Data Integration for Disaggregated Statistics on International Migration: UN Task Force 3 Report

Jason Schachter, External Expert

Task Force Co-Chair

Technical Workshop on International Migration and Temporary Mobility Statistics

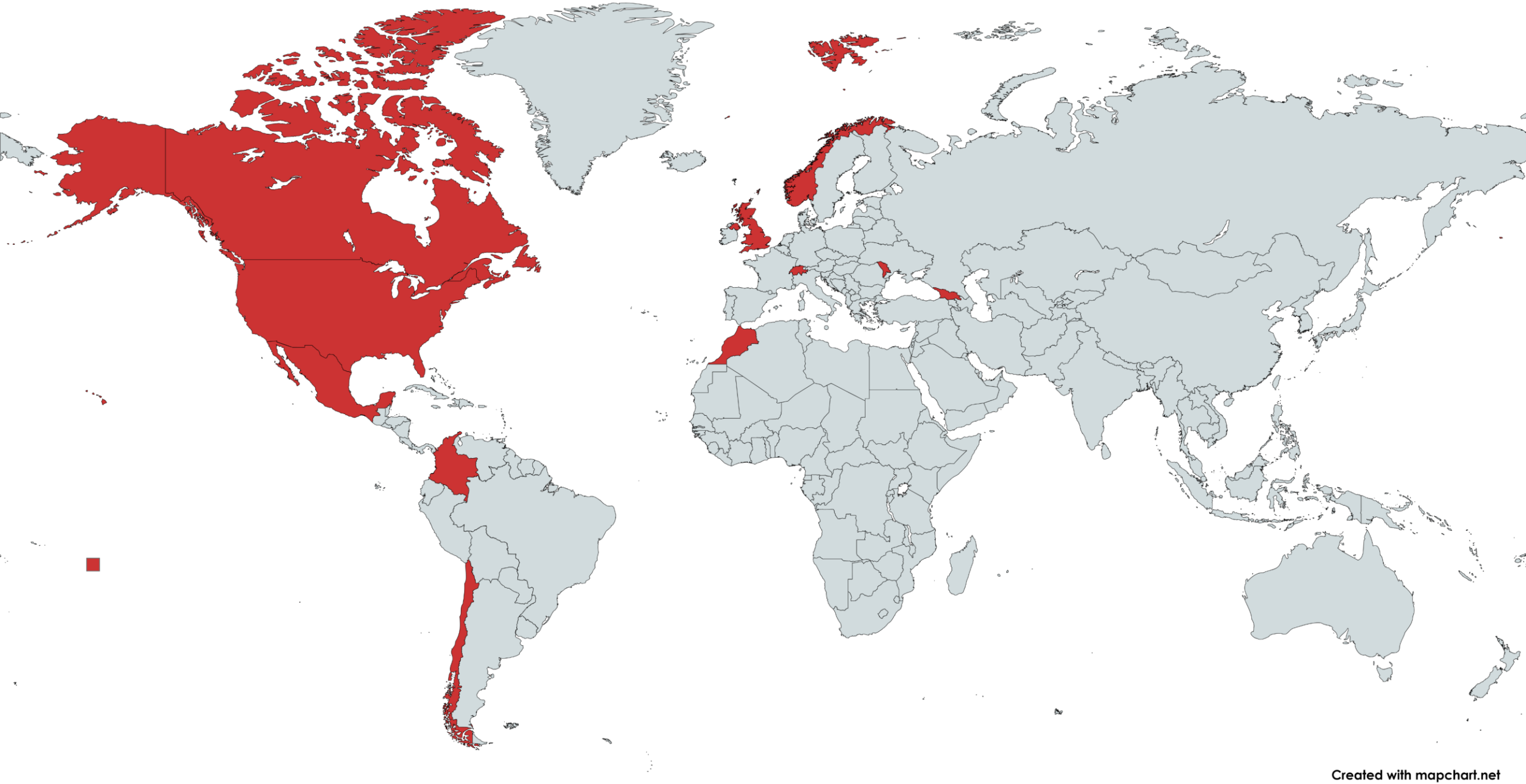
17 July 2024, Warsaw, Poland

Background

- Expert Group on Migration Statistics: Task Force 3
- Contribution to the revision of the 1998 United Nations Recommendations on Statistics of International Migration
- Technical report that supports countries to produce sufficiently disaggregated data for the measurement of international migration by means of integration of micro- and macro-data techniques
- The report covers:
  - (a) variety of methodologies to integrate and analyze migration data on the macro and micro level
  - (b) legal, policy and technical prerequisites
  - (c) major challenges and opportunities in mainstreaming the use of integrated data in official statistics
  - Includes several illustrative case studies

Task Force Membership

- Canada
- Chile
- Colombia
- Georgia (former co-chair)
- Mexico
- Morocco
- Norway
- Republic of Moldova
- Switzerland
- United Kingdom
- United States (co-chair)
- Academia: Australian National University
- International Organizations: Eurostat, ILO, UNSD (co-chair), and WHO



Created with mapchart.net

Chapter 1: Introduction

- Data integration within the context of the new conceptual framework on international migration statistics
  - Need rich data to measure the magnitude/characteristics of migrant populations and changes over time
  - Combining multiple data sources can provide more timely, accurate, and granular data than relying on a single data source
  - Less costly than new data collection; lower operational cost and respondent burden
  - Definitions of data integration
    - Micro-data integration creates new combined datasets, which can produce statistics
    - Macro-data integration creates new statistics from outputs from multiple datasets

Chapter 1, continued

- How data integration can improve migration statistics
  - Coverage, missing characteristics, timeliness (measurement lag), lower levels of geography, hard-to-count populations (e.g. unauthorized), etc.
  - Data quality (“triangulation”)
  - Incorporate new data sources
  - Longitudinal analysis
- Use of outputs derived from integrated data in official statistics
  - More complete counts
  - Challenges: data comparability, use in official statistics, technical expertise

Chapter 2: Macro-Data Integration Methods

- Overview
  - Strategy for combining data from various sources: make missing migration statistics available or provide better coverage and more robust estimates of the target population.
  - Typically, when micro-data integration not possible
- Diverse methods
  - Compilation
    - Pull from various existing subcomponents
  - Statistical adjustment
    - Adjust numbers based on strongly correlated input data
      - For example, trends in one data set inform another (flight, visa, border, etc.)

Chapter 2, continued

- Statistical modeling
  - Information from one dataset used to supplement/enhance data missing or of low quality in a second dataset
  - Shared characteristics in both data sets can be used to estimate characteristics missing from one of the datasets
- Other
  - Hybrid (macro- and micro-)
  - New Methods (e.g. “big data”)
- Challenges for macro-data integration
  - Internal consistency (differences between integrated data sources)
  - Estimate validation

Chapter 3: Micro-Data Integration Methods

- Overview
  - Linking individual data at the record level
    - Lower cost, respondent burden; Better timeliness, geographic coverage. longitudinal.
  - Some countries have been doing this for a long time (e.g. population registers), others more recently
- Creating/enabling the legal framework
  - Transfer of personally identifiable information/data-sharing agreements
  - Data confidentiality
  - Dictates use and retention of linked data
- Creating/enabling the technical framework
  - Financial/technical resources, “Spine” selection/creation, selected variables/events, data updates, data validation, etc.

Chapter 3, continued

- Micro-data integration methodology
  - Deterministic ("exact") vs. probabilistic matching (create unique identifier based on likelihood the same person)
    - Hybrid linkage approaches common
  - Statistical matching for imputation of missing characteristics (synthetic linkages rarely used to this point)
  - Post-linkage evaluation/cleaning of matches (co-mingled data (multiple people combined into one person), multiple records for same person, etc.)
- Challenges for micro-data integration
  - Data access, data privacy, sufficient matching information (PII), unequal coverage, data quality, etc.

Chapter 4: Assessing and Communicating Results

- Overview
  - Often more of a concern for macro-data integration
  - Can combine multiple sources of error (both macro and micro)
- Estimate assessment/validation
  - True value (Census results, Demographic Analysis)
  - Alternative data sources (benchmark data, expert opinion)
  - Model-based error estimation (confidence intervals, Bayesian methods, etc.)

Chapter 4, continued

- Communication with key stakeholders and dissemination of integrated data
  - Builds trust and relevancy of integrated data and resulting statistics
    - Data providers
    - External and internal data users
    - Policy makers
    - General public
- Use of outputs derived from integrated data in official statistics
  - New guidelines and conceptual framework on international migration and mobility

Chapter 5: Conclusions and Future Work

- Data integration in the context of the new guidelines on international migration statistics
  - Improve access to and integration of administrative data sources to produce international migration statistics
  - Develop international data exchanges and collaborations
  - Combine micro- and macro-integration
  - Provide better understanding of new data sources, uses, and limitations

Glossary

- Two-page glossary defining about 30 terms
  - “Big Data” to “Synthetic Data”
  - *Macro-data integration*: methods to produce international migration statistics via the integration of aggregated data from multiple sources
  - *Micro-data integration* combines two or more datasets to create a new combined dataset which can produce international migration statistics

Appendix: Country Case Studies

- Country case studies summarize information from relevant countries' experiences
- Each example of data integration is organized around six main themes:
  - (1) Data needs
  - (2) Collaboration framework for data integration
  - (3) Data sources
  - (4) Integration methodology
  - (5) Data evaluation strategy
  - (6) Dissemination/communication of results
- Country examples: Canada, Chile, Georgia, Mexico, Moldova, Switzerland, United Kingdom, and the United States

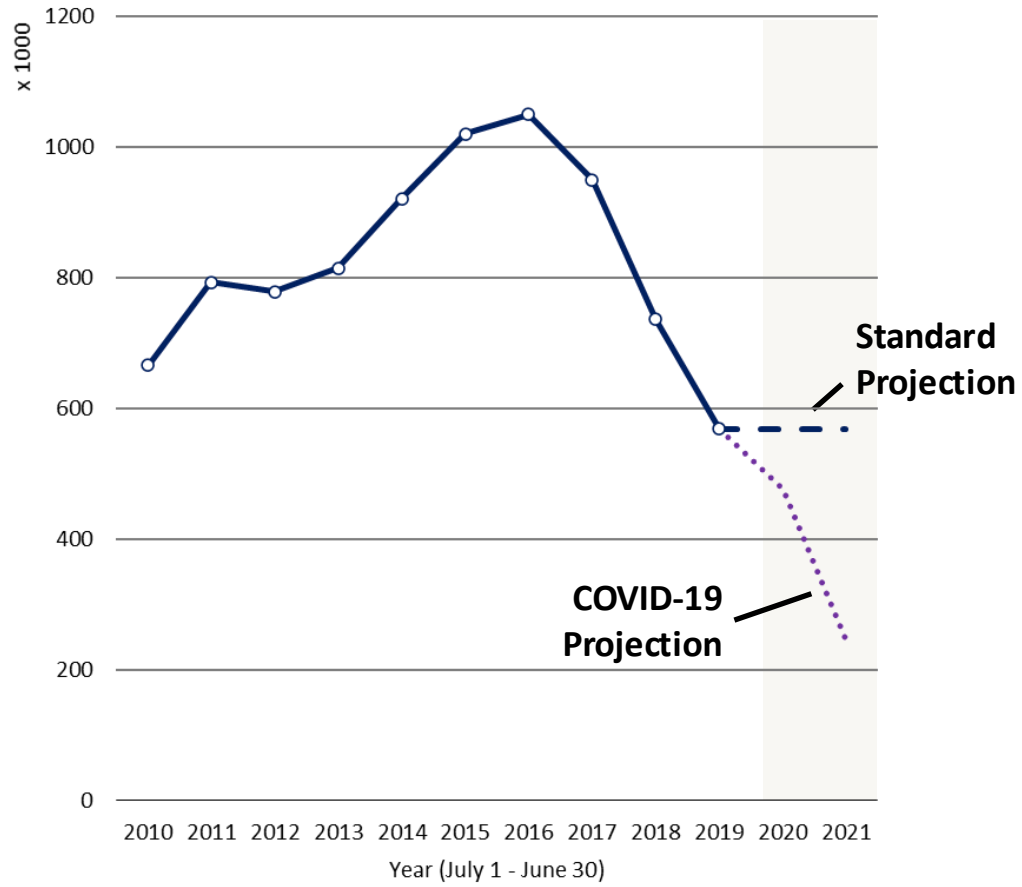
Next Steps

- Report was finalized in late 2023
- Released on UNSD website in early 2024
  - Report: https://unstats.un.org/unsd/demographic-social/sconcerns/migration/task-forces/Task%20Force%203%20Report-Web.pdf
  - Country case-studies: https://unstats.un.org/unsd/demographic-social/sconcerns/migration/task-forces/TF%203%20-%20Appedix_case_studies.pdf
- Next: incorporate into new recommendations

Example of Macro-Data Integration from the United States: Adjustments for COVID-19

- COVID-19 pandemic had a dramatic impact on immigration levels to the United States from March 2020 to the end of 2021
- Would have used 2020 American Community Survey (ACS) to produce Vintage 2021 net international migration (NIM) estimates
  - Data quality concerns (non-response bias, high imputation)
  - Data lag (July 1, 2021 estimate using calendar year 2020 data)
- 2020 ACS produced higher levels of immigration than expected
- Did not use the 2020 ACS, but rather adjusted 2019 ACS data based on trends seen in administrative data between 2019 and 2021

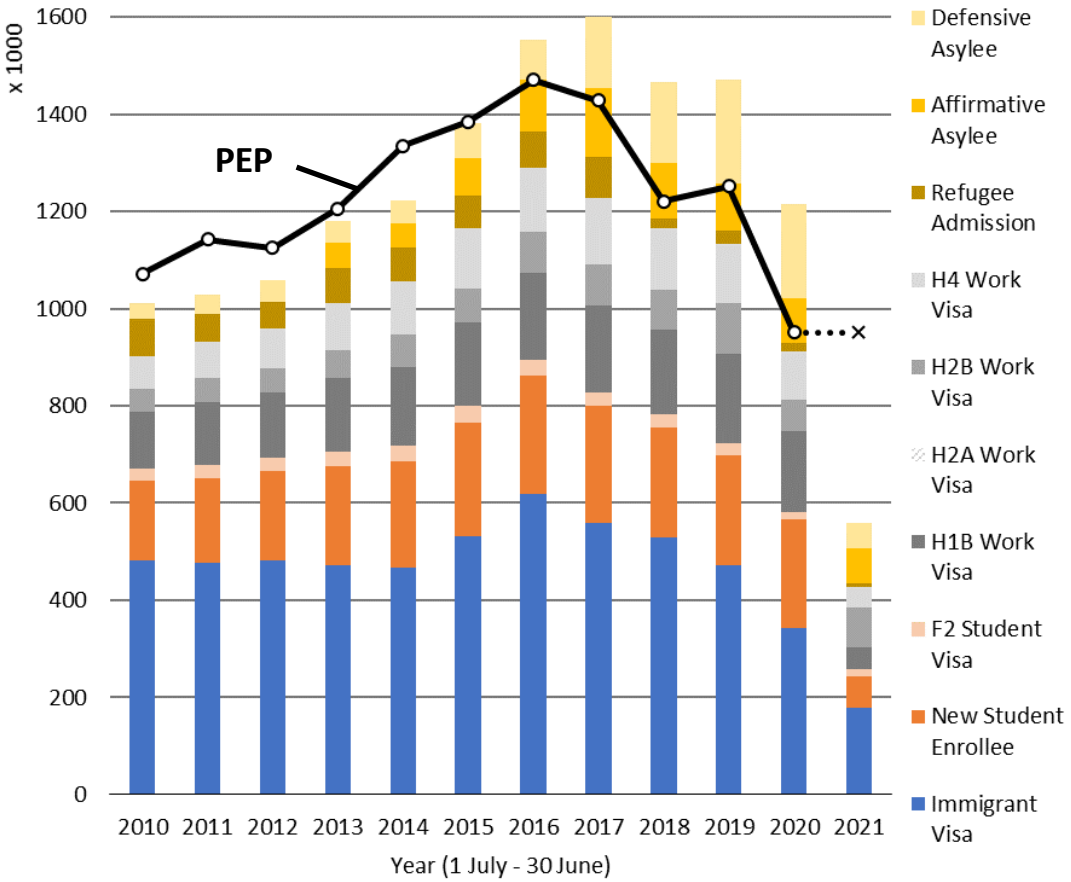
NIM Estimates



Data Sources: U.S. Census Bureau, Vintage 2020 & 2021 Population Estimates-Net International Migration. Internal File

- 2019 was the last year in which non-adjusted ACS data were used for PEP
- Standard projection method assumes that 2019 patterns continue for 2020
- 2020 COVID-19 projection based on public data and special tabulations from BTS, OIS, and Statistics Canada
- **2021 COVID-19 projection is the ratio of 2021 to 2019 benchmark totals applied to the 2019 PEP total to project 2021 PEP**
- Assumed emigration patterns followed similar patterns (confirmed by data from other countries)
- Net air passenger totals, in conjunction with ACS, used for the 2021 PEP estimate of U.S. and Puerto Rico migration

Benchmark Components of Foreign-Born Immigration



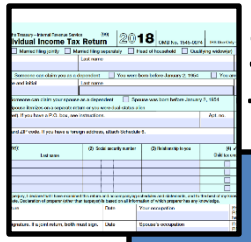
Data Sources: U.S. Census Bureau, Vintage 2020 Population Estimates-Foreign-Born Immigration Component; U.S. State Department, Bureau of Consular Affairs; Institute of International Education; Refugee Processing Center; U.S. Citizenship and Immigration Services (USCIS); U.S. Department of Justice

- Compiled 10 benchmark components of foreign-born movement to the U.S.
- Not an exhaustive list but covers most categories of the foreign born who reside in the U.S. *most of the year*
- Benchmarks and Population Estimates Program (PEP) estimates trend similarly. Benchmarks surpass PEP immigration levels partially due to increased asylees in recent years
- 2020 PEP is a COVID-19 projection of 2019 ACS data
- 2021 PEP is the standard projection assumption, which was not plausible given COVID-19 impacted all of 2021

Example of Micro-Data Integration from the United States: Integrated Database for International Migration (IDIM)

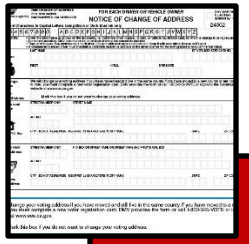
- Developing a linked administrative (and survey) database
  - Administrative data are comprehensive for their population universe, which should allow for more accurate estimation of small geographic areas
  - Administrative data are available much sooner than survey data (3-4 months vs. 1-2 years)
  - Noticeable undercoverage of certain immigrant populations
    - Short-term migrants, undocumented migrants, students, dependents of migrant workers, etc.

Data Sources Currently on IDIM



Internal Revenue Service 1040 Tax Data

- SSN (Social Security Number)
- Name
- Date of Birth
- Address
  - Domestic*
  - Foreign



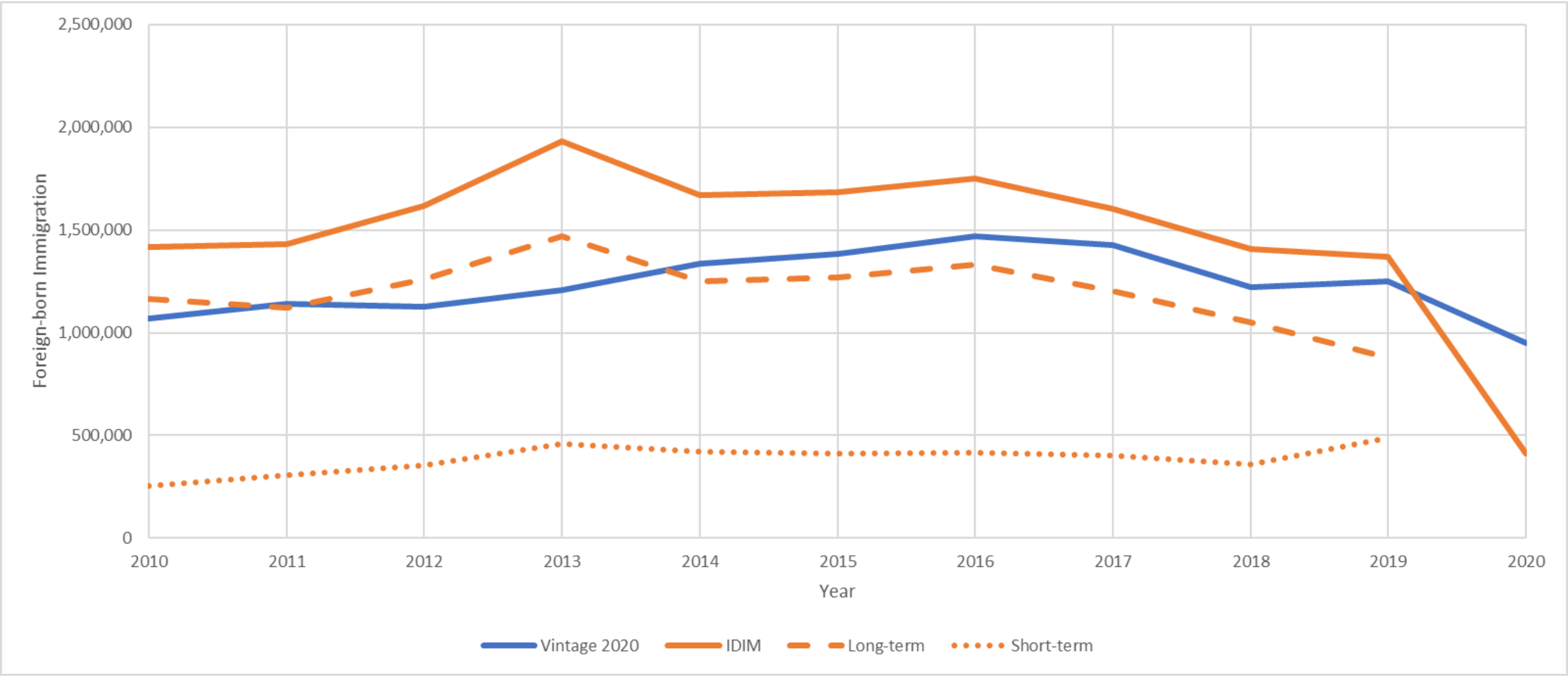
Social Security Administration Numident

- SSN
- Name
- Date of Birth*
- Sex*
- Year of Entry*
- Citizenship status*
  - Native Born
  - Naturalized Citizen
  - Non-citizen
- Place of Birth
- Death Flag

ACS

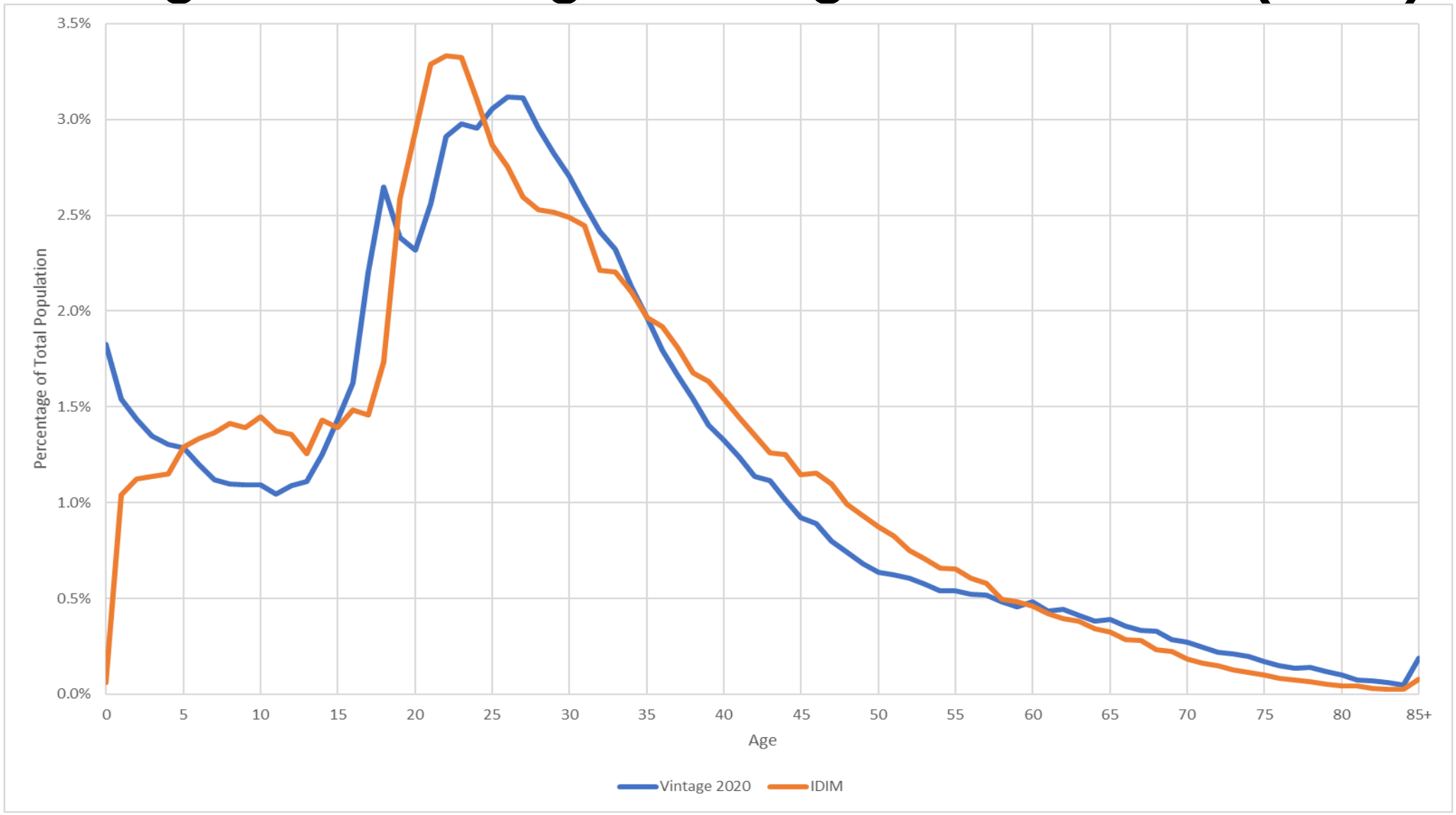
- Name
- Date of Birth*
- Sex*
- Race and Hispanic Origin
- Year entered United States*
- Citizenship Status*
- Socio-economic Characteristics
- Place of Birth
- Address*

Foreign-born Immigration Totals



Sources: Social Security Administration, Internal Revenue Service, U.S. Census Bureau Vintage 2020 Estimates

Foreign-born Immigration Age Distribution (2019)



Sources: Social Security Administration, Internal Revenue Service, U.S. Census Bureau Vintage 2020 Estimates