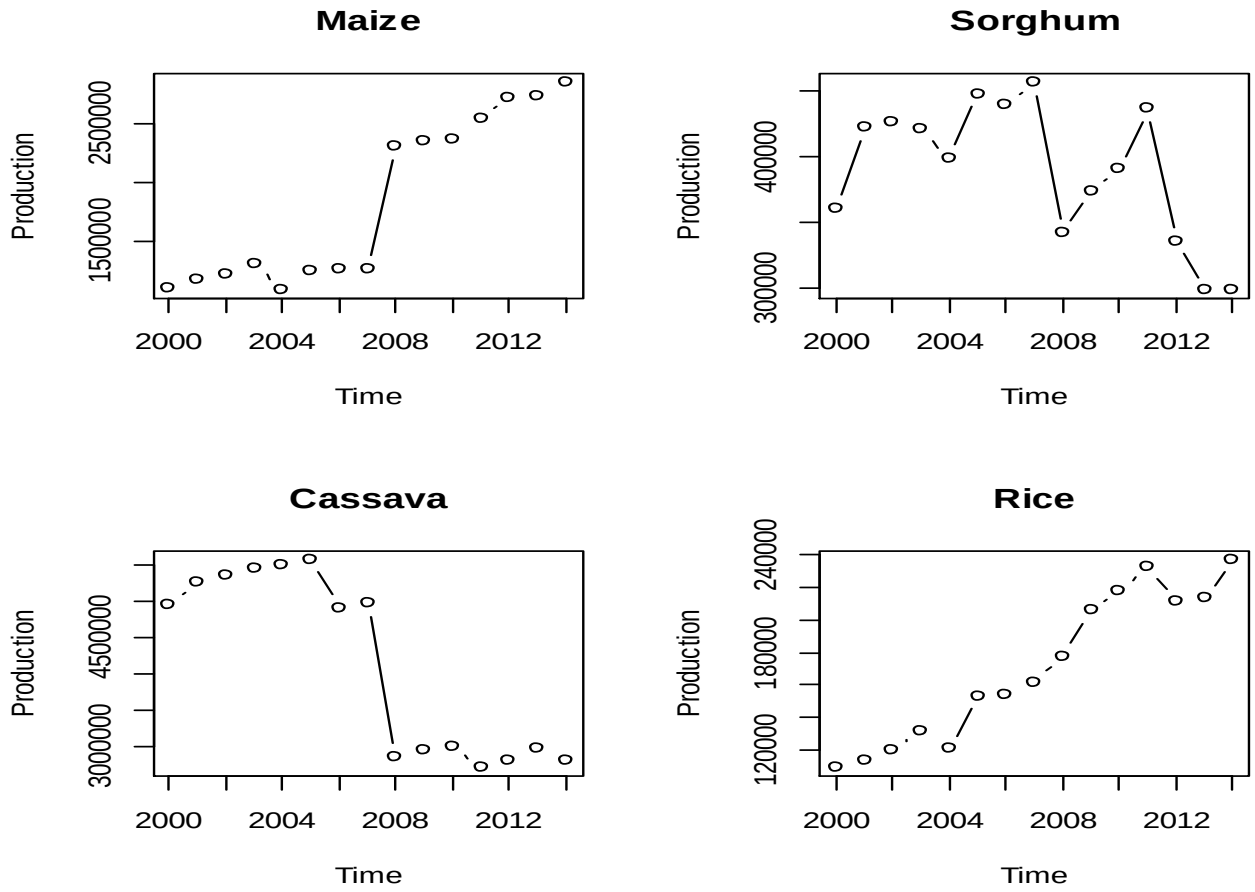

Figure 1. Plot of the time series of different crop. Source: CountrySTAT Uganda.

4.2. Cross Entropy Estimation Method

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.

The estimation approach represents an efficient "information processing rule" using 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.* These additional information is about the adding-up normalization constraint of the probabilities and a moment consistency constraint. Using this method, information from the census can be capitalized to adjust survey sampling weights.

In particular, the model consists of an objective function which is minimized subject to constraints. An extrapolation method can be used on census data in order to obtain a *prior* data in the year of the survey. This extrapolation can be done by using trend estimation.

$\sum$ is the original sampling survey weight for a given statistical unit, is the new sampling survey weight used for the reconciliation, is the prior obtained from an extrapolation based on census data, error weights estimated in the Cross-Entropy procedure, $\sum$, is its prior, $\sum$, is the error support set, represents a general aggregator and a probability or a sample weight.

$$, \quad \ln\left(— + \quad , \quad \ln\left(= \right. \right. \qquad (1)$$

subject to

$$= \quad + \sum \quad , \bar{} , , \quad \epsilon [1, \dots , \ ], \ \epsilon [1, \dots , \qquad (2)$$

and additional adding-up constraints on the error weights

$$= 1, \qquad , = 1 \qquad (3)$$

$$\{\bar{} , \ \} \leq \quad \leq \quad \{\bar{} , \ \} \qquad (4)$$

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:

$$= \quad , \bar{} , ^2$$

, is the prior weights on the error support set.

Assuming a prior distribution with zero mean and a standard error equal to , we used a support set with five terms equal to (-3 , - ,0, , 3). 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 is assumed to have a measurement error. The minimization is performed by a non-linear optimization algorithm. Constraint (4) makes sure that

lies between the original $\sum$ and . The minimization is performed by a non-linear optimization algorithm (Ouedraogo & Nyamsi, 2016).

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. A household consumption data could be used as auxiliary information.

4.3. Post-stratification Method

Post-stratification can be seen as a form of re-weighting. The post-stratification methodology is to re-consider the size of the strata based on the last census, and re-calculate the probability of inclusion, and therefore the sampling weight using new information based on the new census. Based on the information of the new census, new strata are "re-defined" (post-strata), and new weights are calculated considering the part of sample units included in the new strata. This "re-definition" is necessary, if the previous strata are no longer valid.

Let Y be the variable of interest and H the number of post-strata (After the redefinition of strata).

- $_h$ is the number of agricultural holdings in the post-stratum h
- $_h$ is the set of agricultural holdings in the stratum h
- $_h = \sum$ $_h$ is the total of Y in the post-stratum h

- Let $_h$ be the part of the sample included in the post-stratum h. The number of holdings in $_h$ is $_h$.
- $^{-}_h = \frac{1}{_h}\sum$ $_h$ is the mean of the variable of interest in the post-stratum h.

The Post-Stratified Estimator of the total Y is given by

$$= \sum_{h=1} {}_h \, {}^{-}_h$$

If the true value of $_h$ is used this estimator is unbiased.

4.4. Time series Smoothing Techniques

Weighted Moving Average (WMA) could be used to reduce the gap in the time series. WMA assigns a heavier weighting to more current data points since they are more relevant than data points in the distant past. The new value use the WMA method could be compute at a local level and for any strata, and the national total could be obtained.

It could be combined with the cross-entropy method. It could be used to define the relation in (2), in fact could be the total based on WMA method for a given year.

Other technique such as**Exponential Smoothing** could be also used to adjust the trend.

4.5. Handling Misclassification

Abreu*et al.* (2011a) 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.* (2011b) 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. Model-Based and Model-Assisted Methods

We assume that the variable of interest Y (Production or Area) is related to a number of variables observed within the population throughout the years.

We have:

$$Y=f(X)$$

The idea is to assess f based on the census data and impute data for unobserved units during the survey.

The total production can also be written as

$$= 1 \quad + 1 \quad = 1 \quad + 1 \quad (\quad)$$

where 1 and 1 are vectors of n sampled units (sample size) and *N-n* non sampled units respectively. The population matrix of covariates is X = [,] where is the *n x p* matrix for sampled units and is the *(N-n) x p* matrix for non-sampled units.

The estimated weight is

$$= \frac{1 \quad + 1 \quad \hat{} (\quad)}{1} = \frac{_{=1} \quad + \quad \bar{}_{=1} \quad \hat{} (}{_{=1}}$$

The accuracy of the method lies in determining *f*.

a) Spline Regression Method

We are interested to the estimates, not necessary to the *f* itself, therefore the appropriate method is the semi-parametric regression since the OLS regression are influenced by extreme values.

This method uses the regression model = + , ~ (0,) , where *m* is the spline function using a linear combination of truncated polynomials.

(Henry K. & Valliant, 2012) shows that

$$=$$

with

$$= \quad ^{-1} - (\hat{\text{——}} - \hat{\text{——}})/$$

With this non-parametrical model, unit with the same characteristics X will have closed estimates of the variable of interested. The effect dues to misspecification in this case are reduced.

b) Generalized Regression Method (GREG)

This method is a calibration approach that involves minimizing a distance function between the base weights and final weights to obtain an optimal set of survey weights. Here optimal means that the final weights produce totals that match external population totals for the auxiliary variables X within a margin of error.

Specifying alternative calibration distance functions produces alternative estimators. A least squares distance function produces the general regression estimator (GREG)

$$_{GR} = \hat{} + \hat{} (- \hat{})$$

where $\hat{\hat{} = \sum$ is the vector of Horvitz-Thompson totals for the auxiliary variables, $\hat{} =$ is the corresponding vector of known totals, $\hat{}$ is the Horvitz-Thompson estimator used to estimate the total of the variable of interest during the surveys and $\hat{}$ is the regression coefficient estimated from census data.

The term $\hat{} (- \hat{})$ represents the estimate of the difference between the total value of the variable of interest and the HT estimates. This term could be positive, when the HT calculations underestimates the true total value and negative, when it overestimates.

5. Other Methods to Adjust data

Some methods could be performed to adjust data, as required. However, reconciliation of the survey data with the census data may still be necessary after these techniques are applied. These are to be implemented when the survey is being conducted. In fact, it is important take some actions in order to avoid the gap in the time series. Adjustment in data, could be done throughout the intercensal period.

a) dditional 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.

e) Update Sampling Frame

A good system to update the sampling frame should be settled. For instance, a part of the sampling frame could be update each year while implementing the survey. It will allow to minimize discrepancies related to the sampling frame.

6. The Experience of 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.

A) 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.

B) 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.

C) 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.

D) 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.

E) 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)

F) 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.

7. Concluding remarks

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 and trend adjustment may be a basis for data reconciliation. This technical paper has reviewed some of these methods. 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.

An updated sampling frame is one of the keys issue in order to avoid discrepancy in data. It is also necessary to have updated explainable variables. This can be a limitation of that method: The capacity to have accurate and update correlated variable. Some variables as the size of the holding, the number of holding in an area, the variable to the owning of equipment (tractor, etc.) are collected during the census and they can be updated using other data sources (population census, administrative file or other survey).

In some of the examples presented in this paper, 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 could be 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 are being tested to identify the most suitable ones for each type of problem, and to provide countries with effective and workable guidelines.

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., Sang, H., Lopiano, K.K. & Young, L.J. 2011a. 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. 2011b. 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., 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.

Auger, I. E. and Lawrence, C. E. (1989). Algorithms for the optimal identification of segmentneighborhoods. Bulletin of Mathematical Biology, 51(1):39–54.

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

Edwards, A. W. F. & Cavalli-Sforza, L. L. (1965). A method for cluster analysis. Biometrics,21:362–375.

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.

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.

Killick, R., Fearnhead, P., &Eckley, I. A. (2012). Optimal detection of changepoints with alinear computational cost. JASA, 107(500):1590 – 1598.

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

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.

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

Ouedraogo, E. &Nyamsi, U. (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.