Master Sampling Frames for Agricultural Surveys: Brazil overview[1]

Flavio Pinto Bolliger[2]
Marcos Paulo Soares de Freitas[3]
Giuseppe de Abreu Antonaci[3]
Maria Deolinda Borges Cabral[4]

1. Introduction

As the demographic and economic aspects of a society develop there is an ever increasing demand from the government and administrators for information. As consumers they always want more detailed and frequent information on a broader range of topics. As an official statistics agency we are always faced with resources constraints, either in the form of money or time.

The implementation of a master sample frame was a way to answer most of the demands while keeping the associated costs manageable. This is achieved through a more efficient use of resources, mostly on the field, where most surveys will share their sampled area.

The master sample frame also allows us to correlate information from different surveys at the area level, enriching the information produced by the national statistical system.

The focus of this paper is to present IBGE's recent experience in developing a master frame for household surveys, already in full use, and the procedures and difficulties to replicate this effort for agricultural surveys.

2. Development of the Master Sampling Frame for the household survey system

The first problem faced by the master sample was acceptance. Implementing a master sample meant changing the current surveys and the way the field work is coordinated. There were also additional changes to make the definition of concepts the same across all household surveys. All of this made the implementation of a master frame to be met with resistance in several areas.

The Brazilian master frame is composed of enumeration areas created by IBGE itself. They are well defined and cover the totality of the Brazilian territory, so there isn't a problem of completeness or duplication of units. However, there is a problem of classification. The enumeration areas are classified according to their occupation and some are not part of the target population of the surveys. To one of the groups that shouldn't be part of the survey was given a selection probability and a few areas were selected. This has been solved and the frame corrected the following year.

A difficulty faced by the master frame is the size of the enumeration area. Since it is planned for the sampled enumeration areas to remain in the sample for a few years, attending a few surveys, it was necessary for them to have a minimum number of households. However,

---

[1]Paper written for the High Level Stakeholders Meeting on the Global Strategy: From Plan to Action, Rome, 3- 5 December 2012 - Expert Session: Master Sampling Frames for Agricultural and Rural Statistics.

[2]COAGRO – Coordination of Agriculture - IBGE
[3] COMEQ – Coordination of Methods and Quality - IBGE
[4] COSEC – Coordination of Trade and Services -IBGE

mostly at rural zones, where population is sparser, some enumeration areas have fewer than 50 households. In this case a few areas were grouped into a single primary sampling unit.

The Brazilian Institute of Geography and Statistics is developing a new System of Integrated Household Surveys (SIPD, in Portuguese) based on a master sample. This process consists of many steps, such as developing and updating a master frame, designing a sample survey which can attend demands of most household surveys and modifying some current household surveys in order to adapt them to the new master sample.

The master frame is composed of a set of areas called *census enumeration areas*, which do not overlap and cover all the Brazilian territory. These enumeration areas have been allocated in different levels of stratification, adopting administrative and statistical stratification. The administrative stratification was planned to allow publication of results for those levels and the statistical stratification to improve precision. Enumeration areas were randomly selected from this master frame with probability of selection proportional to the size of the enumeration area, considering the number of households, thus composing the master sample.

The variable used in the statistical stratification was the total income of the head of the household for each enumeration area. Certainly there is no variable to be considered the best one for all surveys, so one variable highly correlated with most surveyed characteristics of a household is the best choice in this case. Another important characteristic is that the master sample is well spread over the territory, which had been demanded by the users of official statistics. In this way, national indicators now reflect also the behavior of the countryside, and not just capital cities and metropolitan areas, like before. Subsamples can be drawn from the master sample to attend the needs of each household survey. For example, the Continuous National Household Sample Survey uses a subsample of households in all enumeration areas of the master sample,whilst the Household Expenditure Survey 2008/2009 used a subsample of households in a subsample of enumeration areas. The subsamples selected were designed in order to have from none to total overlap between them. However, it was planned that all surveys share their enumeration areas (which are defined as the primary sample units) with the Continuous National Household Survey. In this way costs involving household listing to update the master sample are shared among the surveys. Additionally, information from different purpose surveys is combined at enumeration area level, improving our knowledge of the surveyed area.

3. Adoption of a Master Sampling Frame for Agricultural Surveys

The deployment project of the National System of Agricultural Establishment Sampling Surveys (SNPA[5], in Portuguese), currently in development by IBGE for the purpose of improvement of agricultural statistics in Brazil, involves the structuring of a master frame. In fact, the project covers many of the goals set out in the Global Strategy (GS). It also follows many of its methodological recommendations, in particular, the adoption of a master frame. Among the alternatives mentioned in the GS, the choice fell on multiple frame, combining agricultural census enumeration areas frame and register of farms from the agricultural census.

It's worth mentioning that this need (or the same methodological solution) for the production of agricultural statistics in the country had long been indicated. In this fundamental aspect of the statistical production process, what is being proposed today was summarized in the same terms in a project with equivalent purpose prepared by King and

[5] See SNPA project at http://www.ibge.gov.br/home/estatistica/indicadores/prpa/default.shtm.

Lindquist back in 1969.

> It is proposed that the IBGE develop a master sample for agriculture (p.32, emphasis added) [...]It was noted above that the agricultural statistics program will have multiple purposes [...]It is therefore recommended an area sample selection with multiple stages [...] using composite measures of farm size for selection of primary sampling units, with probability proportional to size. The entire enumeration will then be concentrated within these area units, except for the additional sample (p.28-29) [...]Before the area sample selection [...] a special list of large and important production units will be developed and defined as an independent universe.They will be sampled separately. (p.32) (King & Lindquist, 1969)

In 2011 and 2012, efforts were spent on studies, planning and defining the SNPA contents, especially the multi-purpose survey for the National Survey on Agricultural Activity (PNAG, in Portuguese), including pre-testing the collection instruments.The deployment schedule foresees pilot tests in 2013 and the beginning of the current surveys in 2014.

This session aims at reporting concepts, methodological choices, difficulties and challenges, and solutions in the structuring of a Master Sampling Frame for Agricultural Statistics in the scope of the SNPA project, with accumulated experience from SIPD and with the resources and information available today.

3.1Starting Information:2006 Agricultural Census and 2010 Population Census

The Master Frame for Agricultural Statistics is based mainly in the 2006 Agricultural Census information, although we currently have more recent information from the 2010 Population Census. Both censuses collected information for the composition of the National Directory of Addresses for Statistical Purposes (CNEFE, in Portuguese), in which the agricultural establishment is a unit, beside the household and others. The 2010 operation included the identification of agricultural establishments, however, for operational reasons, only the ones with buildings were registered.The Agricultural Census enumerated 5.2 million agricultural establishments. Of this total, 2.58 million declared value of goods from buildings, facilities and other improvements; 2.2 million declared area with buildings, improvements and paths, and 3.8 million declared living in the establishment. It is plausible to assume many of the small holdings have not declared the area and the value of the property.

The 2010 Census enumerated only 2.6 million agricultural establishments. It is assumed that this result shows the occurrence of important sub-enumeration. And it may indicate difficulty in ensuring sufficient attention to a secondary aspect in a research focused in households and demographics.

We are now in the middle of a data integration work in order to identify agricultural households that were listed by the Population Census but mistakenly had not had their addresses also associated to agricultural establishments. This work may lead to an improvement and upgrading of the list of establishments in enumeration areas but it will hardly be possible to consider them in the characterization of these areas.

In fact, the work of structuring the area frame considers only the aggregates derived from data from the 2006 Agricultural Census, despite the time spacing.

It is understood that the structural information of the agricultural census is more suitable and remains sufficient as references for sampling purposes throughout the intercensal period.

This decision also took into consideration the multitude of agriculture-specific information offered by the Agricultural Census, allowing not only different alternatives of

scaling variables in the sampling procedure as well as a generous database for analyzing sampling design suitability to multipurpose surveys, such as those provided for in the SNPA.

However, for operational reasons it was considered preferable to structure the operations based in the 2010 area map, already in use at IBGE for other survey systems.This guidance implied the need for appropriate the 2006 Agricultural Census information according to the 2010 area map in the formation of the starting information in order to build the area frame.

3.2Adaptation of the 2006 Agricultural Census into the 2010 area map

To carry out a census operation, it is prepared the operating base that will serve as the main support for all field activities.This preparation is planned based on the previous census geographic base, which is updated to ensure coverage of the entire national territory and to reflect the latest administrative division of the country.

In 2010, the starting point was the 2006 Agricultural Census, and the upgrade process had as its basic principle the maintenance of the enumeration areas, and when it was necessary to change them, the first choice was to divide them.Only in a few exceptional cases there was area aggregation or rearrangement.

The construction of the area frame that will serve the SNPA will be based on the 2010 geographic operational base, and as the agricultural information available is associated to the 2006 areas, it is necessary to establish the relationship between the areas of the two bases, so the information can be associated to the areas that make up the area frame.The 2006 base had 86,350 enumeration areas with agricultural activity. The 2010 uptaded area framework counted up to 105,221.

The information considered refers to the number of establishments and the totals for some variables related to agricultural production.

When the relationship between the bases is 1 to 1, i.e. the 2010 and the 2006 enumeration areas are the same, there is no problem whatsoever. It is a simple association of the information directly to the new area. On the other hand, when there was a division, it needs to be decided how to make this association to the new 2010 areas.

Three alternatives were thought. First, to find out the number of agricultural establishments in each enumeration area -- data from the 2010 Census -- to enable the distribution of the variables of interest totals proportionately among them. But as already mentioned, this 2010 information has some restrictions, which prevented the adoption of this alternative.

Second, to distribute the variables totals equally among the new enumeration areas. And third, still under feasibility study, to allocate the agricultural establishments in the new enumeration areas, using the geographical coordinates collected in the 2006 Census, and only then generate the totals of the variables of interest for each area.

Having in mind the present stage of the SNPA project -- initial sample planning, especially in the definition of an initial sample size of enumeration areas --, it is assumed that the second alternative, which is simpler and easily implemented, will yield good results, which may not be very different from the ones derived from the third option. Even so, it is expected that the third alternative may be adopted for the final construction of the area frame.

This is a recurring problem. Next year, 2013, IBGE starts a new big update of the operational base to be used for the Population Count (2015) and the 2015 Agricultural Census (in 2016). This update will mean new adjustments in the definition of enumeration areas, with

additional subdivisions.

Also considering the objectives of longitudinal analysis in master frame based surveys, the transitional procedures between area frames over time represent a methodological problem not yet sufficiently studied and solved.

3.3Structuring the list of agricultural establishments and producers, and update experience

For sampling studies purposes, the records with data from questionnaire considered valid are being used in the examination of the Agricultural Census. In order to establish a register of establishments and producers for future operations, in contrast, the whole of the records obtained in the census operation was analyzed, with special attention to the review of the registration variables, and updated according to information from the Central Register of Enterprises (CEMPRE, in Portuguese), maintained by IBGE, and the Annual Social Information Report (RAIS, in Portuguese), an administrative record with information regularly provided by employers to the Ministry of Labor and Welfare.

The first version of the Agricultural List Register, assembled from deterministic and probabilistic relations of the 2006 Agricultural Census data (5,218,527 records) and the 2007 CNEFE archive (5,215,380 records), consists of 5,041,613 agricultural establishments with registration information (name of the producer, and name and address of the establishment) and a set of economic variables that allow for selection of future samples.

Of the latter total, 52,319 (1.04%) presented information on the National Register of Legal Entities (CNPJ, in Portuguese), but only 33,159 of them were valid values for comparison.

The study on upgrading procedures for this register used, as the first experience, the Central Register of Enterprises (CEMPRE-2010). It was then created the RAIS-AGRO Register with the local units classified in farming activities, according to CNAE 2.0[6]. This register is composed of 334,863 enterprises and 438,648 local units (ULs), providing variables relating to the registered name and the establishment's name, address and classification of activity.

The match of these two files considered the following variables: unit of the Federation (UF), city, producer`s and agricultural establishment`s names, local unit`s registered and commercial names, as well as the addresses of both files.

The following procedure was used:

1. Standardization of the variables of both files for comparison;
2. Deterministic match considering the CNPJ -- 19,040 records could relate;
3. Probabilistic match by UF and municipality, comparing the producer`s and the establishment`s names with the formal and the commercial names and the addresses of both files.

As a result of this process, it was possible to identify a total of 50,871 agricultural establishments, representing 11.6% of the total RAIS_AGRO Register`s ULs.

[6]National Classification of Economic Activities

The main reasons for this low percentage were the following:

- The CNPJ variable is not frequent in the Agricultural List Register, as agricultural establishments are not required to hold a CNPJ, and furthermore, some of the values actually referred to theTaxpayer Identification Number (CPF) and other unidentifiable numbers;
- The NAME comparison keys were not identical variables, the official and the commercial names do not necessarily coincide with the producer`s or the establishment`s names; and finally,
- Address was the item with fewer results. One possible explanation would be that, whereas rural addresses are not easy to identify, the formal record of CEMPRE has the office`s, the owner of the company, or the accountant`s addresses in the city.

Although advanced linkage techniques have been judiciously applied, the results achieved were quite frustrating. In fact, it is not possible to identify a large number of units present in both registers. This has inhibited the implementation of procedures aimed at the incorporation of new units, as it would result in duplication of non-commensurable records.

This may be due to problems of compatibility and quality of information found in different registers. In fact, experience has shown the need for major investment in the standardization of the information required, the entry system, and, in particular, the entry pattern between the administrative and survey records.

3.4Definition of list register and its impact in the area register

The SNPA planning provides for the division of the population of interest into two parts: a group of agricultural establishments will be part of a list frame and will be sampled from it, and the other group will be considered in the formation of the area frame.

We are discussing the criteria that will determine to which group each establishment will belong. In early studies, these were the criteria for the definition of subpopulation to be sampled at random from the list: (a) property owned by a formally constituted enterprise, i.e., registered at the National Register of Legal Entities (CNPJ), (b) 3 or more permanent employees, (c) 200 ha or more in exploitation.

The list register would consist of establishments that present at least one of these features. The data in the table below, based on the 2006 Agricultural Census, give an idea of the size of sub-population addressed to list frame (almost 300,000) and to area frame (almost 5 million) according to these criteria.

Table1 - Agricultural establishments according to sub-populations of sampling study by list and area - Brazil, 2006

Sub-populations according to type of sampling	farms
List frame	286,713
Only CNPJ	37,247
Only 3 or more permanent employees	55,436
Only 200 ha or more in farm area	140,128
With 3 or more permanent employees and 200 ha or more in farm area	37,498
With CNPJ and 3 or more permanent employees	5,765
With CNPJ and 200 ha or more in farm area	5,596
With CNPJ, 3 or more permanent employees and 200 ha or more in farm area	5,043
Area frame (does not meet any of the 3 conditions)	4,888,923
Total	5,175,636

Source:DPE/COAGRO; 2006 Agricultural Census

It was expected larger establishments and enterprises would be selected, which have more accessible sources of cadastral update. However, the already mentioned difficulties encountered in record linkage and cadastral maintenance recommend to restrict the amount of cases to the list frame. In a second approach, we decided to set specific criteria according to the farm type more relevant and beyond generally applicable criteria, as follows.

Table2 - Farms according to selection criteria for composing sub-population for list sampling

Type	Cut off	Farms	Item coverage (%)
Eggs	10,000 laying hens	1,310	80.0
Chicken	420,000 chickens sold	426	20.6
Pigs	40 sources or 880 pigs sold	10,090	34.3 & 76.7
Milk	50 dairy cows	22,459	14.2
Cattle	1000 heads	20,365	28.3
Coffee	1,000,000 coffee bushes	214	8.6
Orange	300,000 trees or 200 ha of harvested area	425	15.3 & 38.9
Cotton	300 ha of harvested area	313	84.4
Tobacco	100 ha of harvested area	376	12.9
Rice	100 ha of harvested area	3,593	42.5
Corn	300 ha of harvested area	1,633	10.7
Wheat	300 ha of harvested area	64	2.6
Soybean	300 ha of harvested area	10,245	57.6
Beans	100 ha of harvested area	3,147	16.2
Cassava	25 ha of harvested area	8,861	43.8
Wood	1000 ha of forest area planted	213	31.4
Sugar cane	500 ha of area harvested or CNPJ (corporation)	5,329	71.2
	Or		
	Gross production exceeding BRL 150,000 (~ USD 75,000)	25,117	
	Or		
	Storage capacity over 1,200,000 tonnes	2,636	
Total		**115,053**	

Source:DPE/COAGRO; 2006 Agricultural Census

We are evaluating the appropriateness of these criteria, as well as considering an additional criterion of inclusion in the list sampling: establishments with semi-integration contracts for specialized activities, such as tobacco, pigs, poultry and others.

In order to have the definitive areas frame, after all the procedure of relating the geographic bases described in the previous item, the establishments that will be part of lists registration will need to be excluded. Only after this exclusion, the aggregated data necessary for planning the sample and which are related to the agricultural information from the 2006 Agricultural Census will be produced for each enumeration area.

3.5Enumeration areas with few cases

The census maping is built for multiple purposes, being the aspects related to Population Census the most relevant to the delimitation of the area, especially those relating to the number of households.There is, therefore, many enumeration areas with few agricultural establishments, which can be seen in the table that follows.

Table 3 - Frequency of Enumeration Areas (EA) by and size-Brazil-2006

EA size	EA
1to 10 farms	28,374
11 to 50 farms	19,756
51 to 100 farms	18,610
101 to 200 farms	16,075
201 and more farms	3,535
Total	**86,350**

Source: 2006AgriculturalCensus

To the SNPA -- which provides for the permanence of the enumeration areas in the sample for a certain period of time, an area sample rotation and a rotation of selected establishments within each area -- it would not be appropriate to continue with the small areas as they are shaped in the map.We are studying, therefore, a way to aggregate them.

The definition of the methodology to aggregate them takes into account the algorithms that were used in grouping the areas for SIPD (see section 2) and for dissemination of the Agricultural Census results.

The criteria for aggregation in SIPD involved the vicinity and the minimum number of private households, without considering the homogeneity of the demographic characteristics of the enumeration areas. For the Agricultural Census, in addition to the vicinity and of size criteria, the latter measured in number of establishments, it was used a criterion of homogeneity, which took into account the intraclass correlation coefficient of various features investigated in the census.

We intend to adopt a methodology similar to that used in the Agricultural Census, only reducing the amount of features in the definition of homogeneity of the enumeration areas within the groups.We also plan to use the same reasoning applied to the aggregation for the SIPD with regard to the minimum size of each group, which means to consider the number of

establishments selected for the sample within each enumeration area (or group of them) and the mean permanence time in the sample.

In this way, the groups will be formed by enumeration areas that will keep a certain similarity with respect to agricultural characteristics, and it is believed that the number of establishments re-entering the sample, due to the rotation, will be reduced. However, this can lead to the definition of larger area units than desirable.

3.6 Interaction between master samples frames

Another issue that arises in planning the SNPA is its relation with the newly deployed Integrated System of Household Surveys, which will make the integration of the sample survey in the social area. As both systems will have the same register base, the geographical operational base that delineates the enumeration areas, we have been discussing how to handle the overlapping areas (enumeration area or groups of enumeration areas) on selected samples for surveys from the two systems.

The coincidence of enumeration areas being allowed, updating the list of addresses and units becomes easier, as the areas are already known by the field team.Whereas if no coincidence is allowed, a number of enumeration areas would be added -- selected for the SNPA -- and be investigated in order to update the master sample.

A disadvantage of having sample areas overlap in both systems is a possible overload on the informant. Establishments also classified as households can be selected for both the social and the agricultural survey. And as the number of surveys may not be small over time, it would be difficult to avoid the coincidence of these units in the various samples.This could lead to greater difficulty in obtaining the information, with the increasing refusal of informants.

4. Final remarks

IBGE's experience in the structuring of the master sample to conduct integrated household survey, already in use, has been useful to the methodology's application for deployment of the new National System of Agricultural Establishment Sampling Surveys, in development.

The 2006 AgriculturalCensus forms the fundamental basis for structuring the master frame for agricultural surveys.

We highlight the following issues in its preparation:

In the case of Brazil, the Population Census does not seem to result in sufficient or adequate information to an agricultural master frame;

For operational purposes we chose to adopt the most up-to-date enumeration area map, implying suitability procedures. The transitional procedures between area frames over time represent a methodological problem not yet sufficiently studied and solved.

Despite the use of advanced techniques of record linkage, the update of list frame based on administrative records was not successful due to compatibility issues and quality of the information found in different registers. Investment is required in information standardization and quality of filling out information in surveys and administrative records.

It was decided to limit the size of the list frame. The selection criteria for the list need to be carefully examined in the optimization of the multi frame solution.

The geographical distribution of establishments involves the aggregation of enumeration areas with few cases in the area frame to ensure better adjustment to the rotation process. However, this can lead to the definition of larger area units than desirable.

The agricultural master frame has the same register base as that for household surveys. Control of overlapping areas with pros and cons between cadastral update easiness and burden to the informant.The interaction between master frames is considered a matter that requires further studies and knowledge of international experience.

5. Bibliography

CENSO agropecuário 2006. Brasil, grandes regiões e unidades da federação. Rio de Janeiro : IBGE, 2009.

FREITAS, M.P.S. Amostra Mestra para o Sistema Integrado de Pesquisas Domiciliares. Rio de Janeiro: IBGE, Diretoria de Pesquisas. 2007. (Texto para Discussão nº23).

KING, A. J. ; LINDQUIST, M. V. Programa de estatística agrícola para Fundação IBGE. [S.l.] : Agência Norte-Americana para o Desenvolvimento Internacional, 1969. 108 f.

PROPOSTA de Sistema Nacional de Pesquisas por Amostragem de Estabelecimentos Agropecuários – SNPA: concepção geral e conteúdo temático (Texto para discussão – 2ª versão – Outubro 2011). Coordenação de Agropecuária, Diretoria de pesquisas, IBGE. 2011. Disponível em: http://www.ibge.gov.br/home/estatistica/indicadores/prpa/SNPA_concepcao_e_conteudo2av.pdf. Acesso em: novembro 2012.

SISTEMA Integrado de Pesquisas Domiciliares - SIPD. Rio de Janeiro: IBGE, Diretoria de Pesquisas. 2007. (Texto para Discussão. nº24).

WORLD BANK. Global StrategyTo Improve AgriculturalAnd Rural Statistics. ReportNumber 56719-GLB. Washington, DC: World Bank, September 2010. 55p. Disponível em:http://www.icas-v.org/AgStat_GlobalStrategy_10.pdf. Acesso em: novembro 2010.