Designing a Microdata File for Public Use

Lara Cleveland, Research Scientist
7 September 2023

UNESCWA Workshop on Population Projections and Preparing Microdata
Rabat, Morocco
4-8 September 2023

UNIVERSITY OF MINNESOTA IPUMS

1

Desired Characteristics of MICRODATA for Public Use

IPUMS

Useful – Scientific – Principled

- Flexible = neutral design
- Representative of the population
- Well documented
- Sufficiently detailed
- Intact household structure
- Protect privacy of individuals
- Accessible

2

Useful – Scientific – Principled

Metadata
Sample design
Geography and GIS
Data quality
Confidentialization

- Flexible / broadly usable
- Representative of the population
- Well documented
- Sufficiently detailed
- Intact household structure
- Protect privacy of individuals
- Accessible


People illustrations by Storyset


IPUMS

3

Sample characteristics

Keep it simple for flexibility

- **High-density** (10% is standard) data file.
- **Household-level** to include all persons from each sampled household
- **Systematic** design
  (implicitly geographically stratified)
  every nth record across geographically sorted records
- Inflation factors (or **weights** and design variables for complex samples)
- S**ample design documentation**

IPUMS

4

Metadata

Data are only usable or meaningful with good metadata

IPUMS

- **Codebooks**
- **Questionnaires**
- **Enumerator instructions**
- **Census date and methodological notes**
- **Sample design information**

5

Variables and Categories

Full set of census topics

IPUMS

- **Full set of census topics**
- **As many census variables as possible**
- **Single year of age**
- **Geographic codes (subnational units)**
  - Place of residence (or usual residence)
  - Place of previous residence and birthplace (subnational and country)
  - 2nd or lower sub-national administrative level
  - Set lower bound for population per unit (e.g., 20,000)
- **3-digit occupation and industry**, if possible
- Option to omit or collapse sensitive info

6

Metadata

- Codebooks
  - Short and long variable names
  - Value labels
  - Include labels for different types of missing (unknown vs. not in universe)
  - Variable specification (format-dependent)
    - Location and width
- Questionnaires
- Enumerator instructions
- Sample design information

IPUMS

7

Exercises: Data for Research Plans

1. WHO Health Workforce Accounts

This research investigates the location and concentration of health workers, preferably to the second or third subnational geographic level and disaggregated by type of health workers (such as nurses, doctors, midwives, nurse practitioners, medical technicians, etc.).

2. Vulnerability of Older Adults

This study looks at living arrangements of older adults around the world. The research could be useful for understanding vulnerabilities of those living alone, potential familial support systems, and potential caregiving burdens of adult children.

3. SDG 8.6.1 – NEET (% young adults not in education, employment, or training)

We want to measure SDG 8.6.1. The indicator guidelines measure of the percent of young adults (defined as people age 15-24) who are not engaged in education, employment or training to be calculated as

NEET= (total youth - employed youth - youth in school)/(total youth)*100

IPUMS

8

Exercises: Data for Research Plans

For each research plan:

- What information (in the form of variables or census questions) is required for this research?
- Does the census have this information?
- Will published census reports have enough information to do this research?

Extra credit: Once you have learned something about the population of interest, what other information in the census could help you determine how to design policy, target resources, or otherwise determine how to improve lives?

IPUMS

9

WHO Health Workforce Accounts

This research investigates the location and concentration of health workers, preferably to the second or third subnational geographic level and disaggregated by type of health workers (such as nurses, doctors, midwives, nurse practitioners, medical technicians, etc.).

Variables:

- **Occupation**
  with detail (3-digit) about types of health care workers: nurses, doctors, midwives, nurse practitioners, medical technicians, etc.
- **Sub-national geography**
  at least 2nd subnational
-

Extra Credit:

- **Age, ethnicity, language, access to utilities or amenities, disability, fertility, and others**

In census? Yes

Need microdata? Yes

Use International Standard Classifications

IPUMS

10

Emergencies & Understanding Vulnerabilities: WHO on the Health Workforce

National Health Workforce Accounts (NHWA): Definition

A system by which countries progressively improve the availability, quality, and use of data on health workforce through monitoring of a set of indicators to support achievement of Universal Health Coverage, SDGs and other health objectives.

Documentation and tools available here:
www.who.int/hrh/statistics/nhwa/

WHO examples shared courtesy of Dr. Mathieu Boniol
Presentation for IPUMS International Pre-conference Workshop
ISI World Statistics Congress, Kuala Lumpur
August 18, 2019

11

WHO example: measuring the health workforce

Group code			Occupational title
Sub	Minor	Unit	
22			**Health professionals**
	221		**Medical doctors**
		2211	Generalist medical practiti
		2212	Specialist medical practitio
	222		**Nursing and midwifery pr**
		2221	Nursing professionals
		2222	Midwifery professionals
	223		**Traditional and compleme**
		2230	Traditional and compleme
	224		**Paramedical practitioners**
		2240	Paramedical practitioners
	226		**Other health professional**
		2261	Dentists
		2262	Pharmacists
		2263	Environmental and occupa
		2264	Physiotherapists
		2265	Dieticians and nutritionists
		2266	Audiologists and speech th
		2267	Optometrists and ophthalr
		2269	Health professionals not el

Group code			Occupational title
Sub	Minor	Unit	
32			**Health associate professionals**
	321		**Medical and pharmaceutical technicians**
		3211	Medical imaging and therapeutic equipment technicians
		3212	Medical and pathology laboratory technicians
		3213	Pharmaceutical technicians and assistants
		3214	Medical and dental prosthetic and related technicians
	322		**Nursing and midwifery associate professionals**
		3221	Nursing associate professionals
		3222	Midwifery associate
	323		**Traditional and con**
		3230	Traditional and com
	325		**Other health associ**
		3251	Dental assistants an
		3252	Medical records and
		3253	Community health v
		3254	Dispensing optician
		3255	Physiotherapy techi
		3256	Medical assistants
		3257	Environmental and
		3258	Ambulance workers
		3259	Health associate pr

Occupational information
ISCO International Classification at 3-digit or 4-digit level from IPUMS

Some info for 35 countries but detail at preferred level for only 14 countries

Group code			Occupational title
Sub	Minor	Unit	
53			**Personal care workers**
	532		Personal care workers in health services
		5321	Health care assistants
		5322	Home-based personal care workers
		5329	Personal care workers in health services not elsewhere classified
			Additional health-related unit groups
		1342	Health service managers
		1343	Aged care service managers
		2634	Psychologists
		2635	Social work and counselling professionals
		3344	Medical secretaries

IPUMS

12

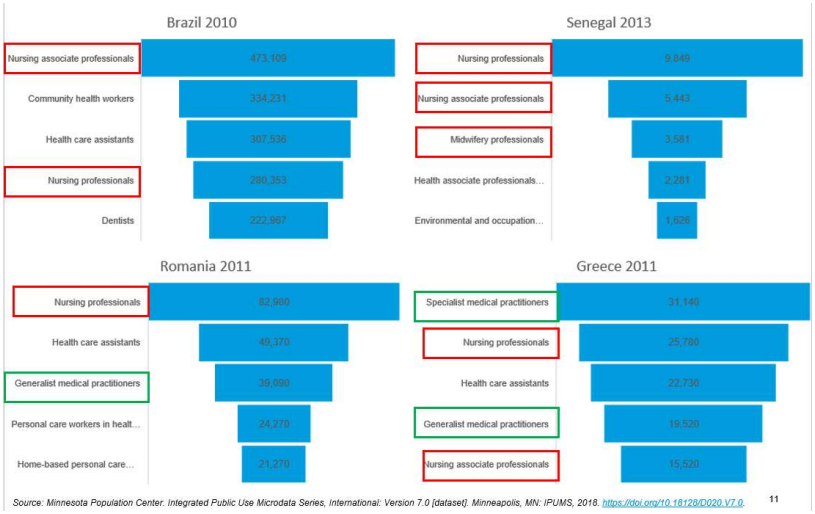
WHO example: measuring the health workforce

Top 5 health occupations for selected countries

Medical doctors

Nursing and midwifery personnel

Wide variability in medical professional availability and distribution



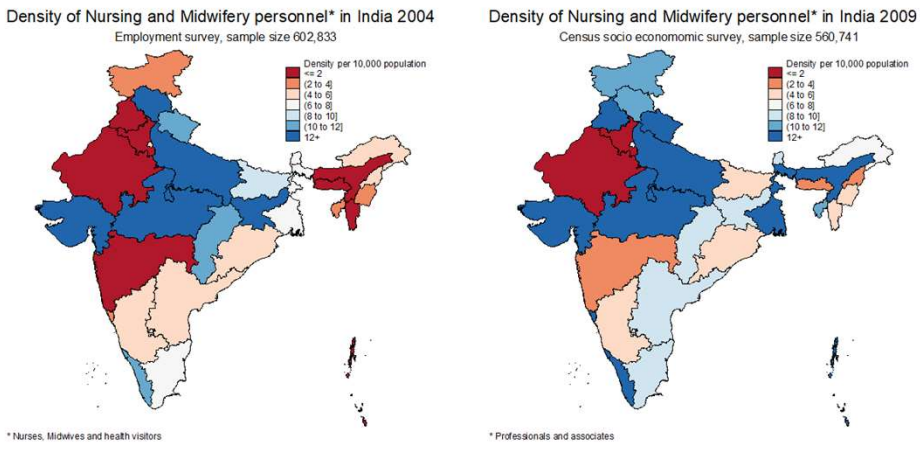
Source: Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 7.0 [dataset]. Minneapolis, MN: IPUMS, 2018. https://doi.org/10.18128/D020.V7.0.

11

IPUMS

13

WHO example: measuring the health workforce

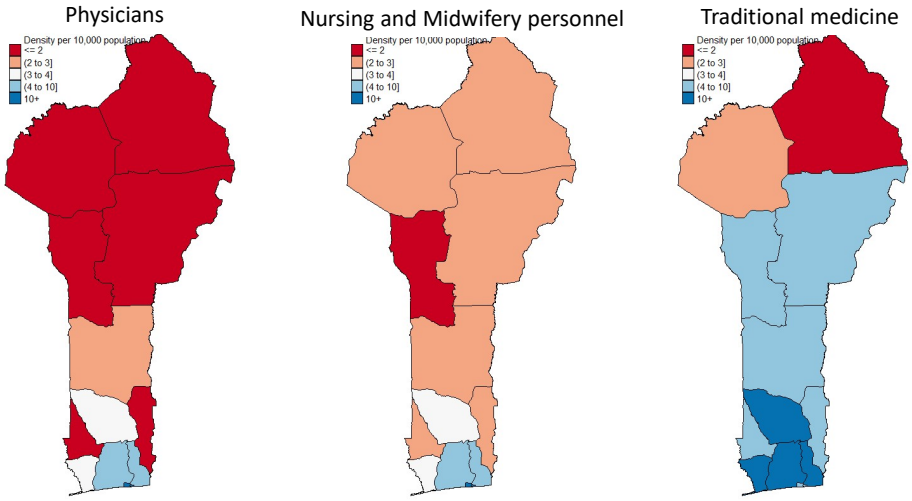


Source: Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 7.0 [dataset]. Minneapolis, MN: IPUMS, 2018. https://doi.org/10.18128/D020.V7.0. Data from the Ministry of Statistics and Programme Implementation, India

IPUMS

14

WHO example: measuring the health workforce



Source: Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 7.0 [dataset]. Minneapolis, MN: IPUMS, 2018. https://doi.org/10.18128/D020.V7.0. Data from the National Institute for Statistics and Economic Analysis, Benin

IPUMS

15

Vulnerability of Older Adults

This study looks at living arrangements of older adults around the world. The research could be useful for understanding vulnerabilities of those living alone, potential familial support systems, and potential caregiving burdens of adult children.

Variables:

- **Age**, especially older age groups
- **# people in the household** (or hh sample)
- **Relationship to head of household**

Extra Credit:

- **Subnational geography, disability, household amenities**

In census? Yes

Need microdata? Yes

Census is often the only viable source of data:

- **study of small groups**
- **household structure**

IPUMS

16

Household Structure of Older Persons

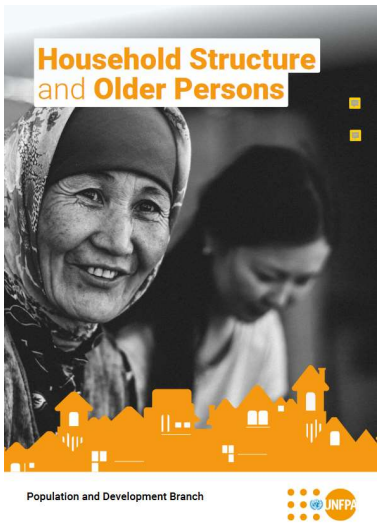
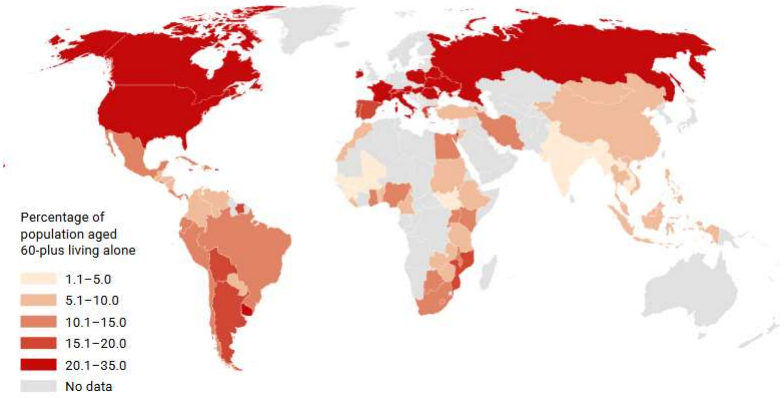


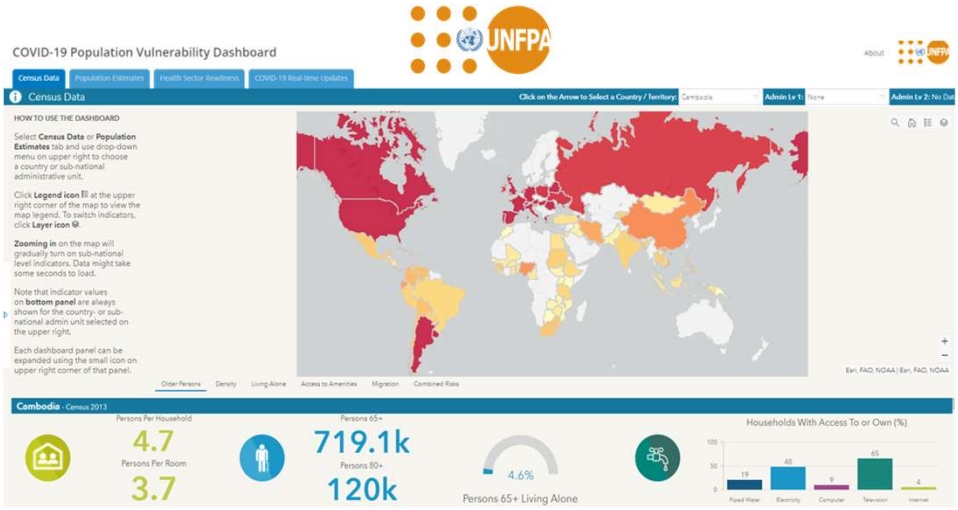
Figure 5. Percentage distribution of older persons living alone in countries with available data, 2000–2017



IPUMS

17

Census Data: Ready for Anything
Adapting to Covid Response Needs



IPUMS

18

Scholarly Research on COVID-19

National age and coresidence patterns shape COVID-19 vulnerability

Albert Esteve[a,b,1], Iñaki Permanyer[a], Diederik Boertien[a], and James W. Vaupel[c]

[a]Center for Demographic Studies, Centres de Recerca de Catalunya, 08193 Bellaterra, Spain; [b]Geography Department, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain; and [c]Interdisciplinary Centre on Population Dynamics, University of Southern Denmark, 5000 Odense, Denmark

Edited by Douglas S. Massey, Princeton University, Princeton, NJ, and approved June 8, 2020 (received for review May 8, 2020)

Based on harmonized census data from 81 countries, we estimate how age and coresidence patterns shape the vulnerability of countries' populations to outbreaks of coronavirus disease 2019 (COVID-19). We estimate variation in deaths arising due to a simulated random infection of 10% of the population living in private households and subsequent within-household transmission of the virus. The age structures of European and North American countries increase their vulnerability to COVID-related deaths in general. The coresidence patterns of elderly persons in Africa and parts of Asia increase these countries' vulnerability to deaths induced by within-household transmission of COVID-19. Southern European countries, which have aged populations and relatively high levels of intergenerational coresidence, are, all else equal, the most vulnerable to outbreaks of COVID-19. In a second step, we estimate to what extent avoiding primary infections for specific age groups would prevent subsequent deaths due to within-household transmission of the virus. Preventing primary infections among the elderly is the most effective in countries with small households and little intergenerational coresidence, such as France, whereas confining younger age groups can have a greater impact in countries with large and intergenerational households, such as Bangladesh.

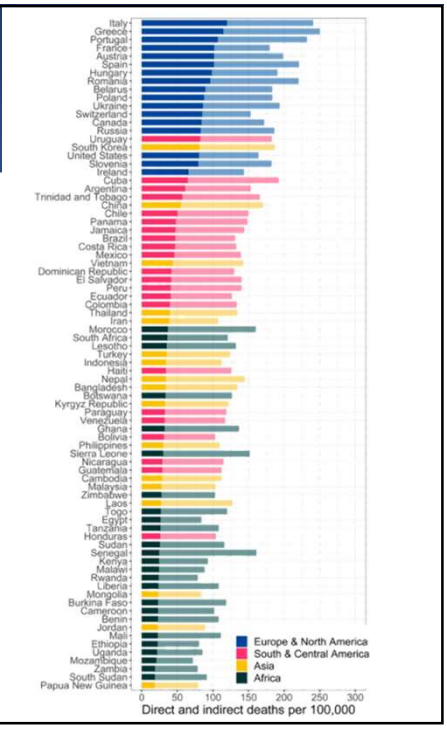
demography | households | COVID-19 | aging | global

The coronavirus disease 2019 (COVID-19) pandemic currently confronts nearly all of the world's countries. A growing number of governments are enforcing or recommending home quarantines to contain the spread of the virus. As the virus can be transmitted outside and within households, the effects of such measures will

infections). Lower rates of household transmission would reduce this number of indirect deaths proportionally. The direct effect depends on the age structure of the population; the indirect effect hinges on the size and age structure of households. Combined, they show how, all else equal, national age and coresidence patterns alter the vulnerability of a country to COVID-19 outbreaks.

The expected direct death rates per 100,000 individuals range from 19 in South Sudan to 120 in Italy. Together with Italy, three southern European countries—Greece, Portugal, and Spain—rank among the top six, followed by the rest of Europe and North America. Latin American countries form a homogenous cluster lower than the European and North American cluster. Asian countries spread all over the range, with estimates as high as 81 in South Korea and as low as 23 in Jordan. African countries tend to experience the lowest direct death rates. Where the elderly comprise a large portion of the population, the direct effect is high, whereas direct deaths are much lower where the elderly are vastly outnumbered by younger people.

Mortality due to intrahousehold contagion (right-hand segment of a bar in Fig. 1) does not follow the same order, because coresidence patterns differ widely across countries, even among those countries with similar age structures (4–7). The ratio between indirect and direct effects is a simple indicator of the importance of coresidence patterns, in particular, of the elderly, the most vulnerable group. For European and North American countries, direct and indirect deaths are roughly equal. In Latin America, indirect deaths could approximately double the number of direct deaths. The ratio between potential indirect and direct deaths in Asia ranges from 1.3 (South Korea) to 3.7 (Laos). In



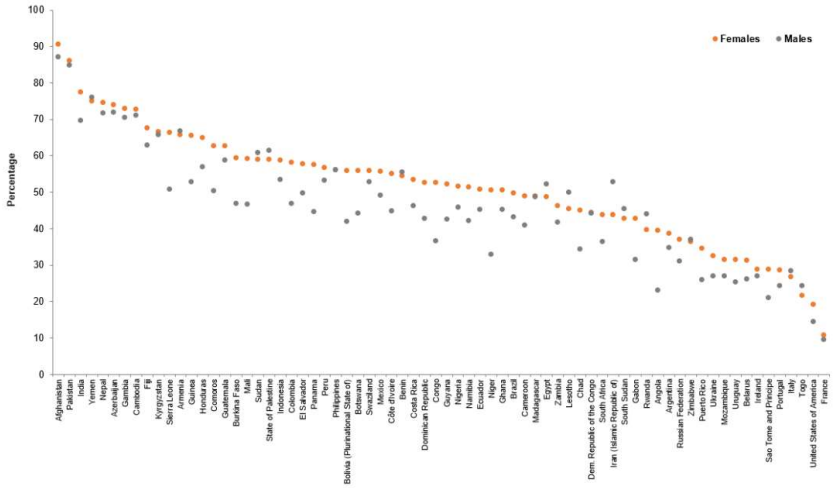
19

Older Adults Living with Children

Disaggregated by sex

Note utility of having household composition information

Percentage of men and women aged 65 years or over who co-reside with children aged 20 years or over, countries and areas with available data



Source: United Nations, Department of Economic and Social Affairs, Population Division (2020a). *Database on the Households and Living Arrangements of Older Persons 2019.*

IPUMS

20

SDG 8.6.1 – NEET

(% young adults not in education, employment, or training)

We want to measure SDG 8.6.1. The indicator guidelines measure of the percent of young adults (defined as people age 15-24) who are not engaged in education, employment or training to be calculated as

NEET= (total youth - employed youth - youth in school)/(total youth)*100

Variables:

- **Employment (in labor force)**
- **School attendance**
- **Age**

In census? Yes

Need microdata? Yes

Extra Credit:

- **Sex, sub-national geography,**

Disaggregation

IPUMS

21

Disaggregation for SDGs

IPUMS

22

Disaggregation and Study of Small Populations: SDG Reporting

Census Microdata

110 of 169 Targets for 11 of the 17 Goals

Multidimentional crosstabulation and investigation

Household
- Household composition
- Dwelling ownership
- Household amenities
- Access to utilities
- Group quarters
- Subnational geography

Person
- Fertility
- Mortality
- Migration
- Education
- Labor-force participation
- Occupational structure
- Ethnicity
- Disability

IPUMS supports the Sustainable Development Goals

IPUMS

23

Disaggregation and Study of Small Populations: SDG Reporting

Census Microdata

To fulfill recommended disaggregation

- Sex and gender
- Age
- Income quintiles or deciles
- Disability
- Ethnicity and indigenous status
- Economic activity
- Geographic location (at least urban/rural)
- Migrant status

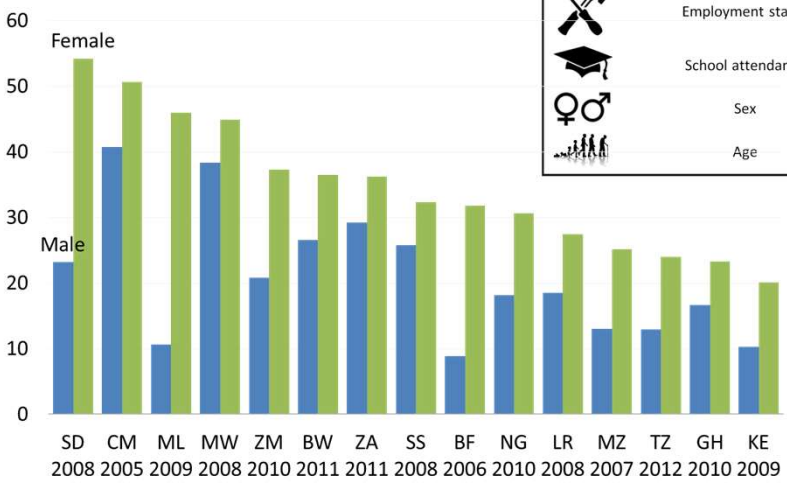
IPUMS supports the Sustainable Development Goals

IPUMS

24

Disaggregation for SDGs

Percentage of young people not in education, employment, or training (NEET), Africa



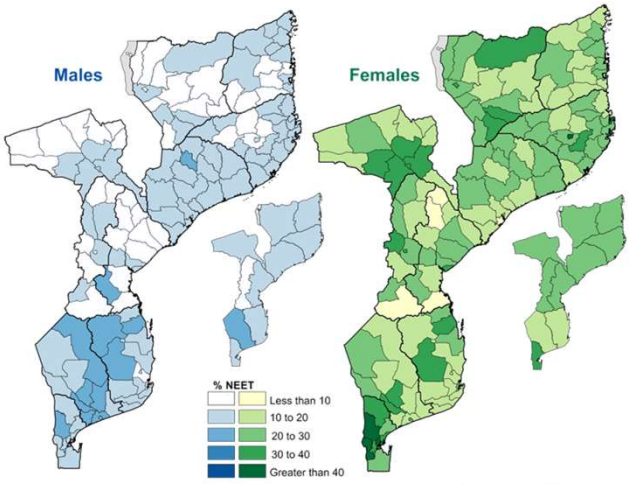
Input	IPUMS variable
Employment status	EMPSTAT
School attendance	SCHOOL
Sex	SEX
Age	AGE

IPUMS

25

Disaggregation for SDGs

Percentage of young people not in education, employment, or training (NEET), Africa



Input	IPUMS variable
Employment status	EMPSTAT
School attendance	SCHOOL
Sex	SEX
Age	AGE

$$\text{NEET} = \frac{\text{total youth} - \text{employed youth} - \text{youth in school}}{\text{total youth}} * 100$$

Mozambique, 2007

IPUMS

26

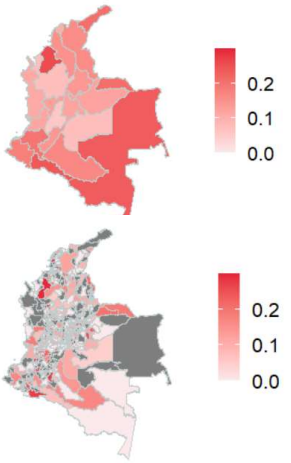
Census to Extend the Power of Other Data Sources: Small Area Estimation (Survey + Census)

Surveys

- Rich topical coverage and detail
- Small sample sizes = limited disaggregation power

Small area Estimation

- Match survey to census on key characteristics
- Extend inference to smaller geographic areas



27

Census to Extend the Power of Other Data Sources: Small Area Estimation (Survey + Census)

Goal 1: End poverty in all its forms everywhere

Target 1.1: By 2030, eradicate extreme poverty for all people everywhere, currently measured as people living on less than $1.25 a day.

Indicator 1.1.1: Proportion of the population living below the international poverty line by sex, age, employment status and geographic location (urban/rural)

Goal 8: Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all

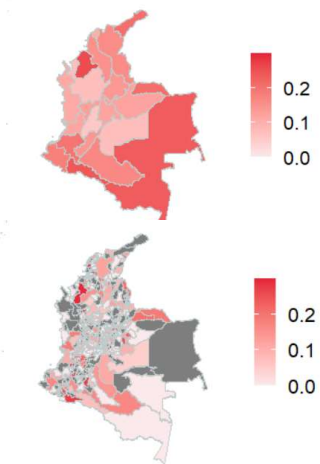
Target 8.5: By 2030, achieve full and productive employment and decent work for all women and men, including for young people and persons with disabilities, and equal pay for work of equal value

Indicator 8.5.2: Unemployment rate, by sex, age and persons with disabilities

Goal 11. Make cities and human settlements inclusive, safe, resilient and sustainable

Target 11.2: By 2030, provide access to safe, affordable, accessible and sustainable transport systems for all, improving road safety, notably by expanding public transport, with special attention to the needs of those in vulnerable situations, women, children, persons with disabilities and older persons

Indicator 11.2.1: Proportion of population that has convenient access to public transport, by sex, age and persons with disabilities



Survey size limits direct estimation (e.g., Colombia from UN SAE Toolkit)

28

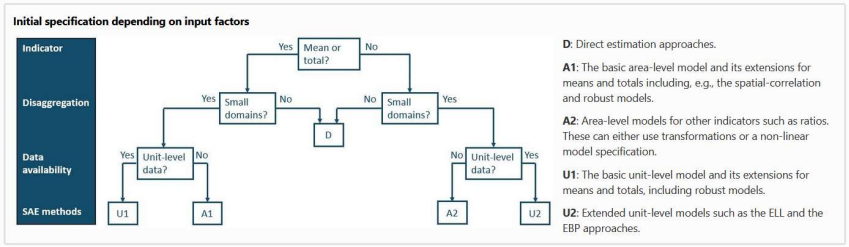
Small Area Estimation
The UN Statistics SAE Toolkit

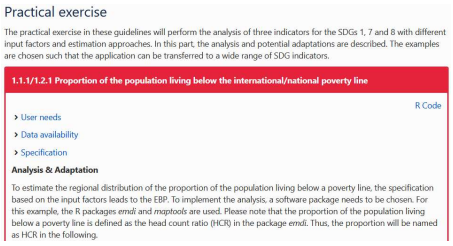
29

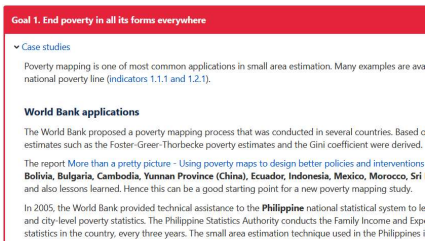
Small Area Estimation
The UN Statistics SAE Toolkit

Many resources available

Methodology
Training materials
References
Practical exercises
Case studies







IPUMS https://unstats.un.org/wiki/display/SAE4SDG/

30

Census to Extend the Power of Other Data Sources: Environmental Pollutants

Vol. 125, No. 9 | Research

Changes in Transportation-Related Air Pollution Exposures by Race-Ethnicity and Socioeconomic Status: Outdoor Nitrogen Dioxide in the United States in 2000 and 2010

Lara P. Clark, Dylan B. Millet, and Julian D. Marshall

Published: 14 September 2017 | CID: 097012 | https://doi.org/10.1289/EHP959 | Cited by: 8

Estimated Changes in NO_2 Environmental Injustice Metrics

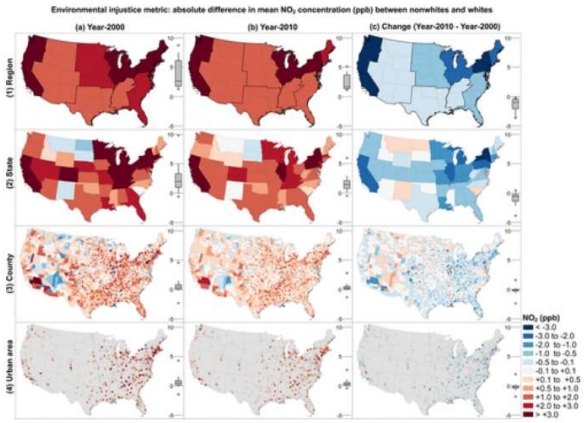
Nationally, on an absolute basis, environmental injustice declined from 2000 to 2010. The nonwhite–white NO_2 disparity decreased from 5.0 ppb in 2000 to 2.9 ppb in 2010 (−2.1 ppb [−42%]; Table 2). However, nationally, on a relative basis, environmental injustice persisted. Nonwhites remained more exposed to outdoor NO_2 air pollution than whites on average in 2010, and there was little change in the relative NO_2 difference between nonwhites and whites between 2000 and 2010: The nonwhite–white NO_2 difference was 33% in 2000 (nonwhites were 40% more exposed than whites) and 31% in 2010 (nonwhites were 37% more exposed than whites).

Table 2 Estimated population-weighted mean NO_2 concentrations (ppb) for nonwhites and whites: year 2000, year 2010, and change over time (year 2010–year 2000).

Race-ethnicity	2000	2010	Change: 2010–2000
Nonwhites[a]	17.6	10.7	−6.9 (−39%)
Whites[b]	12.6	7.8	−4.7 (−38%)
Difference: nonwhites–whites	5.0 (33%)	2.9 (31%)	−2.1 (−42%)

[a]Nonwhites includes all race-ethnicity minority groups (i.e., people who reported any race-ethnicity other than white alone, non-Hispanic).

[b]Whites includes people who reported white alone, non-Hispanic race-ethnicity.



31

IPUMS Award Winners: Census + Climate, Economic Surveys, & More

PLOS ONE

OPEN ACCESS PEER-REVIEWED
RESEARCH ARTICLE

An introduction to DUIA: The database on urban inequality and amenities

Frederico Roman Ramos, Justus Uitermark

Published: June 25, 2021 • https://doi.org/10.1371/journal.pone.0253824

DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

Articles Special Collections for Authors for Readers About the Journal

ARTICLES
» Current Volume
» Older Volumes
» Volume 39
» Editor's Choice
» Replicable Articles
» by Author
» by Subject
» Search
CITATIONS

VOLUME 39 - ARTICLE 29 | PAGES 835–854

Educational selectivity of internal migrants: A global assessment

BY » Aude Bernard, » Martin Bell

DOWNLOAD PDF SUBMIT A RESPONSE LETTER

DATE RECEIVED: 30 Apr 2018
DATE PUBLISHED: 10 Oct 2018
WORD COUNT: 2548
KEYWORDS: » development, » education, » internal migration, » migration, » selectivity
DOI: » 10.4054/DemRes.2018.39.29

ELSEVIER

Global Environmental Change
Volume 65, November 2020, 102183

Climate-Induced migration and unemployment in middle-income Africa

Valerie Mueller, Clark Gray, Douglas Hopping

Show more

Population Studies
A Journal of Demography

ISSN: 0032-4728 (Print) 1477-4747 (Online) Journal homepage: https://www.tandfonline.com/loi/rpst20

Until work do us part: Labour migration and occupational stratification in non-cohabiting marriage

Giulia Ferrari & Ross Macmillan

To cite this article: Giulia Ferrari & Ross Macmillan (2019): Until work do us part: Labour migration and occupational stratification in non-cohabiting marriage, Population Studies

To link to this article: https://doi.org/10.1080/00324728.2019.1583359

32

Questions

IPUMS

Lara Cleveland, PhD
University of Minnesota
clevelan@umn.edu
ipumspartners@umn.edu

@ipumsi

33

34

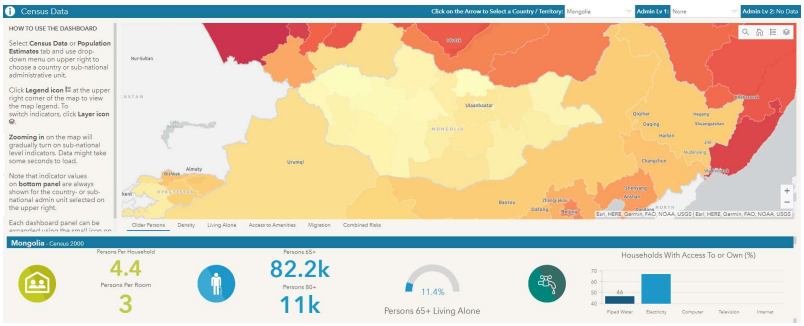
Responding to Emergencies & Understanding Vulnerabilities

Census-based indicators for SDG reporting:

United Nations Population Fund

SUSTAINABLE DEVELOPMENT GOALS

- Fertility
- Sex ratio
- Age composition
- Nuptiality
- Household living arrangements
- Migrant status
- Education level
- Labor force participation
- Disability status



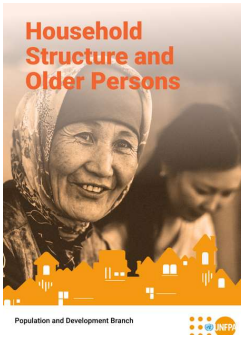
IPUMS provided more than 8 million data cells (without geographic disaggregation)
Included cell suppression where appropriate
Provided statistical confidence intervals.

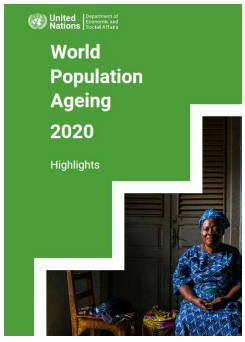
IPUMS

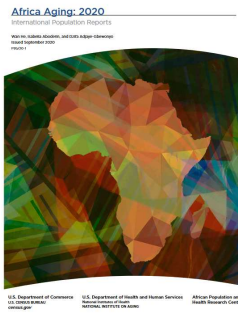
35

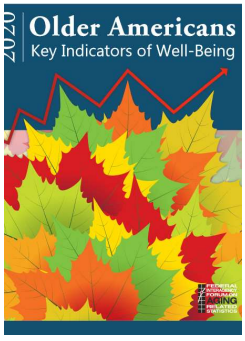
Disaggregation and Study of Small Populations:

Research on Aging









Publications by the United Nations, U. S. Census Bureau, and U.S. Federal Agencies on Global Aging Trends

36