Data Prospects: IPUMS-International

8

Matt Sobek

The IPUMS Projects

The Integrated Public Use Microdata Series (IPUMS) is a suite of population data projects developed at the Minnesota Population Center of the University of Minnesota. The largest of these, IPUMS-International, is the world's most extensive collection of publicly accessible population microdata. Related IPUMS databases consist of U.S. census and survey data and historical census data from Europe and North America (Sobek et al. 2011; Ruggles 2014). All IPUMS data can be accessed at no cost by qualified researchers.

The IPUMS projects share a number of key characteristics. Each record describes a person, and those individuals are organized into households. The projects follow a similar approach to data harmonization, documentation and dissemination. Common variables are coded and labeled consistently, and documentation describes comparability issues for each of these harmonized variables. All of this information is presented via a web interface that limits the display to a user's selected samples of interest. A data extraction system allows researchers to select only the variables and samples they require, defining a customized pooled dataset that they download for analysis on their own desktop. The system delivers the individual-level data on specific persons, not tables or other summary measures. Because variables are harmonized across time and place, IPUMS is optimized to support comparative research.

This chapter focuses on the largest of the Minnesota data projects: IPUMS-International (henceforth, simply "IPUMS"). The database currently contains information on 560 million people in 79 countries. For most countries IPUMS provides information directly relevant to the study of both internal and international migration, such as place of birth, prior residence, and duration of residence. The chapter starts by describing some of the attributes of IPUMS owing to its nature as census data processed for scientific use. The temporal and geographic scope of the series is discussed next, followed by the general topical coverage of the variables. Most of the remaining discussion concerns the specific migration data available in IPUMS. The chapter concludes with a brief note on data quality and a discussion of potential research directions with these data.

Census Microdata

IPUMS is a collection of nationally representative samples of individuals from population censuses around the world (Ruggles et al. 2003; Minnesota Population Center 2014). Each sample

M. Sobek (✉)
University of Minnesota, Minnesota Population Center, Minneapolis, MN, USA
e-mail: sobek@umn.edu

M.J. White (ed.), *International Handbook of Migration and Population Distribution*, International Handbooks of Population 6, DOI 10.1007/978-94-017-7282-2_8

is a unique cross section, and records cannot be linked across them. The samples are large: often five to ten percent of the national population. Because of their size, it is possible to study small population subgroups that cannot be analyzed using other sources. This can be particularly salient for migration studies that focus on specific stocks or flows. Most samples have information on migration, and all of a person's other characteristics are known as well, allowing sophisticated multivariate analysis. The IPUMS data are normally taken from the long-form census schedules when those were used, and most of the samples include housing information. Nevertheless, the data are limited by the range of questions that were asked by a given census, which are typically fewer than those included in social or economic surveys.

IPUMS samples are composed of microdata: each record describes the attributes of a single person. In addition to their personal characteristics, individuals are organized into households. This hierarchical household-person structure gives the data much of its power, making it possible to inter-relate the characteristics of co-resident people in creative ways. To fully exploit this feature of the data, IPUMS constructs "pointer" variables that identify the location within the household of each person's mother, father, and spouse, if they were present (Sobek and Kennedy 2009). The pointers make it straightforward to compare the characteristics of spouses, to attach parents' characteristics to children or vice versa, and to construct unique household or family-level measures. For example, one can make a variable for spouse's migration status, mother's birthplace, or number of own children in school. The household organization of the data makes it well suited to analyses of migration effects on the family economy and household structure.

The IPUMS data have some limitations inherent to their origin as public use samples drawn from censuses. Although large, the data are still samples and sometimes have too few cases to yield reliable results for certain subpopulations. The sample designs also differ, and this can have an impact on variance estimation (Cleveland et al. 2011). The biggest practical limitation for most researchers, however, stems from measures taken to prevent identity disclosure of people in the database. Of these measures, the greatest concern for migration analysis is the suppression of low-level geography. As a rule, IPUMS does not identify places with less than 20,000 population, combining smaller units until they meet that threshold. Some countries impose their own higher minimum population requirements. Fortunately, most countries provide geography for at least their first- and second-level political divisions, such as governorates and districts in Egypt or departments and municipalities in Colombia; but some only provide the first-level divisions—the equivalent of states in the United States or Brazil. The lack of small-area geographic detail can make it difficult to disaggregate cities from their surrounding regions and impossible to specify villages and other small places. Some of these and related limitations will be discussed where relevant below.

Scope of the IPUMS Database

The 2014 version of IPUMS includes 258 census microdata samples from 79 countries, documented in Table 8.1. The database covers the full spectrum of economic development: roughly three quarters of the countries and two thirds of the samples are from the developing world. Seventeen countries are on the United Nations list of Least Developed Countries. The temporal scope of the data series is 1960 to the present, but there is a lag of two or more years before the most recent census conducted by a country becomes available. Because most countries have multiple samples, it is usually possible to analyze change over time at national and sub-national levels. Sample densities typically vary from one to ten percent of the national population. The median sample size is 805,000 records, and the database has roughly 560 million person records in total.

IPUMS is designed to facilitate cross-national and cross-temporal research. The data extract system lets users define pooled datasets that

Table 8.1 Number of IPUMS samples by country (258 total)

Argentina	5		Fiji	5		Malawi	3		Senegal	2
Armenia	1		France	7		Malaysia	4		Sierra Leone	1
Austria	4		Germany	4		Mali	3		Slovenia	1
Bangladesh	3		Ghana	2		Mexico	7		South Africa	3
Belarus	1		Greece	4		Mongolia	2		South Sudan	1
Bolivia	3		Guinea	2		Morocco	3		Spain	3
Brazil	6		Haiti	3		Nepal	1		Sudan	1
Burkina Faso	3		Hungary	4		Netherlands	3		Switzerland	4
Cambodia	2		India	5		Nicaragua	3		Tanzania	2
Cameroon	3		Indonesia	9		Nigeria	5		Thailand	4
Canada	4		Iran	1		Pakistan	3		Turkey	3
Chile	5		Iraq	1		Palestine	2		Uganda	2
China	2		Ireland	9		Panama	6		Ukraine	1
Colombia	5		Israel	3		Peru	2		UK	2
Costa Rica	4		Italy	1		Philippines	3		USA	7
Cuba	1		Jamaica	3		Portugal	3		Uruguay	6
Dominican Rep.	5		Jordan	1		Puerto Rico	5		Venezuela	4
Ecuador	6		Kenya	5		Romania	3		Vietnam	3
Egypt	2		Kyrgyzstan	2		Rwanda	2		Zambia	3
El Salvador	2		Liberia	2		Saint Lucia	2			

include any variables they desire from as many times and places as they choose. Using the extract system it is feasible to build a single dataset containing selected variables for all half-billion persons in the database. If such a dataset would be too large, the system is capable of drawing a systematic subsample of cases. Of course, most analyses are more localized in time and place, but IPUMS offers the unique potential for truly globe-spanning research. This is a practical possibility not only because of the data extract system, but also due to the harmonization of the variable codes and to the documentation system that integrates information at the variable level across samples (Esteve and Sobek 2003). A key feature of the variable documentation compiles the census questionnaire text for all requested samples on one screen, enabling researchers to discern for themselves how question wording might affect comparability. Thus, the primary logistical barriers to cross-national studies are removed, freeing researchers to focus on substantive issues.

IPUMS disseminates data with the permission of each country's National Statistical Office. New samples are regularly added to the database, including additional countries as well as more recent censuses that add chronological depth for existing countries. A majority of the IPUMS samples are not readily accessible, if at all, from other sources. Most of the statistical offices participating in IPUMS lack their own mechanism to develop and distribute public use microdata. IPUMS does include samples that are in distribution elsewhere, but it constructs new technical variables to enhance analytical power, and may conduct minimal data editing in addition to its trademark practice of harmonizing the data to a global standard.

The geographic scope of the data series is ever-expanding, but it is richest in the Americas. Latin America has remarkable coverage, largely because of the efforts of the UN Statistical Office in Chile (CELADE), which has been archiving that region's census data for decades. There are also concentrations of IPUMS countries in Europe and parts of Africa and Asia. Some populous and highly developed countries have yet to be persuaded to join the data partnership, including Australia, Japan and Russia. The data for India and Nigeria are survey data, because an agreement to distribute the censuses has not yet been reached with their National Statistical Offices. It should also be noted that sometimes

the most recent sample for a country is fairly old, stemming from any number of reasons. Non-participating countries are regularly re-approached to join the project.

Geographic Harmonization

Geographic harmonization warrants special mention. Geography is of critical importance for migration studies, but it poses unique comparability challenges because of change over time in administrative boundaries. Principally, this is an issue for studies of internal migration, and more so for the developing world, where changes are more frequent. Most often, political units in older censuses get subdivided because of population growth. Less frequently, areas merge, boundaries move, or small units are reassigned between higher administrative levels during more thorough-going reorganizations.

IPUMS has a two-pronged approach to geography. For each country a harmonized place of residence variable holds geography stable by aggregating places into the smallest units that are consistent over time. For example, if place A divided into A and B at some point, and C later split off from B, then the unit ABC is constructed for all years. The harmonized variable amounts to a least common denominator of geographic detail over time. There are fewer, bigger units, but they define the same geographic space in each sample, and one GIS boundary file will apply across all years. A second variable for each country provides full, unaltered geographic detail for each independent sample year. Places receive the same codes across samples based on their names, but some units may not exist in all years, and their spatial footprints may change.

The internal migration geography variables—birthplace and previous residence—use the second, name-based approach to harmonization. Their codes match the name-based place of residence variable for the country, allowing direct comparisons within samples. Thus, full geographic detail is maintained, but the identified places may undergo boundary changes over time, potentially complicating temporal analysis. GIS boundary files apply to the most recent census year. IPUMS will likely construct temporally-stable birthplace and previous residence variables in the future, at least for the first subnational administrative level.

Geographic variables pose challenges with respect to scale as well as time. At the first administrative level within countries the number of geographic changes is generally limited, and comparisons over time are manageable using the existing name-harmonized variables. But roughly a quarter of samples also report birthplace or previous residence at the second level, such as counties or districts. Analyses over time at the second level may require historical geographic knowledge on the part of the researcher. IPUMS GIS boundary files apply to the most recent sample year for the first administrative level for previous residence and birthplace. In most cases GIS boundary files are not available at the second subnational level, but these will be added where possible going forward.

General Topical Coverage

The topical coverage of the IPUMS samples is dictated by the censuses from which they derive. The samples typically include variables corresponding to most or all of the questions asked in a census, but the lengths of the underlying questionnaires differ. Table 8.2 lists the most common types of variables available in the censuses. Most samples include some migration-related variables, such as place of birth or previous residence, in addition to a fairly consistent core of demographic and socioeconomic variables. All censuses have basic demographic information such as age, sex, and marital status. Questions on education and employment are likewise nearly universal, but they employ a wide variety of classification schemes and sometimes reflect national idiosyncrasies that make comparisons difficult. Fertility information is also widely available—more consistently so in developing countries—and ethnicity, language and religion are fairly common. The most direct economic measure, income, is rarely asked in censuses; thus

Table 8.2 Selected topical coverage of harmonized IPUMS variables

Group	Variable	Group	Variable
Household Variables		**Person Variables (continued)**	
Geography	First administrative level	*Fertility/Mortality*	Children ever born
	Second administrative level		Children surviving
	Urban–rural status		Details of most recent birth
Dwelling	Number of rooms		Parental mortality
	Toilet access	*Birthplace/Nativity*	Place of birth
	Construction materials		Country of birth
	Age of structure		Citizenship
	Living area		Year of immigration
Utilities	Electricity	*Ethnicity*	Religion
	Water		Race
	Sewage		Ethnic group
	Fuel		Language spoken
	Heating		Mother tongue
Amenities	Automobiles	*Education*	School attendance
	Washer		Literacy
	Television		Educational attainment
	Computer		Years of schooling
	Phone	*Work*	Employment status
Other	Home or land ownership		Occupation
	Number of deaths		Industry
	Number of international migrants		Class of worker
Person Variables			Hours worked
Demographic	Age	*Income*	Total income
	Sex		Wage and salary income
	Marital status		Source of livelihood
	Relationship to householder	*Disability*	Disability status
Migration	Residence 1/5 years ago or last move		Type of disability
	Years in current locality		Cause of disability

Selected variables only. Variable availability differs across samples

researchers often use education or occupation as a proxy. There is no information on household consumption or on the attitudes or opinions of household members. IPUMS is exploring the potential for developing one or more international socioeconomic scales based on occupation or education (e.g., Ganzeboom and Treiman 1996).

All samples identify the location of the residence in which a household was enumerated—with varying geographic specificity—and most include housing characteristics describing some physical attributes of the dwelling. Home ownership is widely available, although there is considerable diversity in the tenure arrangements reported. Many samples include access to basic utilities such as electricity, water and sewage, though this is more common among developing countries. Basic dwelling attributes such as number of rooms, presence of a toilet, amount of living area, and age of the dwelling are provided for a varying number of samples. The materials from which the dwelling is constructed are intermittently recorded in several variables, but the categories are difficult to harmonize across countries. Many samples also report the presence of various household assets, such as televisions, automobiles, and computers. IPUMS plans to construct a wealth index for developing countries based on housing characteristics, modeled on techniques developed for the Demographic and Health Surveys (Filmer and Pritchett 2001; Rutstein and Staveteig 2014).

By default, the IPUMS data browsing system displays the variables that have been internationally harmonized. All samples, however, include additional variables that were not suitable for harmonization for a number of reasons: they are sui generis or rare; they use a unique classification that is incompatible with the international standard; or there is something conceptually distinct about the underlying census question that would tend to mislead researchers if put in the context of a harmonized variable. IPUMS does not lose information. All these unharmonized sample-specific variables are available through a selection in the web browsing system. The system also identifies which variables serve as inputs for internationally harmonized variables, allowing researchers to deconstruct the IPUMS recodes and potentially devise their own. The unharmonized variables might have additional detail or different information on the topics listed in Table 8.2, or they might cover subjects not commonly included in censuses, such as contraceptive practice or household ownership of livestock or agricultural implements. As more samples accumulate in the IPUMS database, a critical mass of information on a specific topic occasionally develops, and a new harmonized variable is created to organize this information for researchers.

Migration Data

The IPUMS database contains considerable information on migration. The movement of populations is of great interest to all national statistical offices, and most censuses contain one or more questions on the topic. Table 8.3 shows the availability of key migration variables across the samples in the IPUMS database. The most widely available migration variables are of two general types: place of birth and place of residence at some time prior to the census. Each type records internal as well as international migration. An additional set of less common variables include duration of current residence, year of immigration, urban–rural status of previous residence, nationality, and reason for migration.

Place of birth is an indicator of lifetime migration. One knows the person migrated, but not when they moved or whether they made intervening moves. Place of birth somewhat underestimates lifetime migration, because some people return to their birthplace after living elsewhere. Most IPUMS samples report country of birth, thus identifying the net lifetime immigration of each foreign stock to every region and locality within the recipient country. Some IPUMS samples identify only a handful of specific countries of origin while others may identify a hundred or more; however, significant nations of origin that apply to each country are usually specified. A subset of samples provide only nativity status: they identify the foreign-born without giving their specific country of origin. No censuses record the subnational place of birth of foreign-born persons.

Place of birth for the native-born is even more widely available than country of birth. Although such subnational birthplace information is common, a majority of censuses record only the largest administrative units, such as state or province, limiting opportunities for fine-grained geographic analysis. Political boundary changes over time can be especially problematic for internal lifetime migration. Many decades may have passed for the respondent, with more potential for boundary changes and greater scope for ambiguity with respect to use of historical or modern place names for areas.

Previous-residence data are the most useful for many migration analyses. The data of this type most frequently available in IPUMS report a person's usual residence 1 or 5 years prior to the census. These period data are more likely than birthplace data to be reported at the second administrative level of the country, but the first level is more common. In IPUMS currently, the second-level geographic detail for birthplace and previous residence is available only via the unharmonized source variables. The period migration data also usually indicate the prior country of residence for international migrants. In some cases this may not be the actual country

Table 8.3 Availability of migration variables in IPUMS

Variable	N of Samples
Migration status: 1 year ago	34
Migration status: 5 years ago	93
Migration status: previous residence	75
Major/minor administrative division 1 year ago	37
Major/minor administrative division 5 years ago	83
Major/minor administrative division, previous residence	71
Country of residence 1 year ago	25
Country of residence 5 years ago	50
Country of previous residence	49
Urban status 1 or 5 years ago	12
Urban status, previous residence	17
Years residing in current locality	88
Nativity status	216
Country of birth	160
Major administrative division of birth	191
Citizenship status	133
Country of citizenship	90
Year of immigration	54
Reason for migration	22
International migrant from household	14

Some rows represent multiple variables. The universe is 258 samples

of origin of the migrant, but a step in a longer migration process.

Roughly half the IPUMS samples have period migration data, with about two-thirds of those reporting residence 5 years ago, and most of the rest one year ago. The 1- and 5-year retrospective migration data are not directly comparable. The more common 5-year variable offers a longer time window in which intervening moves, return migration, and mortality could occur. To its benefit, it tends to yield roughly five times as many migrants for study as the 1-year measure. It is also worth noting that a small number of samples have longer period migration variables of 10 years, or ones pegged to the previous census, which can be convenient for intercensal measurement.

In contrast to the specific-period migration questions, a substantial number of samples provide a person's previous residence without imposing any time frame on the question. Similar to birthplace data, one can tell the person is a lifetime migrant, but not when they moved, unless the census also asked a duration question. One does, however, know the most recent place from which the person migrated with no potential for intervening moves. In combination with birthplace, this variable can provide two data points for a given migrant, allowing study of return migration. A small number of samples have both previous residence and a fixed-date residence variable, potentially offering three data points for recent migrants. Assuming there are enough cases, it offers the possibility of studying migrants who enter a country by passing through another.

IPUMS constructs migration status variables that summarize the previous residence information. The variables record if a person migrated within the time frame of the variable between minor administrative units (when possible), between major administrative units, or between countries. These summary variables do not distinguish between different migrant streams, but they do identify short and long-distance migration as they are often operationalized. Enterprising researchers can further delineate migrants into those moving between contiguous and non-contiguous units, but IPUMS does not construct that information for users.

Data on duration of residence are a different type of migration information broadly available in censuses. The internal migration variables take two forms: years in current locality and years in current dwelling. The "locality" in this context can differ considerably in size across countries. Most samples refer to movement at the village or municipality level, but some only report migration between larger administrative units. The years-in-dwelling variable is limited in scope, but it is the only migration variable not subject to any measurement issues regarding reference periods or geographic scale. In combination with other variables it can identify short-distance moves that did not involve a change in locality: mobility as opposed to migration.

International migrants report their year of arrival in the country of residence in a number of samples. From this information IPUMS also calculates the number of years since immigration, subject to some months of rounding error, depending on the date of the census within the calendar year. In a few cases the information is restricted to non-citizens, but the data are otherwise fairly consistent in recording all foreign-born persons' date of arrival to take up residence. For each of the duration variables the data are sometimes reported in intervals rather than individual years. To make the data easier to use across samples, IPUMS converts the grouped data into pseudo-continuous form by recoding to the midpoint or the first year of the interval. Researchers must therefore take care when making certain comparisons or when calculating age at migration. The comparability documentation for the variables specifies the samples that were converted from intervals. For the samples with truly continuous data, the duration migration variables can reveal whether family members migrated together or within close proximity to each other.

Citizenship status for foreign-born persons is reported in roughly half the IPUMS samples. A number of those also distinguish naturalized citizens and stateless persons. A sizeable subset of samples indicates the specific nationality of residents; although, as with birthplace, the number of identified categories varies greatly from one sample to the next.

Over twenty IPUMS samples report urban–rural status prior to migration, almost all of them from developing nations. Countries define urban status differently, but the census migration question usually depends on the respondents, who are likely to have a fairly colloquial interpretation of "urban." It behooves the researcher to examine the census form to see exactly how the data were obtained. At this writing, IPUMS has not created internationally harmonized urban–rural migration variables. These data can nevertheless be accessed as unharmonized source variables specific to the various samples.

Approximately ten percent of IPUMS samples, all for developing countries, report a person's reason for migration. Most of these samples also include information on the number of years since migration, aiding in the interpretation of the data. All samples identify work, family and study as reasons for migration, with different types of labor migration often being delineated. Marriage is often indicated as a cause, and sometimes divorce and widowhood. A variety of other reasons are listed in various samples, with a concentration in types of forced migration due to war, disaster, or insecurity.

A handful of IPUMS samples provide a different class of migration data that does not fit within the normal IPUMS data structure: individual records for people who migrated abroad in some span of time prior to the census. These individuals do not receive regular person data in the IPUMS, because they are no longer residents in their households, or even in the country. Because these migrant records do not conform to the basic IPUMS data scheme, they are available as separate stand-alone files that can be downloaded and linked to data extracts using the household serial number. The information in the migration records is limited, so these files are not especially rich objects for investigation in themselves without linking to their households of origin. Most records indicate the age and sex of the migrant, when they left, where they went,

and possibly their reason for migrating. Because the IPUMS data are samples of ten percent or lower density, the records typically number at most a few tens of thousands. There are measurement issues as well. If a whole household migrated or dissolved, then migrant data do not exist for those persons.

Data Quality

IPUMS does not provide summary measures of data quality, but there are plans to do so. A key challenge is the difficulty devising measures that can be calculated for the entire database despite differences across samples. And there is some concern regarding the potential of specific indicators to convey a mistaken impression of the utility of the microdata samples as general purpose scientific-use datasets. For example, coverage error, such as a population undercount may not be problematic for the kinds of multivariate analyses conducted by most researchers. Content errors affecting particular variables, on the other hand, may pose more serious problems. They can stem from poor reporting or flawed data processing. Some of the latter type of error can be corrected by IPUMS when there are identified.

It is relatively easy to calculate summary measures of the quality of age-sex reporting in the IPUMS samples. The Whipple's and Myers' Indices are measures of digit preference in age reporting: the former gauges preference for digits ending in 0 or 5, and the latter for any digits. The general impression of IPUMS samples from the age quality measures is not surprising: the older samples—those from the 1960s and 1970s—are typically of lower quality than those from more recent decades; and the data from developed countries on average appear more accurate than those from developing countries. Table 8.4 presents Whipple's index values for Latin American samples in IPUMS. The measures are broadly consistent with impressionistic observations from IPUMS data processing. The samples with poor age reporting were also more prone to data structure problems like malformed households or errors in technical variables, presumably because of the limited computing resources available in decades past. But this is only a generalization, and there are outliers among old and new samples and rich and poor countries.

More sophisticated demographic evaluation methods employing information drawn from outside the census would be a significant undertaking to apply across the collection of IPUMS samples. A more limited approach to assessing data quality with respect to migration is to conduct internal consistency checks across selected variables within a sample. For example: how many persons under age five report a residence 5 years ago; or what proportion of people reporting foreign citizenship are native-born? It may also be instructive to look at the incidence of large residual categories and missing values. Such quality appraisals must, however, contend with the issue of data editing by national statistical offices. Most samples do not have detailed information on how they were processed, but inferential evidence suggests a number of them were edited for missing values. Any sample with no missing data among the basic demographic variables such as age, sex and relationship-to-householder undoubtedly underwent some level of editing. Among these samples, most do not provide flags indicating where edits occurred.

Additional consistency-type quality checks are possible where multiple samples are available for a country (see Moultrie 2012). Figure 8.1 shows completed fertility by birth year for women in four censuses of Thailand. No attempt has been to smooth the data. For any given birth year, the figures from all the censuses should be nearly identical, net of some mortality and migration effects. The data for 1990 and 2000 are highly congruent, apart from the noisiness one sees in all the samples for the earliest birth years, corresponding to elderly respondents. The 1970 and 1980 samples exhibit a similar general trend as 1990–2000, but 1980 is roughly a half-child higher per woman than the later years, and 1970 is a half-child higher than 1980. We can conclude that at least two of these censuses misreport fertility, although it would take further investigation to determine which are at fault.

Table 8.4 Whipple's index for selected Latin American samples

Sample	Index	Category		Sample	Index	Category
Argentina 1970	104	Very accurate		Costa Rica 1973	121	Approximate
Argentina 1980	107	Fairly accurate		Costa Rica 1984	108	Fairly accurate
Argentina 1991	103	Very accurate		Costa Rica 2000	110	Fairly accurate
Argentina 2001	103	Very accurate		Ecuador 1962	176	Very rough
Bolivia 1976	145	Rough		Ecuador 1974	137	Rough
Bolivia 1992	124	Approximate		Ecuador 1982	127	Rough
Bolivia 2001	113	Approximate		Ecuador 1990	133	Rough
Brazil 1960	143	Rough		Ecuador 2001	112	Approximate
Brazil 1970	126	Rough		Mexico 1960	175	Very rough
Brazil 1980	111	Approximate		Mexico 1970	148	Rough
Brazil 1991	102	Very accurate		Mexico 1990	125	Approximate
Brazil 2000	104	Very accurate		Mexico 1995	123	Approximate
Chile 1960	131	Rough		Mexico 2000	118	Approximate
Chile 1970	123	Approximate		Mexico 2005	119	Approximate
Chile 1982	104	Very accurate		Panama 1960	126	Rough
Chile 1992	100	Very accurate		Panama 1970	121	Approximate
Chile 2002	100	Very accurate		Panama 1980	112	Approximate
Colombia 1964	144	Rough		Panama 1990	109	Fairly accurate
Colombia 1973	140	Rough		Panama 2000	103	Very accurate
Colombia 1985	139	Rough		Venezuela 1971	115	Approximate
Colombia 1993	118	Approximate		Venezuela 1981	102	Very accurate
Colombia 2005	106	Fairly accurate		Venezuela 1990	110	Fairly accurate
Costa Rica 1963	125	Rough		Venezuela 2001	103	Very accurate

Only certain variables are amenable to this technique, but such consistency checks can offer additional perspective on overall census quality.

Although more an inherent limitation than a data quality issue, per se, the geographic detail in the microdata has notable implications for certain migration applications. For practical reasons, internal migration is usually defined as movement across administrative divisions within a country. These migration-defining units vary substantially in size and population between and within countries. Migration distance implied by moves between adjacent physically large Amazonian municipalities can be quite different from migration between adjacent units within a metropolitan area. Moves that occur entirely within a physically large geographic unit will not be recorded as migration in most cases, whereas relatively short-distance moves that cross a boundary will be reported. The measurement issues can be especially acute for comparative analyses including multiple countries (see Bell and Muhidin 2011). Table 8.5 reports the median population of the smallest geographic units identified in each country's most recent microdata sample. The numbers reflect differing political geographies combined with varying degrees of geographic suppression for confidentiality. The physical expanse of the units can be calculated from GIS boundary files available for most countries' highest administrative level, but that geographic information is typically lacking for lower level units.

Research Directions

The size and scope of the IPUMS database offer unique opportunities for migration research. Much of its potential lies in comparisons: between places, over time, and between different subpopulations. The database encourages researchers to think big—to look for patterns and interrelationships that cannot readily be explored with other data sources. The following discussion describes a number of research areas

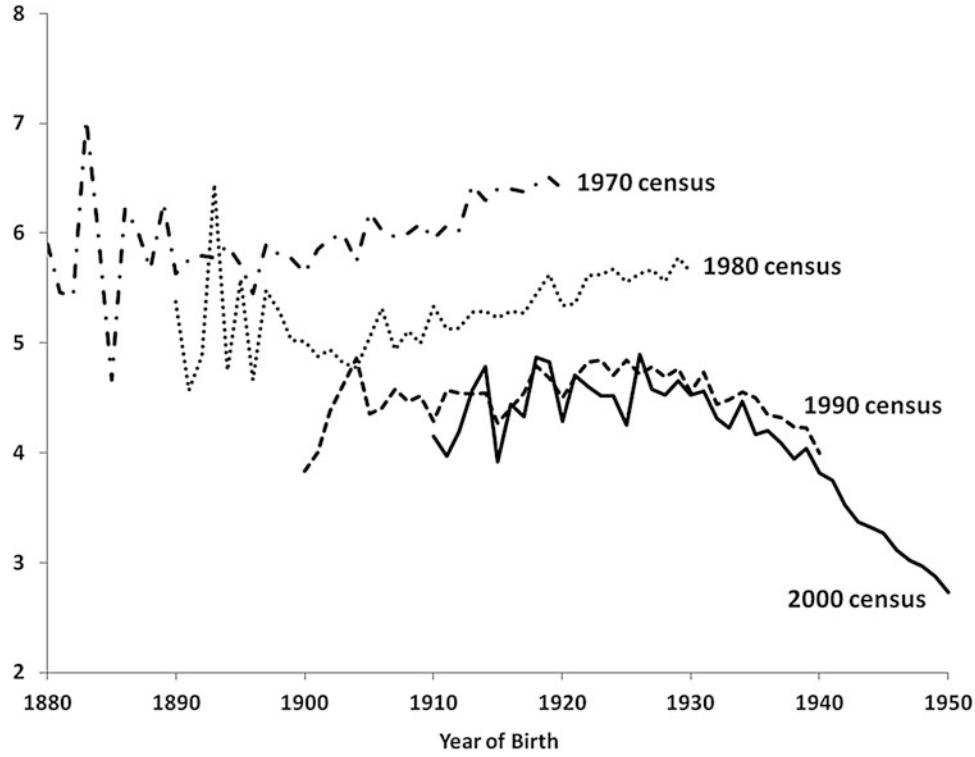

Fig. 8.1 Completed fertility by year of birth, Thailand 1970–2000
Children ever born to ever-married women age 50–90

that have attracted attention from IPUMS users and others for which the database is well suited. The international IPUMS project began in 2000, but its country coverage was fairly sparse until near the end of the decade. There are still many unexplored and underexploited aspects of the data series.

As part of the IPUMS registration process, researchers must describe how they intend to use the data. Approximately 15 % of IPUMS users identify migration or immigration as a significant component of their research. The topical distribution is difficult to quantify, but a selective listing of themes includes the following:

- Immigrant adaptation
- Socioeconomic attainment
- Migration and aging; life course
- Gender and migration; fertility
- Migration and education
- Skilled worker migration; brain drain
- Migration of children
- Forced migration
- Labor market effects

Migrants' adaptation and their status relative to non-migrants are among the areas of research for which the IPUMS data have considerable potential. Census data provide ample cases to study human capital, employment, fertility, family structure, and other characteristics. For any of these topics, migrant groups can be compared to one another and to the non-migrant population at the national or even sub-national levels. Socioeconomic status is a common basis of comparison, with educational attainment the most straightforward indicator widely available in the censuses. Despite differences in national education systems, there is considerable consistency in identifying completion of primary, secondary or tertiary schooling. Figure 8.2 shows secondary education rates for native and foreign-born adults

Table 8.5 Median population of smallest geographic unit, by country (in 000s)

Mexico	13		Uruguay	82		Senegal	253
Colombia	31		Mongolia	87		Indonesia	255
Sierra Leone	34		South Africa	96		Tanzania	301
Mali	36		Malaysia	98		Israel	332
Burkina Faso	36		Slovenia	100		Malawi	351
Nicaragua	37		South Sudan	102		Iraq	365
Brazil	41		Vietnam	110		Romania	414
Philippines	41		Puerto Rico	121		Morocco	459
Panama	43		Haiti	121		Ireland	481
Dominican Rep.	43		Zambia	126		Thailand	634
Liberia	43		United States	130		Cuba	712
Ecuador	44		Ghana	134		Rwanda	731
Spain	46		Saint Lucia	134[a]		Portugal	776
Costa Rica	47		Uganda	137		Canada	963
Ukraine	47		Armenia	142		Pakistan	1,000
Bolivia	48		Jamaica	146		Belarus	1,413
Venezuela	48		Palestine	167		Iran	1,509
El Salvador	48		Guinea	168		France	1,812
Greece	50		Sudan	174		Italy	2,114
Chile	55		Turkey	200		China	2,985
Argentina	59		Egypt	208		Germany	3,763
Fiji	61		Kenya	215		Nigeria	4,279
Cambodia	63		Austria	225		United Kingdom	5,143
Jordan	65		Switzerland	233		India	8,635
Peru	65		Nepal	240		Hungary	10,210[a]
Kyrgyzstan	70		Bangladesh	243		Netherlands	15,986[a]
Cameroon	71						

[a]No subnational units are identified
Figures refer to the most recent available sample in each country

for selected Asian countries around the year 2000. Immigrants typically have higher educational attainment, but there are exceptions and significant variation in degree. Although generally more difficult to work with, occupation data offer further opportunity to gauge migrant socioeconomic success. One can consider migrants' relative educational attainment within occupations (Thomas 2010) or even convert detailed occupations into a continuous measure of socioeconomic status to make broad cross-national comparisons (Spörlein and van Tubergen 2014).

The grouping of individuals into households in the census microdata enables analysis of migrant living arrangements. Studies can compare household structures of immigrant populations to non-migrants in their origin country (Van Hook and Glick 2007), or evaluate living arrangements of specific diaspora streams in two or more destination countries (Burr et al. 2012). The IPUMS pointer variables identify each person's co-resident spouse, facilitating study of migration effects on marriage patterns. By comparing the individual attributes of spouses, one can assess the propensity of migrants to marry within their respective socioeconomic group, such as their education stratum (Choi and Mare 2012), or to form unions with persons of other ethnicities (Qian et al. 2012). Figure 8.3 shows the proportion of married foreign-born people in Europe and the United States in a union with a native-born person. The data reveal marked differences in endogamy between countries, with intermarriage in the U.S. lower than in Europe. The data would

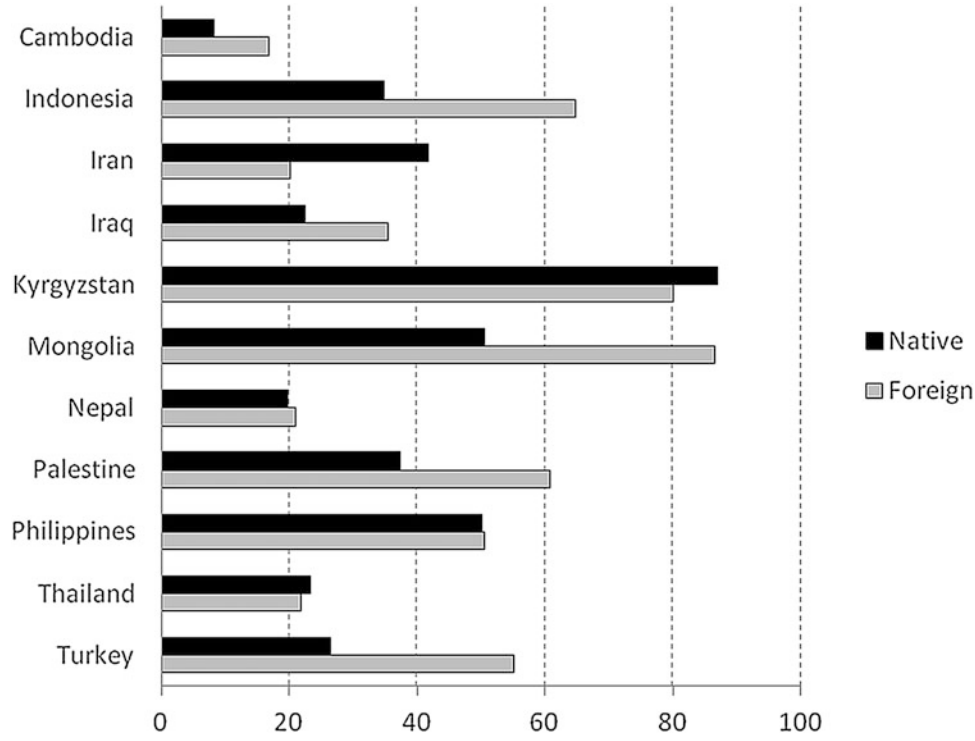

Fig. 8.2 Secondary education by nativity status, selected Asian countries (%)
Persons age 20–59

support further exploration of potential cohort effects and variation among immigrant groups both within and across countries. The census data also have considerable potential for research on migration and the life course, including elderly and retirement migration (Bernard et al. 2014; Bradley and Longino 2009).

The interplay of migration and gender is one of the most active areas of research using IPUMS. The changing sex composition of international migration flows can be explored at a multi-national scale across decades (Donato 2010). Where samples are available for both sending and receiving countries, the selectivity of migrants with respect to various criteria can be analyzed with respect to gender (Feliciano 2008). Skilled worker migration is among the many phenomena that have a distinct gender dimension (Docquier et al. 2009). Another perspective on gendered migration concerns the demographic and economic effects on the sending country, with altered sex ratios potentially distorting marriage and labor markets (Raphael 2013; White and Potter 2013).

The effect of migration on children and youth is another topic that has attracted considerable attention from researchers. The effect is usually measured in terms of school attendance or employment of migrant children (Rendall and Torr 2008). Schooling is not difficult to measure with the census, but child employment can be problematic due to differing minimum ages for reporting work and the degree to which unpaid family labor may go unreported. Rather than the migrants themselves, one can focus on the children left behind when family members seek work abroad (Halpern-Manners 2011). The characteristics of the receiving area with respect to public services, such as the prevalence of housing units with electricity and sewage, can be a factor affecting the propensity of families to migrate with children (Archambault et al. 2012). Children migrating without relatives are another topic that can be explored with the

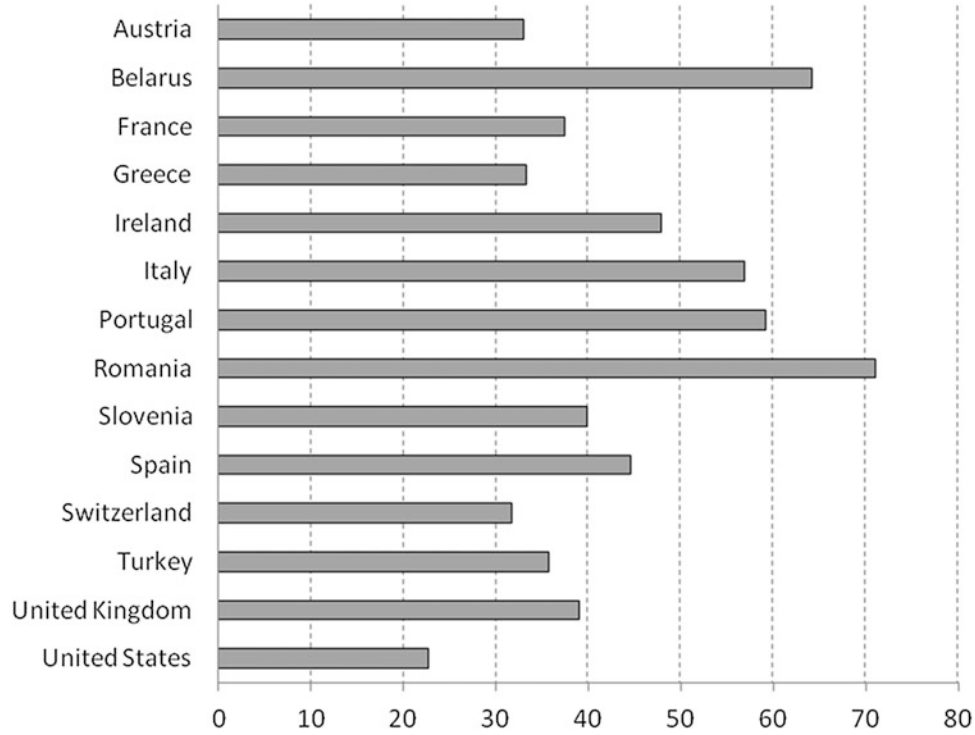

Fig. 8.3 Foreign-born persons married to natives: Europe and United States (%)
Persons in a marriage or consensual union with a co-resident spouse
Data are from the most recent sample for each country

census microdata, owing to large sample sizes and detailed relationship information on co-resident persons (Yaqub 2009). Figure 8.4 gives school attendance rates for internal migrant children by presence of relatives across developing countries on three continents. In nearly all countries, children residing without relatives are significantly disadvantaged. The ease with which the IPUMS data enables international comparisons is reflected in some large-scale profiles of child migrants (Barker et al. 2013; McKenzie 2008).

Census data are well suited to studying the relationship between education and migration. The brain drain—the flow of highly educated persons from developing to developed countries—is one of the more popular topics indicated by researchers applying for access to IPUMS. The scope of the database allows globe spanning studies (Dumont et al. 2010), and its temporal depth offers the opportunity to explore the historical trajectory of skilled worker migration (Docquier and Marfouk 2006). The data support studying internal skilled migration as well as international (Clemens 2009). Work variables offer another perspective on skilled migration and provide the opportunity to explore potential education-occupation mismatches among migrants. Employment status and industry provide further perspective on migrant outcomes relative to educational attainment. Education can also be considered within the broader context of factors affecting the propensity to migrate (Aguayo-Téllez and Martínez-Navarro 2013). By pairing data from two countries, one can consider the educational attainment of migrants in relation to the non-migrant population they left behind (Feliciano 2005). Figure 8.5 shows the proportion of Brazilian-born adults residing in various destination countries who have completed secondary education. The data suggest distinct

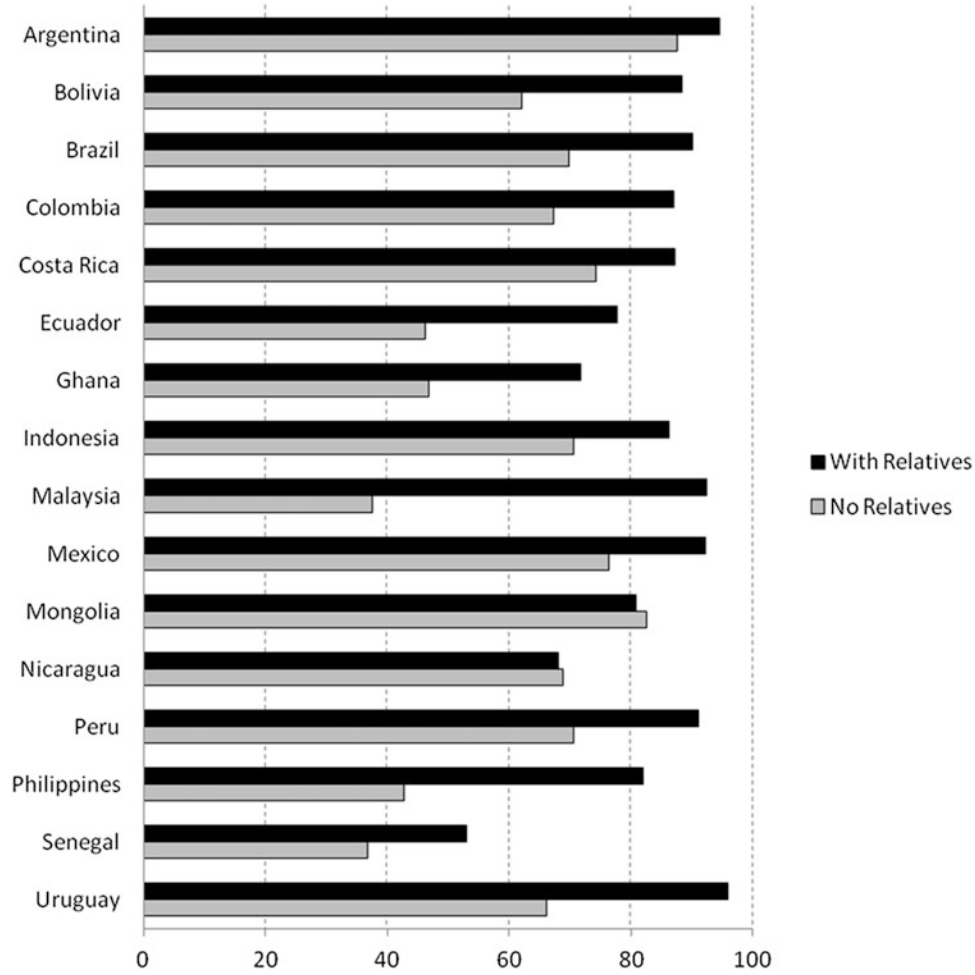

Fig. 8.4 School attendance of migrant children, by presence of relatives in the household (%)
Children age 6–15 who migrated internally within the past 5 years. Samples circa 2000

migration streams exhibiting limited selectivity with respect to distance, with internal Brazilian lifetime migrants being only modestly more educated than non-migrants. In most cases one could extend such comparisons over time using multiple censuses.

Return migration can be identified using birthplace and prior residence information, which are broadly available in the IPUMS samples. There is scope for examining the determinants of return migration, although at the individual level one cannot know the type of work people were performing in their old location (Medina and Posso 2013). The socioeconomic attainment of return migrants relative to non-movers can be assessed through their education, employment and housing characteristics (Thomas 2008, 2009; Thomas and Inkpen 2013).

Fourteen IPUMS samples, mostly from Latin America, indicate the number of household members who migrated internationally in some specified period prior to the census. Half of those samples provide individual records for each migrant, which can be linked to data extracts. As mentioned above, these records allow finer analysis of the characteristics of migrants, their reasons for migrating, and the structure and status of the sending households. To this point, the migrant records are a relatively underutilized aspect of the IPUMS data series. Unfortunately,

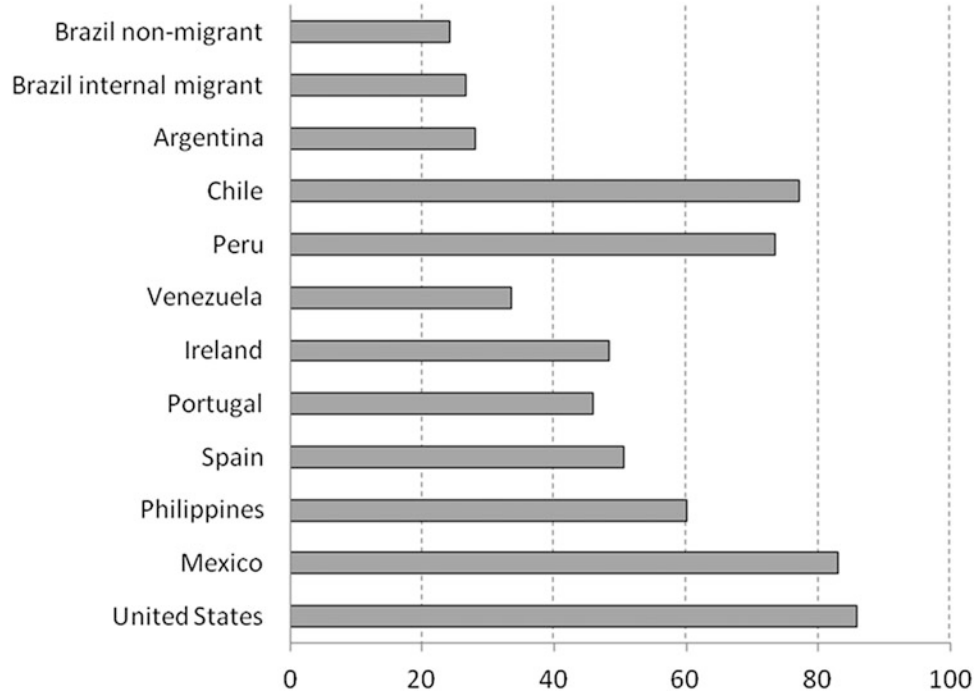

Fig. 8.5 Brazilians at home and abroad: secondary education by country of residence (%)
Persons born in Brazil, age 18+
Data are from census closest to year 2000 for each country

despite its importance in many developing countries, the receipt of remittances by a household is reported in only a few samples in Africa and Latin America.

There is some scope for using the census microdata to analyze the effect of migration on labor markets in both source and destination areas (Aydemir and Borjas 2007; Mishra 2007). This is most straightforward in the limited instances—such as Mexico and the United States—where income data are available in the census. The earnings data are also useful for analyzing push and pull factors from the perspective of potential migrants (Davila et al. 2009).

The census data can be used to develop multilevel models exploring the effect of locality on migration determinants and outcomes (Loebach and Korinek 2012; Spörlein and van Tubergen 2014). Contextual information can be calculated from the census: for example, the concentration of certain industries, housing opportunities, or immigrant groups might all be tabulated at the province or district level and used with individual-level variables to analyze migrant behavior. By bringing in outside data, the multi-level approach has been applied as far afield as assessing rainfall effects on propensity to migrate (Nawrotzki et al. 2013). In combination with other sources, IPUMS has also been used for health-related research, such as exploring the connection between migration and malaria transmission in Africa (Pindolia et al. 2013, 2014). Any merging of data sources, however, depends ultimately on harmonizing their geographies to IPUMS, which can sometimes be challenging. A new data project at the Minnesota Population Center, Terra Populus, promises to significantly reduce the barriers to combining IPUMS with environmental data in the future (Minnesota Population Center 2013).

The IPUMS data also offer great potential for overtly spatial analysis. Using GIS boundary files, researchers can calculate migration distance and direction, population density and other measures. The GIS files provided by IPUMS define areas corresponding to geography variables in the microdata—typically political units. They are not

point data. Boundary files are available at the country and the first administrative levels (e.g., states or provinces) within countries. Boundaries for the second administrative level will be added as they are processed in the future.

The majority of IPUMS migration research, including the majority of studies referenced above, is oriented to international moves. But most of the aforementioned topics can be explored in terms of internal migration. The IMAGE project (Internal Migration Around the Globe) is an ambitious effort to investigate internal migration cross-nationally at a global scale. The project makes extensive use of the IPUMS samples in its efforts to inventory the world's data and develop consistent measures of internal migration, among other goals (Bell and Muhidin 2009; Bell and Charles-Edwards 2013; Bernard et al. 2014).

The foregoing is a selective list of potential applications of the IPUMS data to migration research. IPUMS continues to expand in geographic coverage and temporal depth. This growth means new research possibilities for the database are continually arising. But even heavily mined fields of study can yield new insights using novel approaches, making fresh comparisons, or combining the census data with evidence from other sources.

References

Aguayo-Téllez, E., & Martínez-Navarro, J. (2013). Internal and international migration in Mexico: 1995–2000. *Applied Economics, 45*, 1647–1661.

Archambault, C. S., de Laat, J., & Zulu, E. M. (2012). Urban services and child migration to the slums of Nairobi. *World Development, 40*, 1854–1869.

Aydemir, A., & Borjas, G. (2007). Cross-country variation in the impact of international migration: Canada, Mexico, and the United States. *Journal of the European Economic Association, 5*, 663–708.

Barker, K. M., Temin, M., Engebretsen, S., & Montgomery, M. R. (2013). *Girls on the move: Adolescent girls & migration in the developing world. A girls count report on adolescent girls*. New York: Population Council.

Bell, M., & Charles-Edwards, E. (2013). *Cross-national comparisons of internal migration: an update on global patterns and trends* (Technical Paper No. 2013/1). New York: UN Department of Economic and Social Affairs, Population Division.

Bell, M., & Muhidin, S. (2009). *Cross-national comparison of internal migration* (Human Development Research Paper 2009/30). New York: UN Department of Economic and Social Affairs, Population Division.

Bell, M., & Muhidin, S. (2011). Comparing internal migration between countries using Courgeaus K. In J. Stillwell & M. Clarke (Eds.), *Population dynamics and projection methods: Understanding population trends and processes – Volume* (pp. 141–164). Dordrecht: Springer.

Bernard, A., Bell, M., & Charles-Edwards, E. (2014). Life-course transitions and the age profile of internal migration. *Population and Development Review, 40*, 213–239.

Bradley, D., & Longino, C. (2009). Geographic mobility and aging in place. In P. Uhlenberg (Ed.), *International handbook of population aging* (pp. 319–339). Netherlands: Springer.

Burr, J. A., Lowenstein, A., Tavares, J. L., Coyle, C., Mutchler, J. E., Katz, R., & Khatutsky, G. (2012). The living arrangements of older immigrants from the former Soviet Union: A comparison of Israel and the United States. *Journal of Aging Studies, 26*, 401–409.

Choi, K. H., & Mare, R. D. (2012). International migration and educational assortative mating in Mexico and the United States. *Demography, 49*, 449–476.

Clemens, M. A. (2009). *Skill flow: A fundamental reconsideration of skilled worker mobility and development* (Center for Global Development, Working Paper 180). Washington DC: Center for Global Development.

Cleveland, L., Davern, M., & Ruggles, S. (2011). *Drawing statistical inferences from international census data* (Minnesota Population Center, Working Paper 2011–01). Minneapolis: University of Minnesota.

Davila, A., Mora, T. T., & Hales, A. D. (2009). Earned income along the U.S.-Mexico border. In M. T. Mora & A. Davila (Eds.), *Labor market issues along the U.S.-Mexican border* (pp. 107–120). Tuscon: University of Arizona Press.

Docquier, F., & Marfouk, A. (2006). International migration by education attainment, 1990–2000. In C. Ozden & M. Schiff (Eds.), *International migration, remittances and the brain drain* (pp. 151–199). New York: Palgrave and Macmillan.

Docquier, F., Lindsay Lowell, B., & Marfouk, A. (2009). A gendered assessment of highly skilled emigration. *Population and Development Review, 35*, 297–321.

Donato, K. M. (2010). U.S. migration from Latin America: Gendered patterns and shifts. *Annals of the American Academy of Political and Social Science, 630*, 78–92.

Dumont, J.-C., Spielvogel, G., & Widmaier, S. (2010). *International migrants in developed, emerging and developing countries: An extended profile* (Social, Employment and Migration Working Paper 114). Paris: OECD.

Esteve, A., & Sobek, M. (2003). Challenges and methods of international census harmonization. *Historical Methods, 36*, 66–79.

Feliciano, C. (2005). Educational selectivity in U.S. Immigration: How do immigrants compare to those left behind? *Demography, 42*, 131–152.

Feliciano, C. (2008). Gendered selectivity: U.S. Mexican immigrants and Mexican nonmigrants, 1960–2000. *Latin American Research Review, 43*, 139–160.

Filmer, D., & Pritchett, L. (2001). Estimating wealth effects without expenditure data—or tears: An application to educational enrollments in states of India. *Demography, 38*, 115–132.

Ganzeboom, H., & Treiman, D. (1996). Internationally comparable measures of occupational status. *Social Science Research, 25*, 201–239.

Halpern-Manners, A. (2011). The effect of family member migration on education and work among nonmigrant youth in Mexico. *Demography, 48*, 73–99.

Loebach, P., & Korinek, K. (2012). Crossing borders, crossing seas: The Philippines, gender and the bounding of cumulative causation. *International Migration, 50*, 1–15.

McKenzie, D. J. (2008). A profile of the world's young developing country international migrants. *Population and Development Review, 34*, 115–135.

Medina, C., & Posso, C. (2013). South American immigrants in the United States of America: Education levels, tasks performed and the decision to go back home. *Journal of Economic Studies, 40*, 255–279.

Minnesota Population Center. (2013). *Terra Populus: Beta version [machine-readable database]*. Minneapolis: University of Minnesota.

Minnesota Population Center. (2014). *Integrated Public Use Microdata Series, International: Version 6.3* [Machine-readable database]. Minneapolis: University of Minnesota.

Mishra, P. (2007). Emigration and wages in source countries: Evidence from Mexico. *Journal of Development Economics, 82*, 180–199.

Moultrie, T. A. (2012). General assessment of age and sex data. In A. Hill, K. Hill, I. Timaeus, B. Zaba (Eds.), *Tools for demographic estimation* (http://demographicestimation.iussp.org). New York: UNFPA.

Nawrotzki, R. J., Riosmena, F., & Hunter, L. M. (2013). Do rainfall deficits predict U.S.-bound migration from rural Mexico? Evidence from the Mexican census. *Population Research Policy Review, 32*, 129–158.

Pindolia, D. K., Garcia, A. J., Huang, Z., Smith, D. L., Alegana, V. A., Noor, A. M., Snow, R. W., & Tatem, A. J. (2013). The demographics of human and malaria movement and migration patterns in east Africa. *Malaria Journal, 12*, 397.

Pindolia, D. K., Gracia, A. J., Huang, Z., Fik, T., Smith, D. L., & Tatem, A. J. (2014). Quantifying cross-border movements and migrations for guiding the strategic planning of malaria control and elimination. *Malaria Journal, 13*, 169.

Qian, Z., Glick, J. E., & Batson, C. D. (2012). Crossing boundaries: Nativity, ethnicity, and mate selection. *Demography, 49*, 651–675.

Raphael, S. (2013). International migration, Sex ratios, and the socioeconomic outcomes of nonmigrant Mexican women. *Demography, 50*, 971–991.

Rendall, M. S., & Torr, B. M. (2008). Emigration and schooling among second-generation Mexican-american children. *International Migration Review, 42*, 729–738.

Ruggles, S. (2014). Big microdata for population research. *Demography, 51*, 287–297.

Ruggles, S., King, M., Levison, D., McCaa, R., & Sobek, M. (2003). IPUMS-international. *Historical Methods, 36*, 60–65.

Rutstein, S. O., & Staveteig, S. (2014). *Making the demographic and health surveys wealth index comparable* (DHS Methodological Reports No. 9). Rockville, Maryland: ICF International.

Sobek, M., & Kennedy, S. (2009). *The development of family interrelationship variables for international census data* (Minnesota Population Center, Working Paper 2009–02). Minneapolis: University of Minnesota.

Sobek, M., Cleveland, L., Flood, S., Hall, P. K., King, M. L., Ruggles, S., & Schroeder, M. (2011). Big data: Large-scale historical infrastructure from the Minnesota Population Center. *Historical Methods, 44*, 61–68.

Spörlein, C., & van Tubergen, F. (2014). The occupational status of immigrants in western and non-western societies. *International Journal of Comparative Sociology, 55*, 119–143.

Thomas, K. A. (2008). Return migration in Africa and the relationship between educational attainment and labor market success: Evidence from Uganda. *International Migration Review, 42*, 652–674.

Thomas, K. A. (2009). The human capital characteristics and household living standards of returning international migrants in eastern and southern Africa. *International Migration, 50*, 85–106.

Thomas, K. A. (2010). Racial and ethnic disparities in education-occupation mismatch status among immigrants in South Africa and the United States. *Journal of International Migration and Integrations, 11*, 383–401.

Thomas, K., & Inkpen, C. (2013). Migration dynamics, entrepreneurship, and African development: Lessons from Malawi. *International Migration Review, 47*, 844–873.

Van Hook, J., & Glick, J. E. (2007). Immigration and living arrangements: Moving beyond economic need versus acculturation. *Demography, 44*, 225–249.

White, K., & Potter, J. E. (2013). The impact of outmigration of men on fertility and marriage in the migrant-sending states of Mexico, 1995–2000. *Population Studies, 67*, 83–95.

Yaqub, S. (2009). *Child migrants with and without parents: Census-based estimates of scale and characteristics in Argentina, Chile and South Africa* (Innocenti Discussion Papers 2009-02). New York: UNICEF.