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
- Simple Integration
- Integration Scenarios

Introduction

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

  Increased information due to:

  – units of analysis

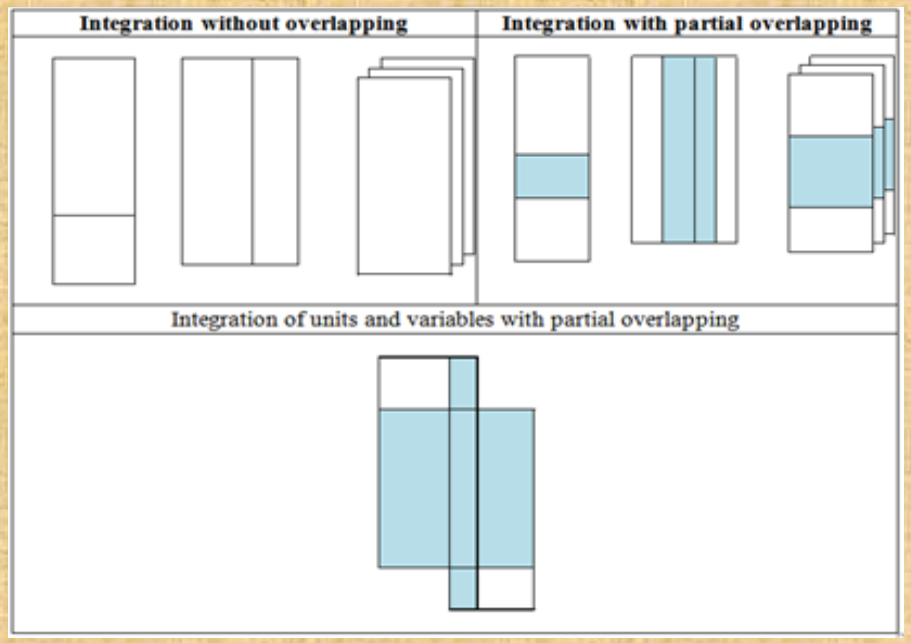
  – variables

  – temporal occasion

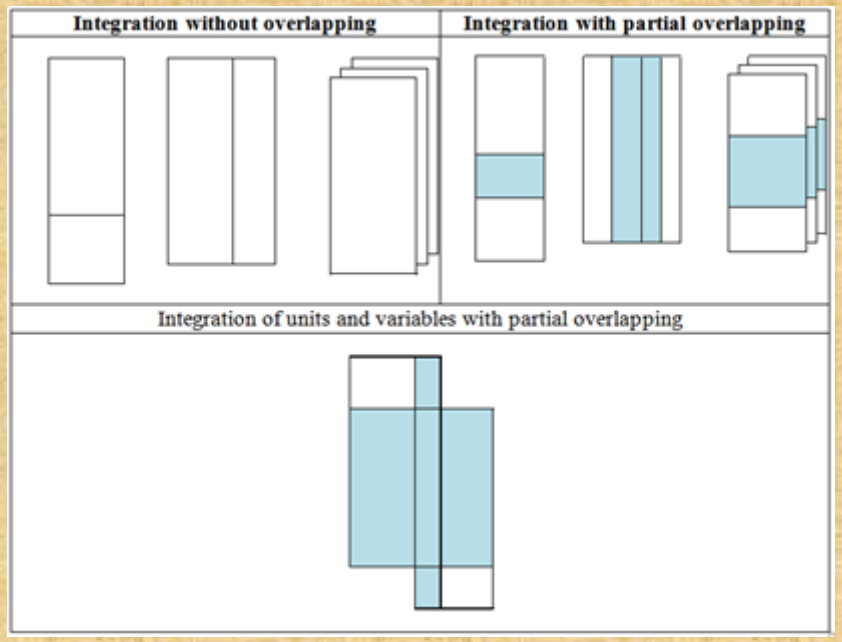
Population

- Information collected at the same level of aggregation in different surveys over time
  - information to be integrated may or may not overlap
  - Units and variables of independent variables on same enumeration units
  - Repeated units

Simple Integration

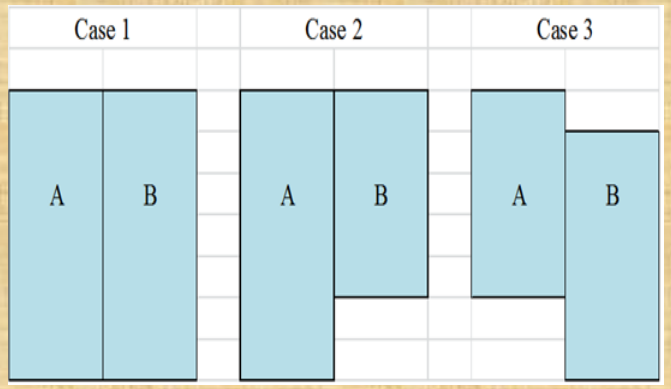


Simple Integration



- Simple integration of only one of the three elements i.e units, enumeration areas, variables

Integration scenarios



3 Cases

1. **Every unit in dataset A is also in dataset B and vice versa.**
2. **Every unit in dataset B is in dataset A, but there are individuals that appear only in dataset A.**
3. **Some units appear in both datasets A and B, and other units appear in only one of the two datasets.**

Integration scenarios

Imperfect information from the 3 cases

1. Duplicated units
2. Omitted units
3. Missing data
4. Errors in data
5. Timing differences

Fixing imperfections

1. Careful preparation of input files before integration
2. Integration procedure to be dependent on data reliability

Integration scenarios

1. **Intersection of the two databases**

Yields different outputs

i. Case 1 combined units,

ii. Case 2 intersection depends on ratio of units in A and units in B,

iii. Case 3 intersection dependent on number of units belonging to only one of the two data sets

Integration scenarios

2.Union of the two databases

Case 1

- match the intersection of the two databases

Cases 2&3

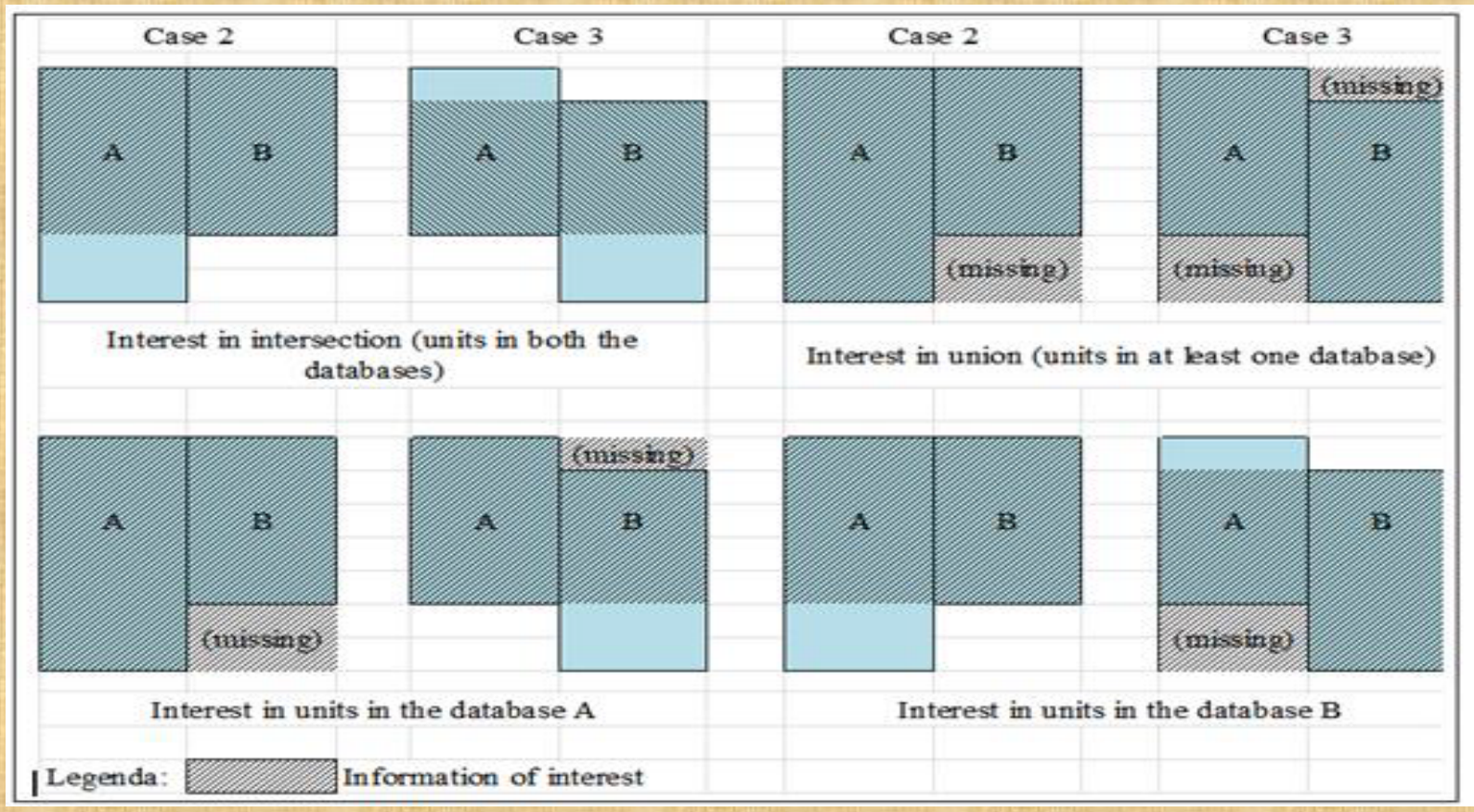
- take into consideration many more units than the intersection;
- some information is missing.

Integration scenarios

3. **Only database A, or only database B:**

1. Additional information arising from the integration of the two databases is relevant
2. **Possible to obtain information on the units that represent the intersection**
3. **But data for the other units of interest are missing**

Integration scenarios



Information from different Units

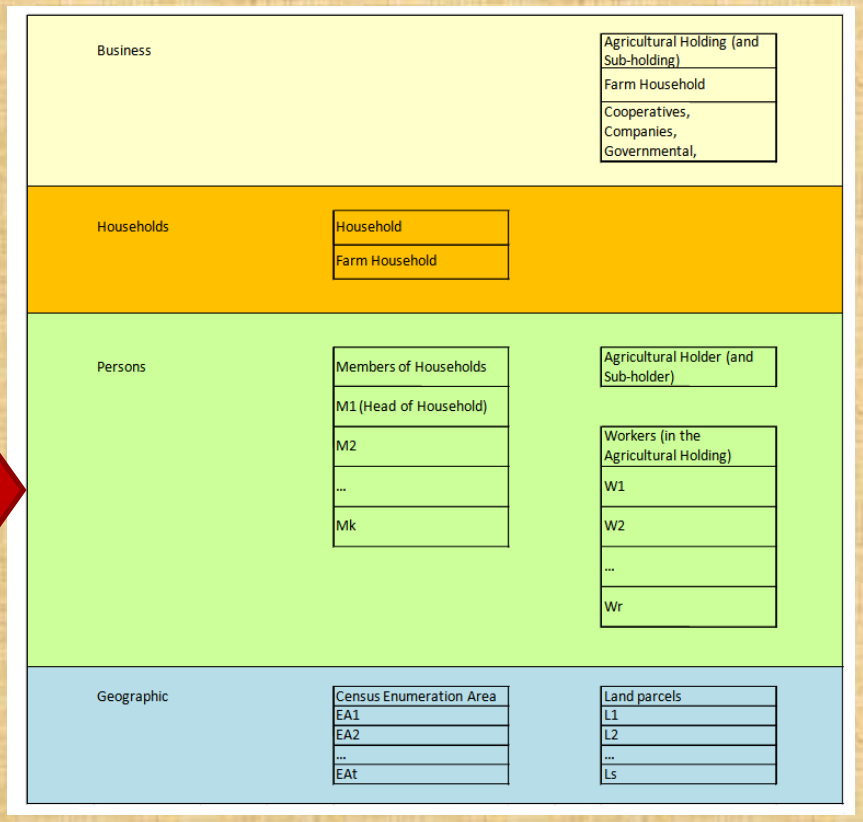
- Statistical units that are not necessarily of the same type or level of aggregation.
- Data collected from different surveys

- Individuals
  - Members of household, workers on a farm holdings, Farmholders etc
- Households/ Farm Households
- Land parcels, administrative/geographical units
- Businesses

Some scenarios

- Income surveys from different enumeration units
  - Agricultural holdings
  - Individuals
  - Farm Households

Integr ate Units



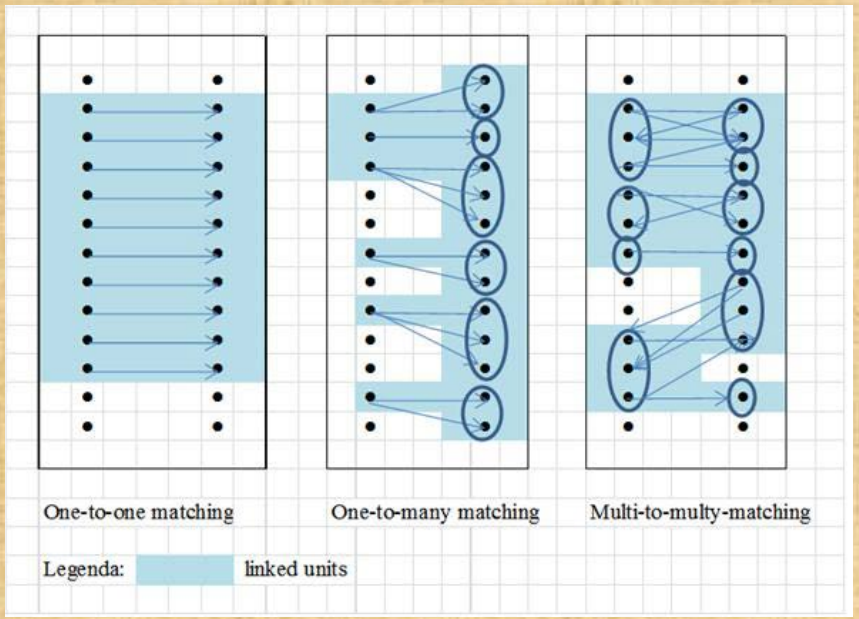
Data Linkage Process

- Key considerations

1) One to one Matching
2) Many-to-one Matching
3) Many-to-many Matching

Data Linkage Process

- One to One
  - One record of one database links to only one record of another database
- Many-to-one
  - One record from database linked to many records in another database
- Many-to-many
  - similar to many-to-one
  - possible for records on both databases to be linked but rare



Data Linkage Process

Population Censuses

- Primary statistical unit is household
- Linkage of these units forms the basis integrated agricultural statistical system

Agricultural Censuses/Surveys

- Primary statistical unit is agricultural household
- Sometimes these households are not the same

Other integration requirements

- Integration of metadata
  - Concepts, classifications, data collection methods etc
  - Preconditions for an integrated database
- Efficient organisation, trained personel and sustainable budgetary allocations
- Cooperation amongst agencies, producers and archiving

Thank You