UNITED NATIONS ESCAP SIAP Statistical Institute for Asia and the Pacific

Session6B:
From administrative data to register-based statistics;
Data Linkage and Matching

Regional Training on Producing Register-based Population Statistics in Developing Countries
27– 31 October 2013

Arman Bidarbakht-Nia

Statistician/lecturer, UNSIAP

Sources:

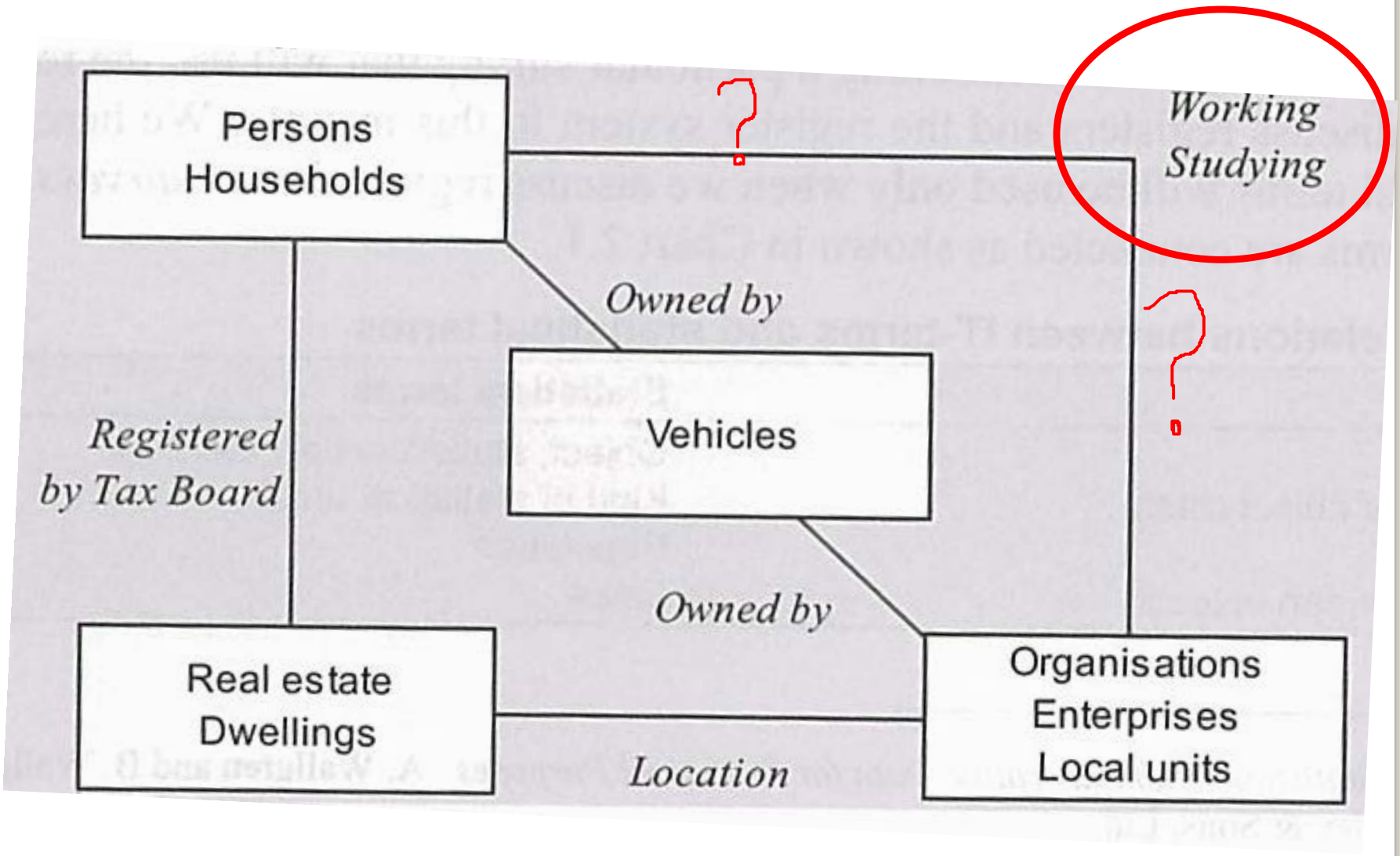
-United Nations (2011): Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices
-Wallgren, A. and Wallgren, B. (2007), *Register-based Statistics: Administrative Data for Statistical Purposes*, John Wiley & Sons, Ltd, Chichester, UK.

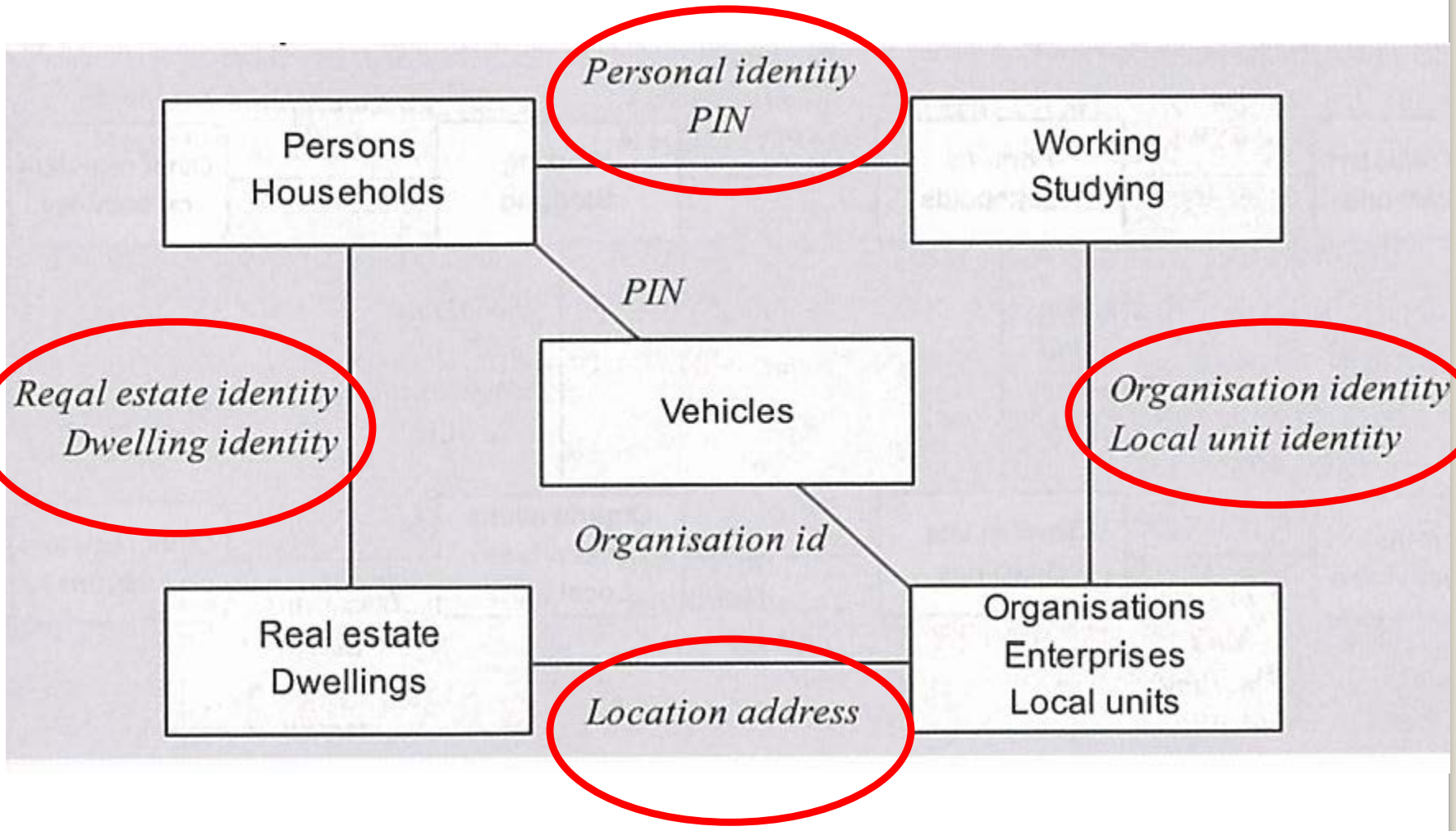
Need for Linkage between multiple sources

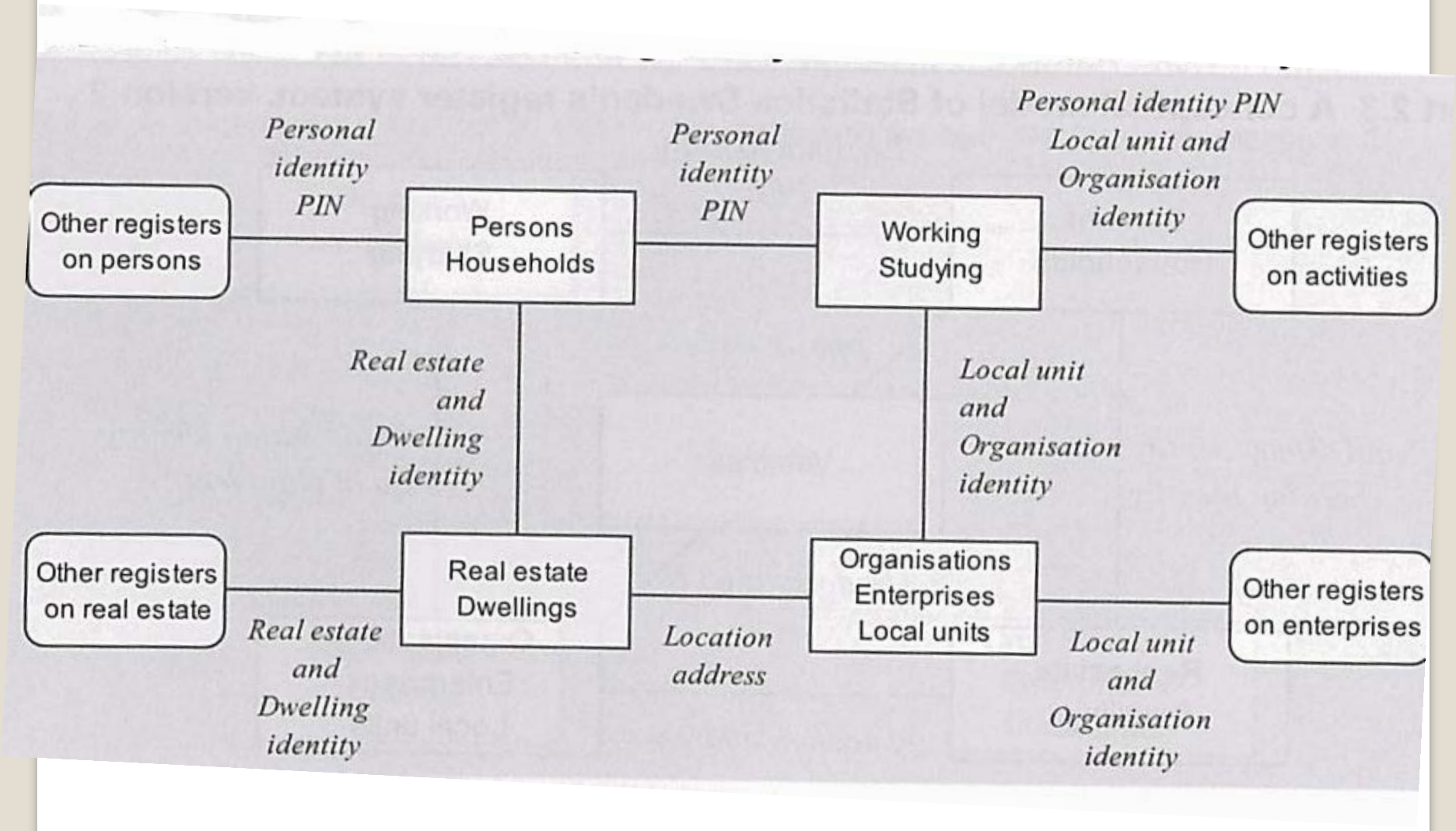
- Creating a statistical register
- Using auxiliary variables from administrative data
- Using multiple administrative sources as frame
- Administrative sources complement sample surveys
- …

Links are created based on relations

- Between persons, enterprise and local area
  - *person works/studies at an enterprise/ organization located in a local area*
- Between person and property/dwelling
- Between local area and property
- Between person/enterprise and vehicle


Persons
Households
Working
Studying
Owned by
Registered
by Tax Board
Vehicles
Owned by
Real estate
Dwellings
Organisations
Enterprises
Local units
Location


Personal identity
PIN
Persons
Households
Working
Studying
PIN
Reqal estate identity
Dwelling identity
Vehicles
Organisation identity
Local unit identity
Organisation id
Real estate
Dwellings
Organisations
Enterprises
Local units
Location address


Other registers on persons
Personal identity PIN
Persons Households
Personal identity PIN
Working Studying
Personal identity PIN Local unit and Organisation identity
Other registers on activities
Real estate and Dwelling identity
Local unit and Organisation identity
Other registers on real estate
Real estate and Dwelling identity
Real estate Dwellings
Location address
Organisations Enterprises Local units
Local unit and Organisation identity
Other registers on enterprises

How to do it?

- A link between two units consists of "*one or several*" common *linkage variables* that contain the information needed to *identify relations* between different types of units.

- This typically takes a form of *matching*

  - Exact
  - Probabilistic

How to do it? Exact matching

- When a *common identifier* exist
- Heavily depends of the quality of matching variables
- Even if common identifiers exist, it may not be possible to rely only on exact matching because of errors

How to do it? Probabilistic matching_1

- When *common identifiers* don't exist/have poor quality
- Using variables common to sources involved; *Matching keys*
- Matching keys may not present in all sources. They may be derived (*size of the enterprise by ISIC code from turnover per head*)
- Most common variables: name, address, DOB, occupation

How to do it? Probabilistic matching_2

Choice of variables

- Distinguishing power: *uniqueness of the values of the matching key*
  - High distinguishing power: *reference number, full name, full address*
  - Low distinguishing power: *sex, age, city, nationality*
- Distinguishing power also depends on the level of details. So careful choice and standardization

How to do it? Probabilistic matching_3

Possible scenarios

1- Match: A=A

2- Non-Match: A≠B

3- Possible match: A=a ?

4- False match: A=B

5- False Non-match: A ≠A

Basic matching techniques

Clerical

- Expensive
- Inconsistent
- Slow
- But; Intelligent

Automatic

- Cheap
- Consistent
- Quick
- But; of limited intelligence

Best is a mixed method with minimizing clerical

How to maximize automatic matching?

1. **Standardization**
   - Spell out abbreviations
   - Standardize common variations of names *(of places, persons,..)*
   - Remove “noise” words *(street, road,..)*
   - Standardize postal code

It may reduce quality of data or even chance of matching. Best is to keep the original variables

How to maximize automatic matching?

2. **Parsing;** *an extension to standardization*

- *Increase chance of matching by converting text from a form that is recognisable by humans, to a form that is more logical for computer processing; e. g.*
- Letters with similar sounds to a common string *("f", "v", and "ph" to "f")*
  - Remove silent letters *("h" from "Johnson")*
  - Vowels to single character *("ee" in "steel")*
  - Remove vowels from the end *("ea" from "Andrea")*
  - Replace double letters to single *("nn" in "Anna")*

How to maximize automatic matching?

3. **Blocking;** *For large files*

- *Break the file down into smaller "blocks" to save processing time;* *e.g. do the matching only in the corresponding town rather than whole country*

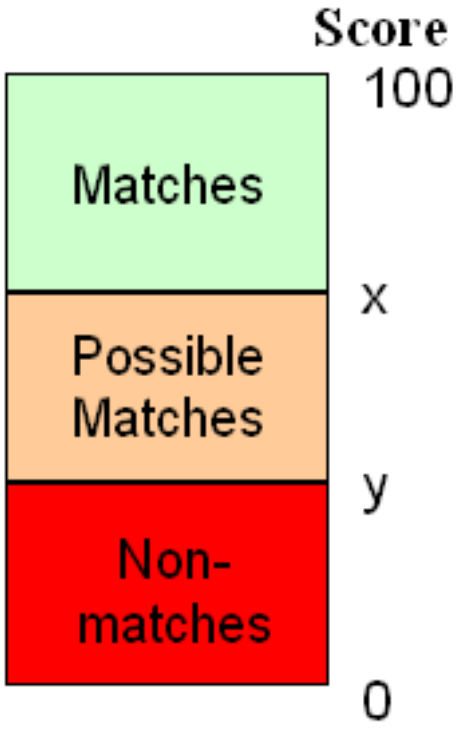
- *can improve the cost-efficiency of the matching process*

- *Can be applied in several stages with less restriction in later stages for non-matched records*

How to maximize automatic matching?

4. **Scoring;** *the likelihood for matching*

- *whether a pair of records is considered to be a definite match, a possible match or a non-match.*



The x and y values are determined based on a probabilistic model

How successful is the combined matching?

The best example is internet search engines

Google Custom Search

siap workshop registr census

Search Advanced Search

About 320 results (0.21 seconds)

Statistical Institute for Asia and the Pacific (**SIAP**)

Regional **Training** on Producing **Register**-based Population Statistics in **...** International University (TIU) under the **SIAP**-TIU Agreement Academic Year 2014.

www.un**siap**.or.jp/

Statistical Newsletter - Third Quarter, 2008, ESCAP Statistics Division<

Oct 3, 2008 **...** **SIAP** meetings/**workshops**: **.....** Mr Davender Kumar Sikri, **Registrar** General and **Census** Commissioner of India, is the Head of the **Census ...**

www.unescap.org/stat/nl/nl-Q3-2008.asp

2012

SIAP Training activities and News in 2012 / 2013 **...** Regional **Training Workshop** on Use of Population and Housing **Census** Data for Sub-national **...** collaboration among statistical agencies, ministries of health and civil **registration** authorities.

www.un**siap**.or.jp/news/enewsletter_2012.php