Data Integration at the U.S. Census Bureau: Recent Initiatives to Improve Estimates of International Migration

Jason Schachter
Chief, Net International Migration Branch
Population Division
U.S. Census Bureau

International Forum on Migration Statistics
January 16, 2018

Net International Migration (NIM) Estimates

- Produce annual estimates of international migration flows to and from the United States
  - National, state, and county
    - By age, sex, race/Hispanic origin
- NIM components
  - Foreign-Born Immigration
  - Foreign-Born Emigration
  - Net Native Migration
  - Puerto Rico flows to/from U.S.
  - Net Military Movement

Challenges for improving international migration estimates

- The American Community Survey (ACS) is our primary data source
  - Annual survey of about 2 million households
- Sample size limitations for characteristics, particularly at the subnational (county) level
  - Pool multiple years of data (impacts recency of estimates)
- Increased concern about non-response
  - Legal status not collected
    - Refugees, Irregular migrants, etc.
- Data quality
  - Ex: Year of Entry: heaping, first or most recent move, etc.
- Recent events impacting migration (e.g. Hurricane Maria and Puerto Rico)

U.S. Census Bureau experience

- U.S. Census Bureau pushing to better leverage pre-existing administrative data, but so far not well utilized for international migration statistics
  - No Population Register
  - Decentralized Federal Statistical System
    - Census Bureau, Department of Homeland Security, State Department, Department of Justice, Health and Human Services, etc.

Developing work to address challenges

- Data Integration
  - Linking Census to Administrative Data
    - Demographic Characteristics File (DCF) combines Census, IRS, and Social Security Administration (SSA) data
    - State Department (refugee data)
- Modeling
  - NIM subnational geographic distribution
    - Combine ACS (larger geographies) and administrative data (smaller geographies)
    - IRS data on US citizens abroad
  - Assign refugee status to ACS from administrative data (data linkage or probabilistic/stochastic matching)

Data Linking

- The Census Numerical Identification File (NUMIDENT) is a dataset of unduplicated SSA records, containing one record for every person ever issued an Social Security Number (SSN)
- Process to match persons across Census surveys and Federal data
  - Person Identification Validation System (PVS)
  - Probabilistic matching to match person data from an incoming file to a reference file (derived from the NUMIDENT)
    - Name
    - Date of birth
    - Address
- Each matched person record is assigned a Protected Identification Key (PIK)
  - Unique identifier for each individual
  - Ensures confidentiality
  - Person linkage key
- PIKs link individual records between datasets
  - IRS, Decennial Census, and the NUMIDENT linked to create DCF

DCF

NUMIDENT	IRS	Census 2010
• Age • Sex • Country of Birth	• Geography	• Race • Hispanic origin *Race and Hispanic origin are imputed for the foreign born who entered after 2010.

DCF Country of Birth Information

- "Country of Birth" (COB) collected from SSA (included on NUMIDENT)
- Two steps to edit and clean COB:
  1. Maximize record count for each country by combining records
     - Ex: Vietnam (VM), North Vietnam (VN), and South Vietnam (VS) recoded to Vietnam (VM)
  2. Correct erroneous country codes
     - Ex: Records coded to China (CH), that list city of birth as Santiago, recoded to Chile (CI)

DCF usage

- Used to assign missing race and Hispanic origin to post-2010 Census entries, which includes new births and immigrants
  - Use country of birth as part of the race imputation process
- Examined secondary domestic migration patterns of the foreign born within the United States
  - Paper presented at 2016 conference
- Other projects currently under development at the Census Bureau

Application: NIM subnational distribution methodology

- Current methodology distributes most NIM components based on recent foreign-born stock population (5-year ACS file)
  - Lacks recency and accuracy of characteristics at the county-level
- Possibilities being researched
  - DCF to estimate national characteristics and subnational totals and characteristics (completely replace ACS)
  - DCF to estimate totals and characteristics of counties below a certain population threshold, while continuing to use ACS for larger counties and all states (combine DCF and ACS)
    - DCF to estimate county totals and characteristics controlled to state totals measured by ACS
    - DCF to estimate county totals and characteristics, plus some states with smaller international migrant populations
  - Keep current methodology, but model county age distribution (or other characteristics) based on DCF results

DCF Research Questions

- Evaluation of DCF foreign-born coverage
  - Missing recent/irregular migrants?
  - How does DCF compare to ACS?
  - Adjust DCF to account for “new arrivals”?
- Do we still have to use a DCF proxy universe (stocks) or can we estimate flows directly?
- Evaluation of ITIN tax data (tax IDs given to those ineligible for Social Security numbers)
  - Can we include them on the DCF (currently on NUMIDENT)?
  - How much would this improve coverage?

IRS Tax Records

- U.S. citizens required to file Federal tax returns when earning income while residing outside the United States
- Information on “address outside the United States” is included on the NUMIDENT, but not used in our processes
  - US citizens living abroad (stock)
  - Current address abroad and previous address in the U.S. (emigration flow)
  - Current address in the US and previous address abroad (immigration flow)
- Could use information on migration flows to/from abroad to distribute our subnational net native migration estimates
- Questions about coverage and accuracy of address filing outside the U.S.

Assigning refugee status to the ACS

- ACS does not collect information on immigrant or legal status, thus no direct method to measure refugees
- Use State Department data on number of refugees resettled in the United States as part of our foreign-born immigration estimates
  - Would need to subtract refugees from the ACS to avoid double counting.
    - Direct matching method (link to ACS by age, date of birth, address)
    - Indirect probabilistic matching method (use information about refugee characteristics to impute onto ACS)

Considerations when using administrative data to improve international migration statistics

- Data sharing mechanisms need to be in place
  - Communication/Cooperation between agencies is critical
- Data protection
- Data linking procedures
  - Deterministic or probabilistic
- Data quality issues
  - Data cleaning necessary
- Coverage (under and over)
- Operational/Definitional
  - How migrants are defined across data sources
    - Duration of stay, usual residence, actual vs intended stay, temporary vs. long-term
- Possibilities for longitudinal analysis
- Emigration still problematic to measure

General questions for discussion

- How to make different integrated data sources compatible?
- How to improve coordination between national agencies?
  - Legal mandates needed?
- Administrative Data=Big Data?
  - How to leverage?
    - Remote sensing, social media, mobile phones, financial transactions, postal codes, etc.
  - Examples relate to tourism/commuting
  - Possible to apply “change of usual residence” definition?
    - New definitional paradigm needed?

Thank you!

Jason Schachter

Chief, Net International Migration Branch

Population Division

U.S. Census Bureau

Jason.P.Schachter@census.gov