Types of data integration
Linking Units, Some Scenarios and Outcomes

Fourth RAP Regional Workshop on Building Training Resources for Improving Agricultural and Rural Statistics: Survey Methods for Agricultural Statistics- Current Practices and International Recommendations

14-18 December 2014, Tehran, Iran.

Alick Nyasulu
Statistical Institute for Asia and the Pacific (SIAP)

Content

- Introduction
- Context of Integration
- Cases of Integration

Introduction

- Integration is generally based on a procedure that merges information originating from multiple surveys or archives.

Increased information due to:

– units of analysis

– variables

– temporal occasion

Contexts of integration

- **Objects of Integration**

❖ **Statistical Data**

collected through total or sample surveys, with the adoption of statistical standards

NSO main source

- **Objects of Integration**

❖ **Administrative Data**

collected through archives or registries created for administrative purposes, or in compliance with laws or regulations

Non-NSO sources

Contexts of integration

National Statistical System

- Institutional mandate
- Data production capabilities
- Obligations to adopt/adhere to international and national standards

Administrative sources

- Dual purpose of registers
- Planned in consistent manner, definitions
- Ex ante/ex-post integration of data collection
- Integration at same or different levels of collection

Case 1: Multipurpose Survey

Integration through a single multipurpose survey!

- Single subject holder
- Detect plurality of information from previous different surveys designed for different purposes/subjects
- Collection of known parameters/relationships about known units-*ex ante*

Case 1: Multipurpose Survey

Integration through a single multipurpose survey!

Positives

- Ability to study the relations between different phenomena, previously investigated through different surveys on different units ;
- Reduction of the overall sample size and consequent reduction of the overall cost and statistical burden;
- Additional resources freed and can be employed to improve the quality of the survey, in terms of coverage, accuracy, or timeliness.

Case 1: Multipurpose Survey

Integration through a single multipurpose survey!

Drawbacks

- Non-optimal timing for the detection of various phenomena may lead to bias in the estimates, especially as regards changes of phenomena over time.
- Sample fatigue and greater statistical burden for individual respondents. Possible adverse consequences in terms of accuracy (measurement errors, total and partial non-response.
- Simplification of the questionnaire and subsequent loss of information.
- Difficulties with the interview protocol (change of respondents) and consequent possible non-sampling errors (non-response and measurement errors).

Case 2: Different Surveys, Archives

Ex post integration of data from different surveys or archives

- Conceptually simple, if the different datasets present the same type of enumeration units (individuals, households, business, etc);

- Technically simple, if the units are identified by the same unique identifier (UID) or by a combination of variables uniquely defined and available in the different datasets (key variables or linking variables);

- Operationally feasible, if the linkage is in compliance with the policies governing the dissemination of the results of the various surveys, and the owners have a common goal.

Case 2: Different Surveys, Archives

- *Ex post integration of data from different surveys or archives*

Quality issues

- Quality of linking dependent on information of individual data sets;
- Different surveys with same enumeration units
- Different enumeration units for different surveys/archives
  - Macro-level integration, domains uniquely defined by different EAs, hold relevant information for estimation
  - EAs based on logical units i.e individuals from same family, land parcels from same area

Case 2: Different Surveys, Archives

- *Ex post integration of data from different surveys or archives*

Positives

- Greater consistency of the direct estimates of variables from different databases
- Increased efficiency of the estimates of the variables present in the various integrated databases

Case 2: Different Surveys, Archives

- *Ex post integration of data from different surveys or archives*

Drawback

Effects of mismatches on estimates.

Positives

- Ability to estimate parameters of the relationship between phenomena not jointly collected in any of the integrated surveys.
- Availability of new or richer sample frames, from which more efficient sampling designs can be defined.

Case 2: Different Surveys, Archives

Positives

Ex post integration of data from different surveys or archives

- Ability to estimate parameters of the relationship between phenomena not jointly collected in any of the integrated surveys.
- Availability of new or richer sample frames, from which more efficient sampling designs can be defined.

Case 3: Planning integration

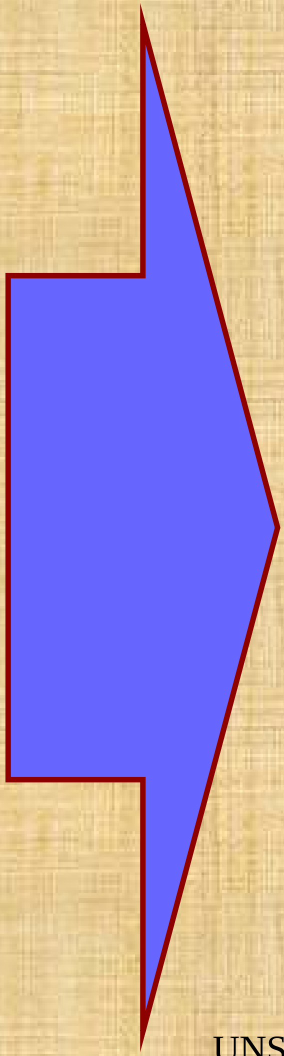
Planning data integration on the basis of different surveys or archives

- Integration is a process designed ex ante;

- Each survey or archive is designed bearing in mind common goal of integration;

Case 3: Planning integration

Planning data integration on the basis of different surveys or archives



- Designed with proper consideration of existing database that may be linked;

- Each survey maintains its own ownership and autonomy in response to specific needs.

Case 3: Planning integration

Planning data integration on the basis of different surveys or archives

Positives

- Each survey maintains its autonomy in response to specific needs;
- Cost reduction,
- Optimization of the use of the overall technical, organizational and financial resources available;

Drawbacks

- ➢ **Need for a minimum time gap between the various data collection exercises;**

Case 3: Planning integration

Positives

- Good quality of record linkage, due to the special attention given to the adoption of common definitions, UIDs and to the data collection for the linking variables planned.
- Achievement of the planned level of quality of the estimates, establishing a coherent sample size, a sample selection technique and an estimation method of the target parameters

Drawbacks

➢ **Loss of specificity and flexibility due to the need to link the units.**

Conclusion

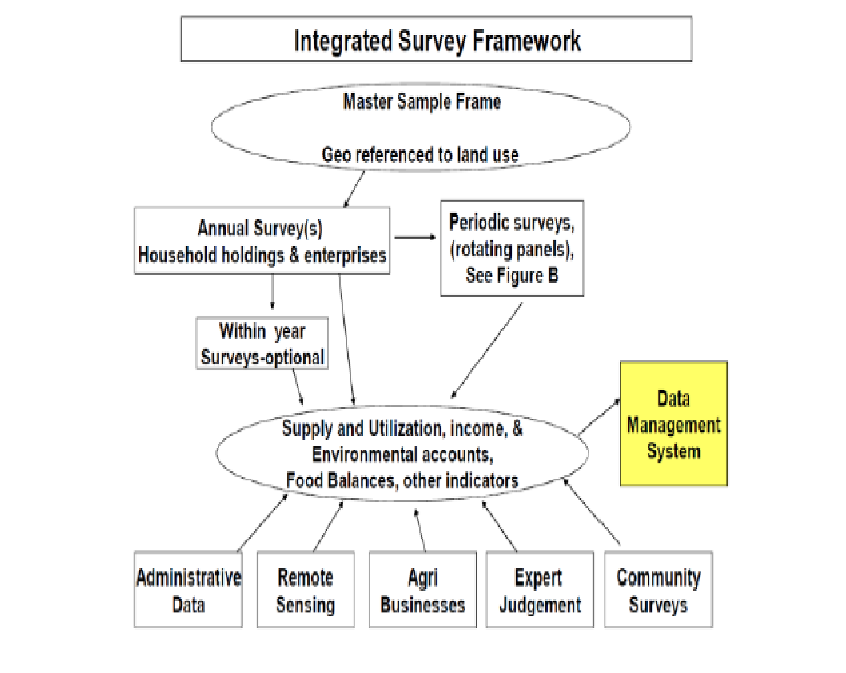

Integrated Survey Framework
Master Sample Frame
Geo referenced to land use
Annual Survey(s)
Household holdings & enterprises
Periodic surveys, (rotating panels), See Figure B
Within year Surveys-optional
Supply and Utilization, income, & Environmental accounts, Food Balances, other indicators
Data Management System
Administrative Data
Remote Sensing
Agri Businesses
Expert Judgement
Community Surveys

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

- Integrated data system for agricultural statistics considered in the Global Strategy is example of ex ante integration
- Different data sources i.e surveys, adminstrative records, expert evaluations, remote sensing

Thank You