Globally Estimating the Population Characteristics of Small Geographic Areas

Tom Fitzwater

U.S. Census Bureau | Population Division

What we know

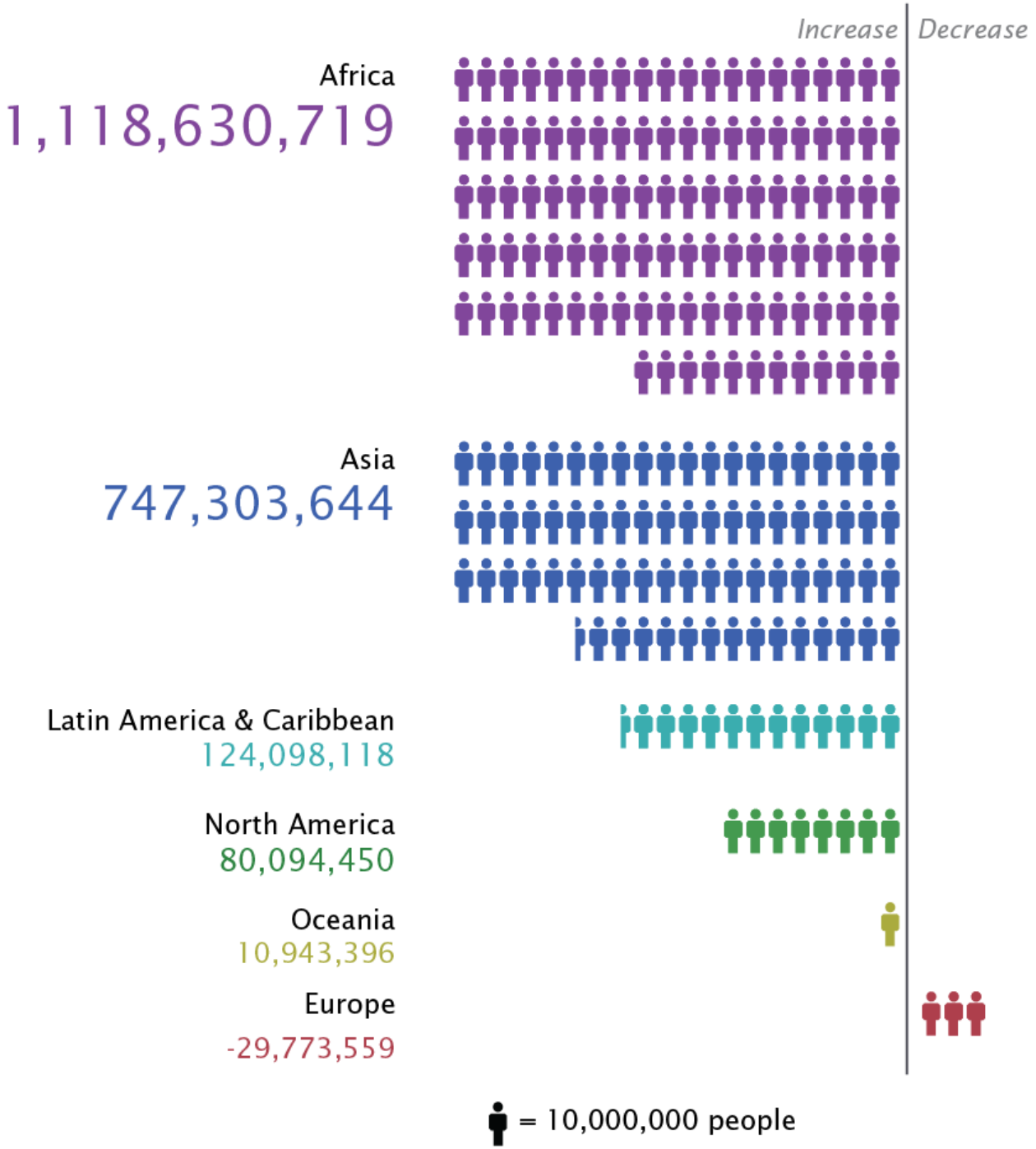
Where do people live?

- **Difficult to measure and quantify.**
- Characteristics:
  - Age and sex distribution, poverty, literacy, access to services, etc.
- Measuring at many scales:
  - Country
  - Region
  - City
  - Neighborhood
  - Person

Data users want answers at **every scale**.

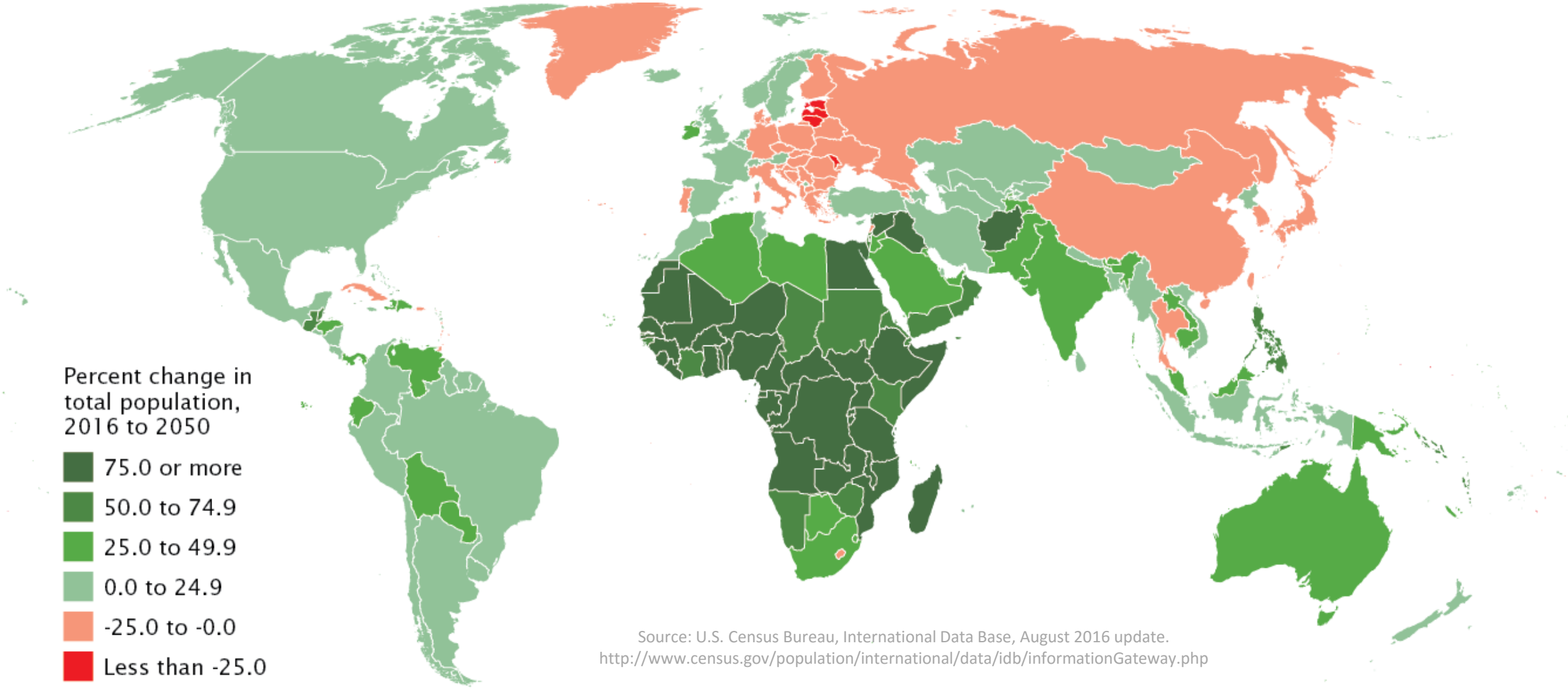
Regional population change

Projected population change: 2016 to 2050



Source: U.S. Census Bureau, International Data Base, August 2016 update. http://www.census.gov/population/international/data/idb/region.php?N=%20Results%20&T=6&A=aggregate&RT=0&Y=2016,2050&R=110,120,130,140,150,160&C=

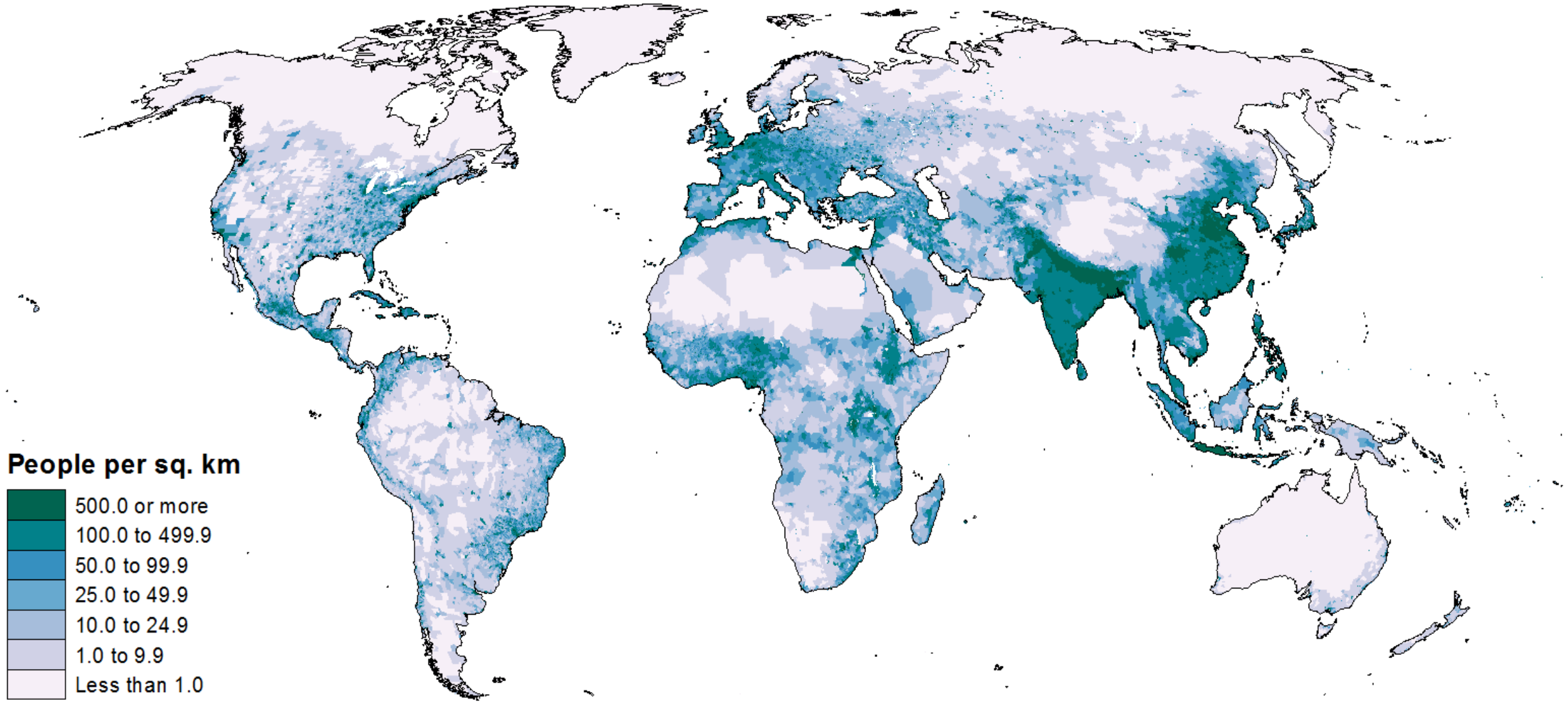
Projected population change by country



Source: U.S. Census Bureau, International Data Base, August 2016 update.
http://www.census.gov/population/international/data/idb/informationGateway.php

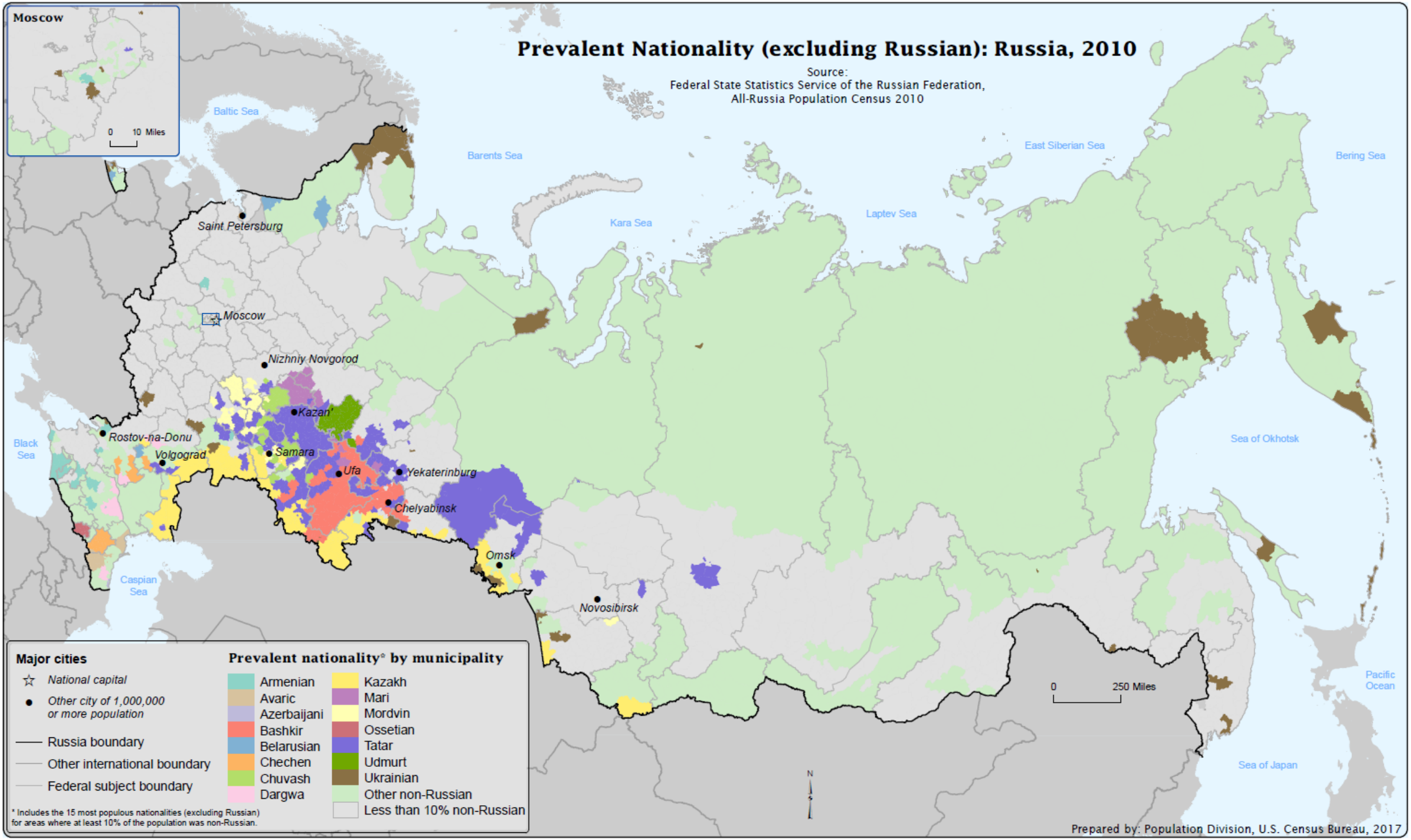
Learn more at: census.gov/population/international

Population distribution



Source: U.S. Census Bureau, subnational population estimates, 2011.

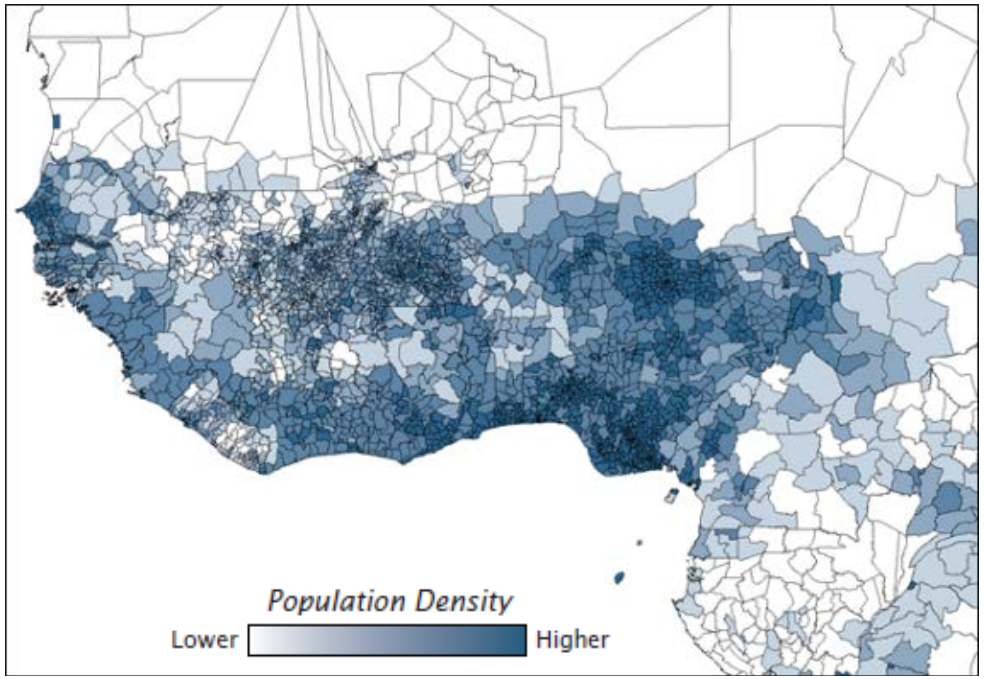
Learn more at: census.gov/population/international

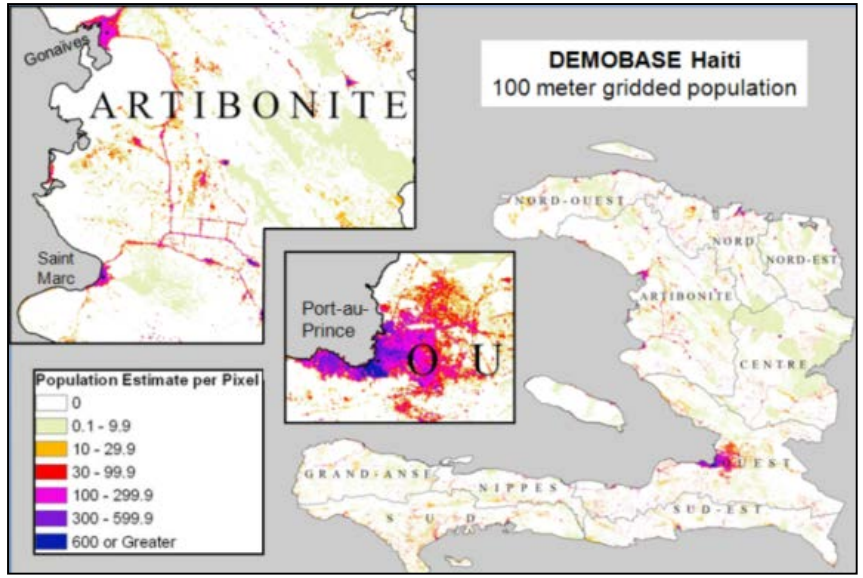

Prevalent Nationality (excluding Russian): Russia, 2010
Source:
Federal State Statistics Service of the Russian Federation,
All-Russia Population Census 2010
Moscow
0 10 Miles
Baltic Sea
Barents Sea
Kara Sea
Laptev Sea
East Siberian Sea
Bering Sea
Sea of Okhotsk
Sea of Japan
Pacific Ocean
Black Sea
Caspian Sea
Saint Petersburg
Moscow
Nizhniy Novgorod
Kazan'
Rostov-na-Donu
Volgograd
Samara
Ufa
Yekaterinburg
Chelyabinsk
Omsk
Novosibirsk
0 250 Miles
N
Major cities
National capital
Other city of 1,000,000 or more population
Russia boundary
Other international boundary
Federal subject boundary
Prevalent nationality* by municipality
Armenian
Avaric
Azerbaijani
Bashkir
Belarusian
Chechen
Chuvash
Dargwa
Kazakh
Mari
Mordvin
Ossetian
Tatar
Udmurt
Ukrainian
Other non-Russian
Less than 10% non-Russian
* Includes the 15 most populous nationalities (excluding Russian) for areas where at least 10% of the population was non-Russian.
Prepared by: Population Division, U.S. Census Bureau, 2017

Our methods

International Projects

- **Applied research and analysis:**
  - International population estimates and projections
  - HIV/AIDS research tools
  - Global aging issues
- **Geospatial products:**
  - Subnational population estimates
  - Gridded products (e.g., Demobase)





What we produce

- Use **multiple inputs** for guidance, including censuses and surveys.
- Estimates and projections:
  - Evaluate and separately estimate base **population**, **fertility**, **mortality**, and international **migration** for each country.
  - Analyze and **project trends** in above variables which change rates and age/sex composition **over time**.
- **Additional indicators**

Censuses

Infrequent
Detailed geography

Surveys

More frequent
Less detailed geography
More characteristics

How many?

Geospatial

Administrative boundaries
Statistical boundaries

Where?

Common census and survey variables

- No country releases individual records, except in the form of "sanitized" public **microdata**.
- All censuses have **total population, age, and sex** variables.
- Other variables differ among countries.
  - **Personal:** ethnicity, language, health, fertility, mortality
  - **Family:** marital status, household size
  - **Migration:** birthplace, citizenship
  - **Education:** literacy, level of education
  - **Work:** employment, job categories
  - **Economic:** income, poverty
  - **Housing:** wall/roof materials, building age
  - **Land use:** urban/rural, agriculture
  - **Access to services:** water, electricity, transport, communications

How we evaluate data

- In our products, we account for census inaccuracies or unexpected events, such as:
  - Systematic **undercounting** of women and children.
  - **Major events**, such as disasters, war, disease.
  - Unexplained or **unlikely population growth**.
  - **Differential coverage** between censuses.
  - Discrepancies in age reporting, such as **age heaping**.
  - Problems reported in a **post-census survey**.
- These considerations are critical to ensure proper analysis of a country's population characteristics.

Studying change over time within a country

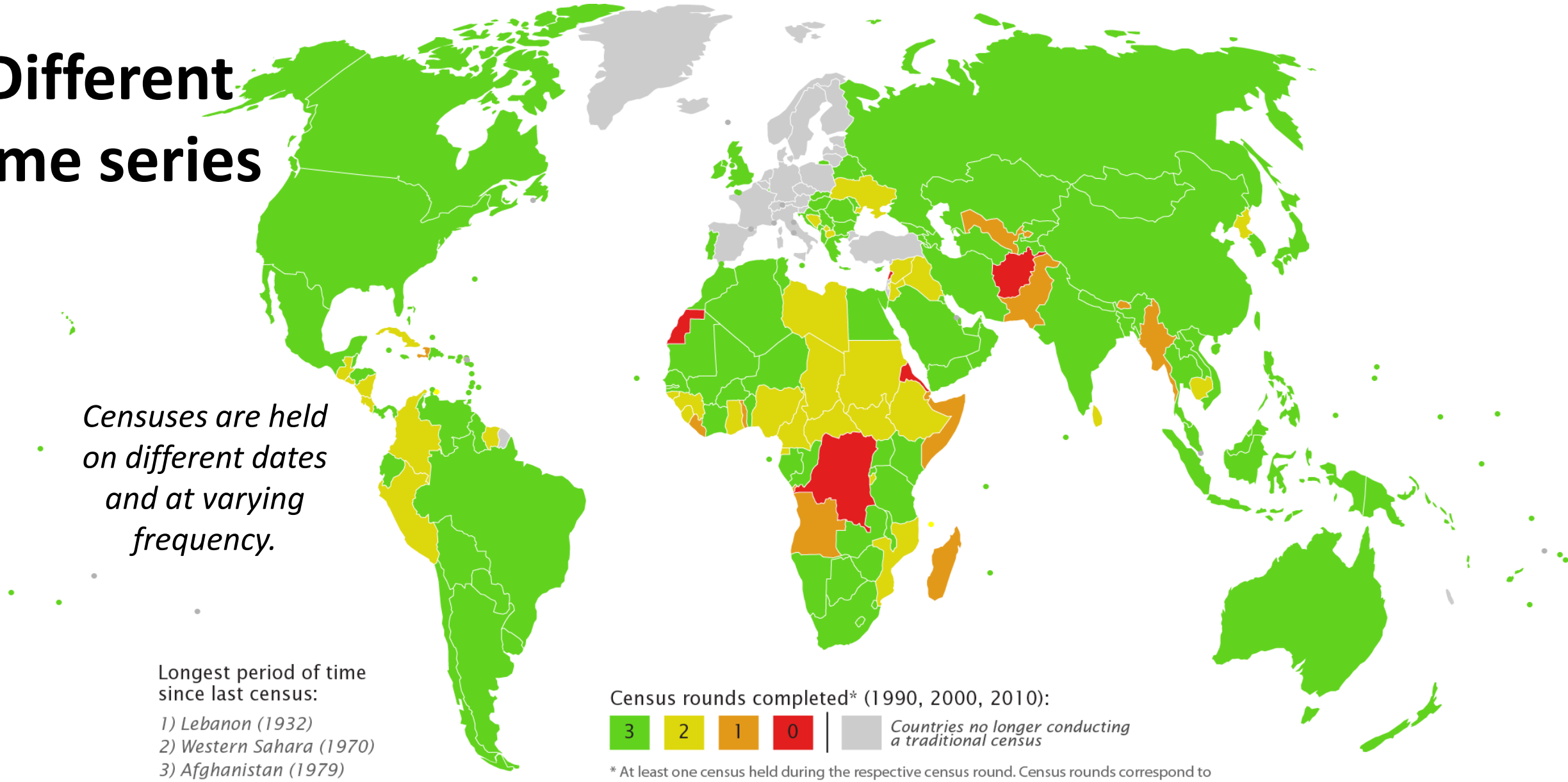
- Comparability:
  - Even if an indicator is present in 2 censuses/surveys, **we do not assume they are directly comparable.**
  - Methods, definitions, or wording of the question may have changed, rendering comparisons misleading.
- Evolving geography:
  - Some countries redraw boundaries frequently.
  - **We match census or survey data to available geographic boundaries.**

Comparing data between countries

- Working across borders is a significant challenge.
- From a geospatial perspective, boundaries from one country's government **may not match** boundaries from neighboring country.
- Census and survey variables **may not be comparable** from country to country due to differing definitions.
  - Example: How do you define ***urban areas***?

Different time series

Censuses are held on different dates and at varying frequency.



* At least one census held during the respective census round. Census rounds correspond to the years 1985-1994 (1990 round); 1995-2004 (2000 round); and 2005-2014 (2010 round).

Source: United Nations 2020 World Population and Housing Census Program <unstats.un.org/unsd/demographic/sources/census/censusdates.htm>

Geospatial data

- **International land boundaries**
  - Large-Scale International Boundaries (LSIB) (geonode.state.gov)
- **Shorelines and inland water bodies**
  - World Vector Shoreline (WVS), or national sources.
- **Internal/subnational boundaries**
  - National sources, GADM, open data repositories.
- **Persistent challenges:**
  - Multiple interpretations for international boundaries.
  - Representing the area of enumeration.
  - Inland water bodies.
  - Finding freely licensed subnational data.
  - Matching subnational boundaries with statistical data.

New tools and methods

New tools and methods

- We leverage our agency's **strengths**:
  - One of the **largest holders** of geographic information in the U.S. government.
  - **Center for excellence** in production and understanding of human geography and population data.
  - Continually **innovating** to conduct U.S. censuses and surveys more efficiently and effectively.
- We apply tools and methods built for domestic purposes to international projects.

Automated data collection & categorization

- Objectives:
  - Index datasets and publications from key websites.
  - Categorize datasets as useful or not useful.
  - Structure datasets regardless of format (PDF, XLSX, CSV, API, HTML, etc.).
  - Save staff time and increase data availability.
- Primary toolset:
  - Apache **Nutch** (web scraper), Python **Natural Language Toolkit** (language), Python **scikit-learn** (machine learning classifier)

Automated data collection & categorization

- Build language-specific **training datasets**.
- **Extract text** from datasets or documents (initial focus: PDF).
- Identify common features based on *n*-grams, or sequences of *n* words.
  - e.g., *2*-gram = 2 word pair like "demographic survey"
- Tested Naïve Bayes & **Support Vector** classifiers.
  - Promising early results: **~93%** accuracy (SVM) for English-language documents, compared to human classification.

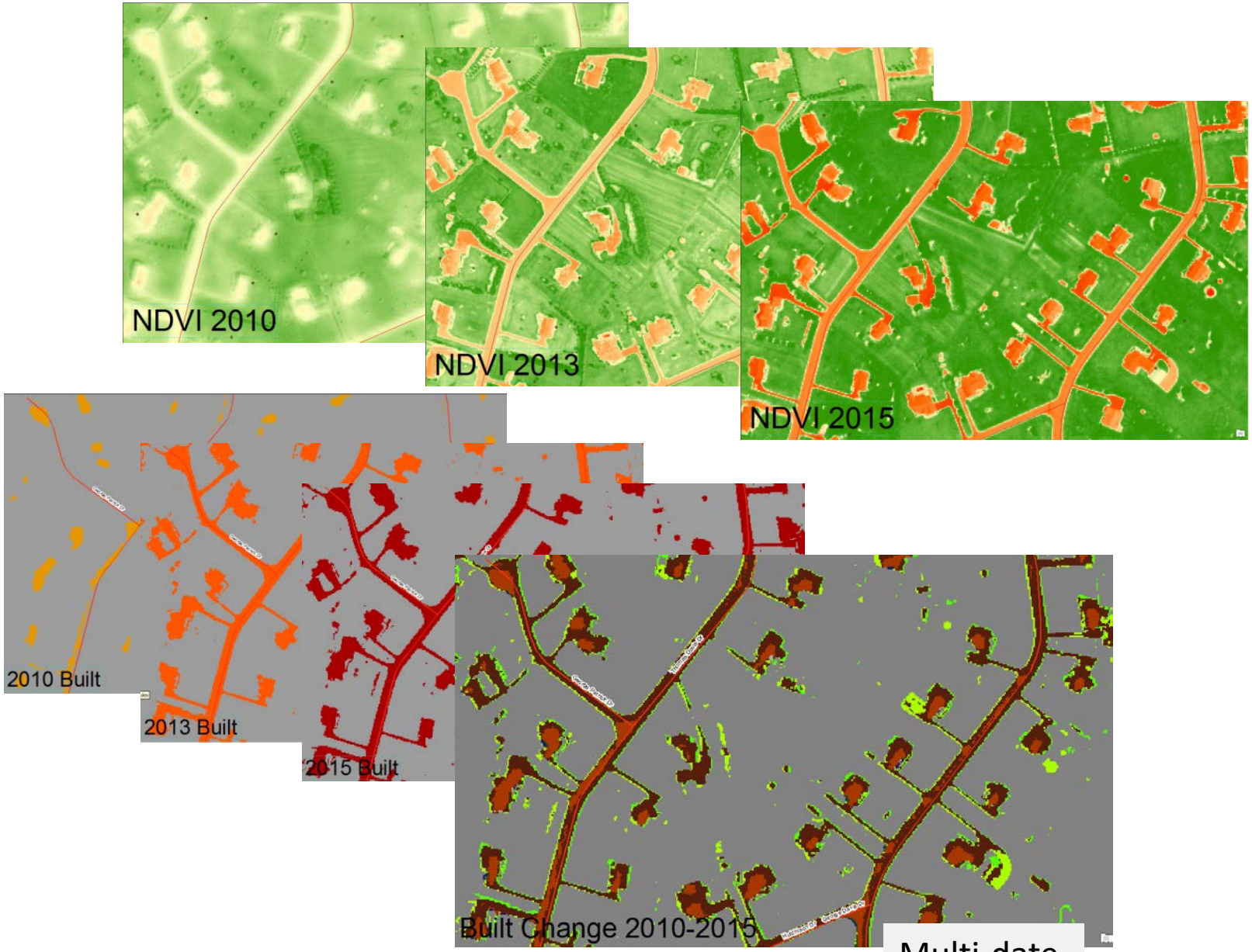
- **Future considerations:**
  - Continuously search for new data.
  - Tag data by subject, source, location.
  - Automatically extract and structure data tables.
  - Develop feedback process to improve classification.
  - Explore geospatial data identification and extraction.
  - Expand to other file formats and APIs.
  - Cloud architecture to scale up/down easily.

Building extraction

- Identifying **physical features** from **remotely sensed** data.
- Use a combination of:
  - High resolution imagery (aerial or satellite)
  - Light detection and ranging (lidar)
  - Vector geospatial data (e.g., road networks)
- Work under way to enhance U.S. housing unit location accuracy in support of census/survey operations.
  - Census Bureau holds 150 million+ housing unit locations for U.S.
- Considering applying similar methods to other countries to refine estimates of population distribution.

Constructing built-up layer

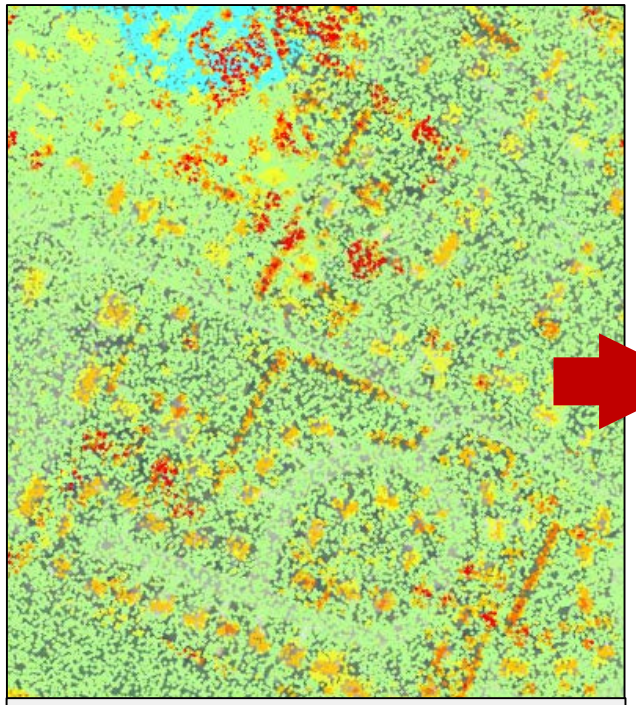
- Built/not-built derived from normalized difference vegetation index (**NDVI**).
- Change identified by comparing multiple time periods.



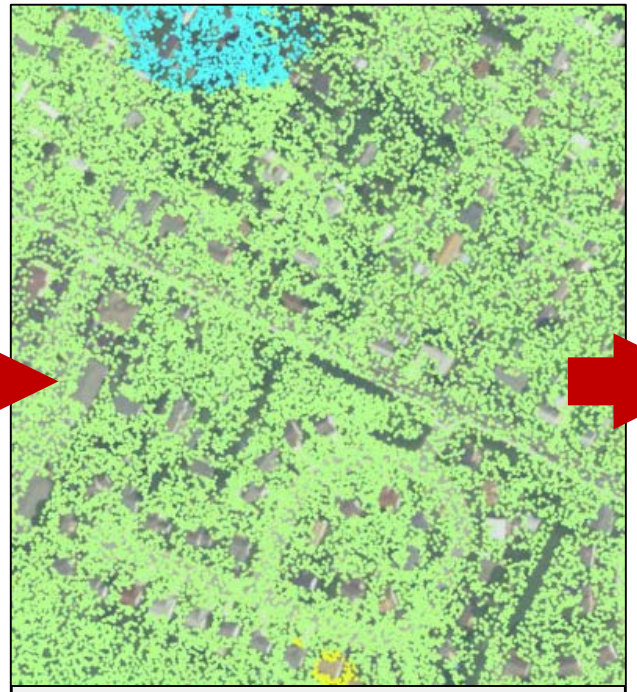
Multi-date Built layer

Learn more at: census.gov/population/international

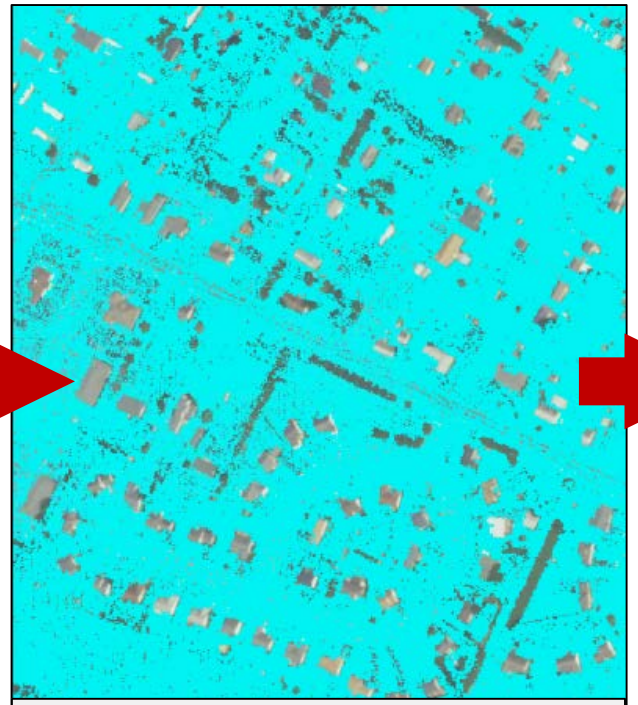
Extracting ground from lidar



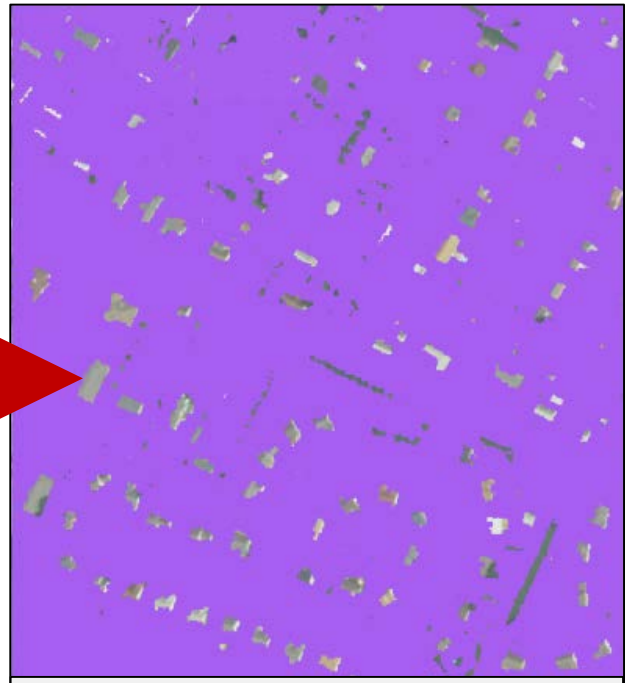
Lidar point cloud.



Queried for ground points.

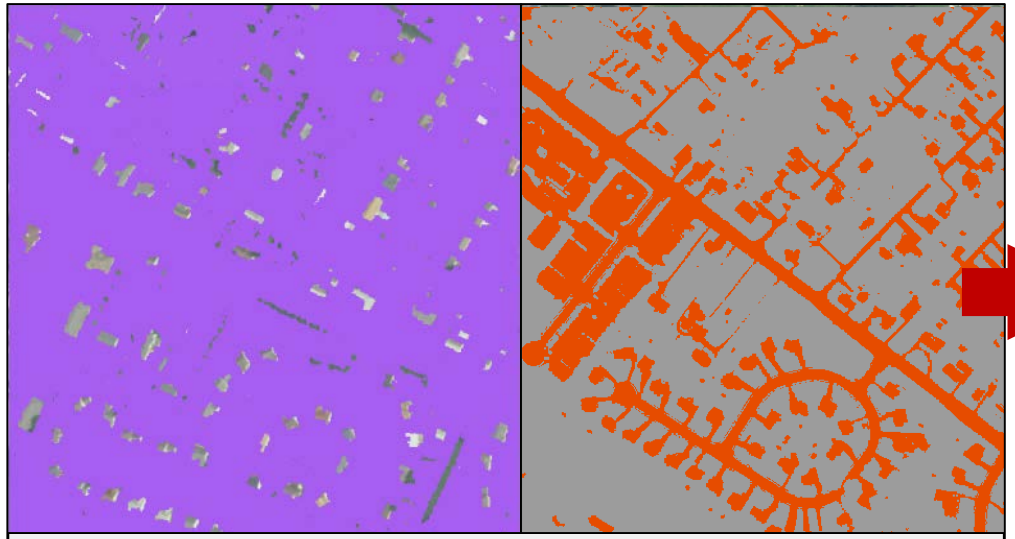


Ground points converted to raster layer.

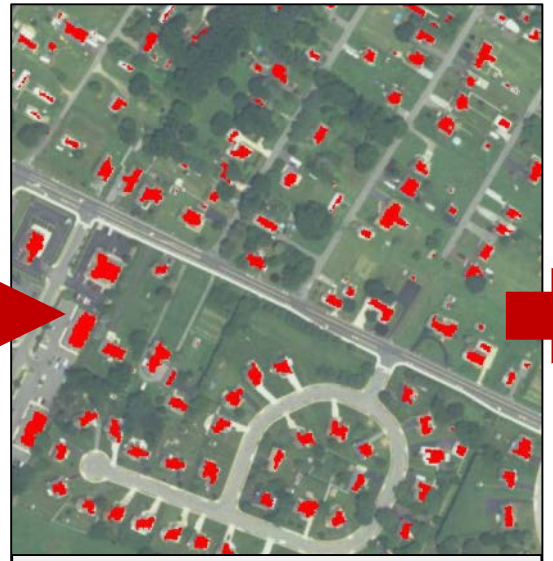


Expanded ground raster layer.

Extracting buildings



Expanded ground raster (left) used as mask to remove low areas of Built/Not Built layer for latest year (right).

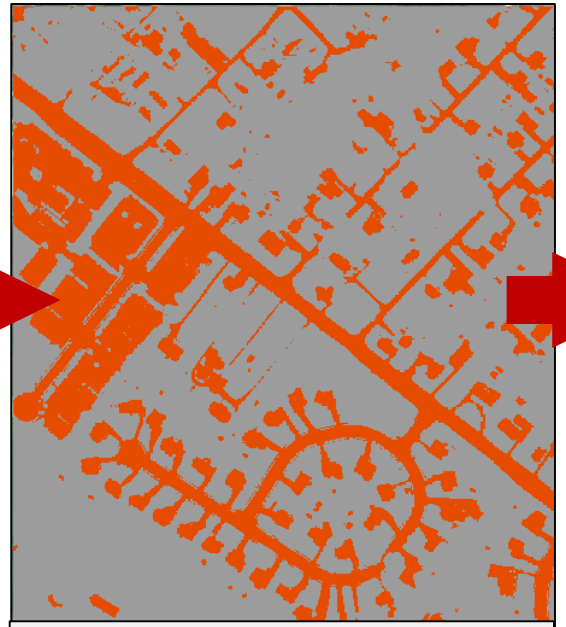


Resultant raster layer represents primarily **buildings**.

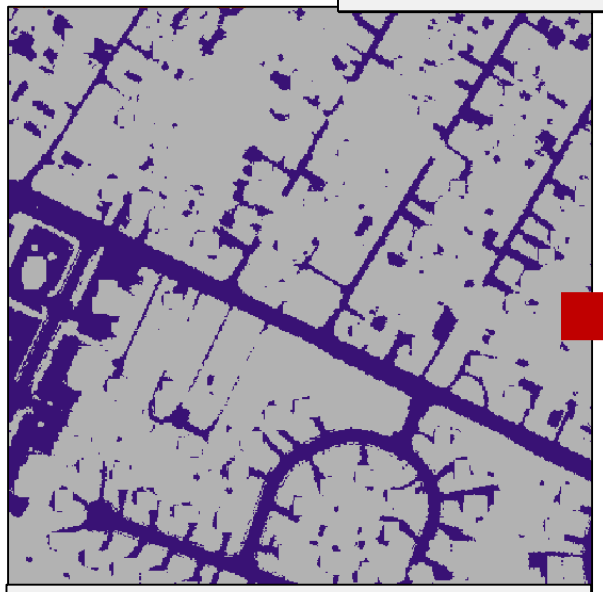
Building polygons generated and **compared** with existing point features in collection. **Footprints** and **heights** captured.

Extracting roads

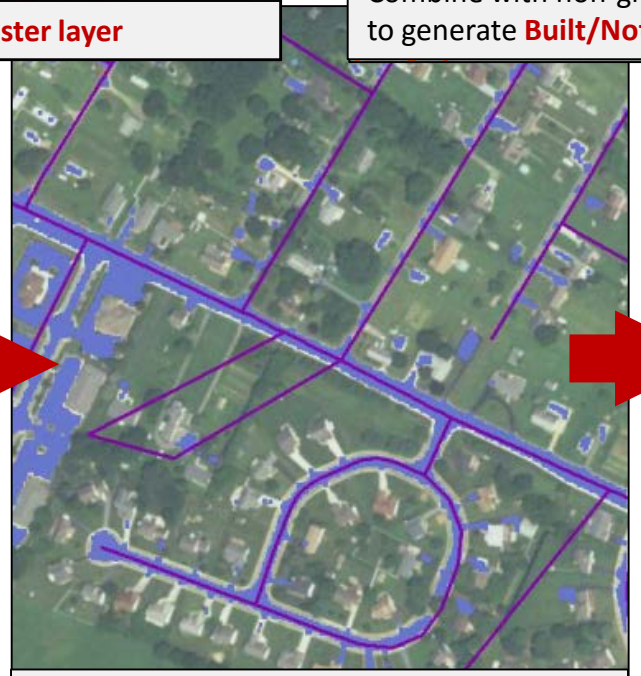
Start with **buildings raster layer**



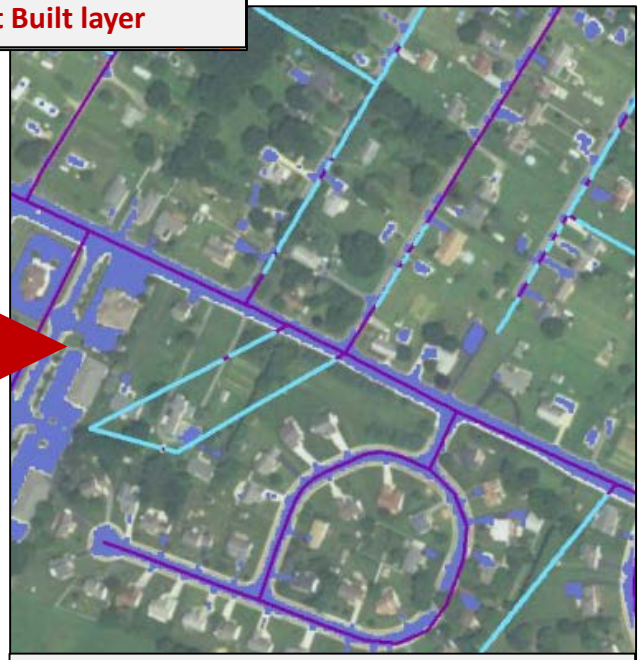
Combine with non-ground point cloud to generate **Built/Not Built layer**



Contraction and dilation to separate roads from parking lot to get **paved area layer**



Small road segments eliminated and intersected with existing road vectors.



Misaligned road segments identified and saved.

Population/settlement surfaces

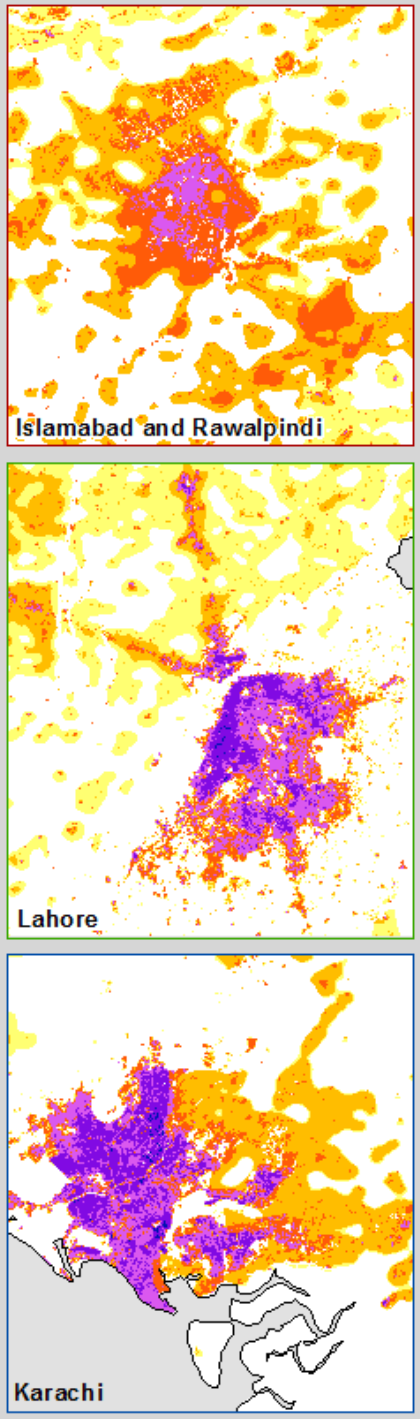
- Increasing number of datasets are providing estimates of **population** and **settlements** in **raster/gridded** format.
- Advantages:
  - Custom geography (not dependent on aggregate results).
  - Works across many scales (from neighborhoods to continents).
  - Ideal for scientific applications.

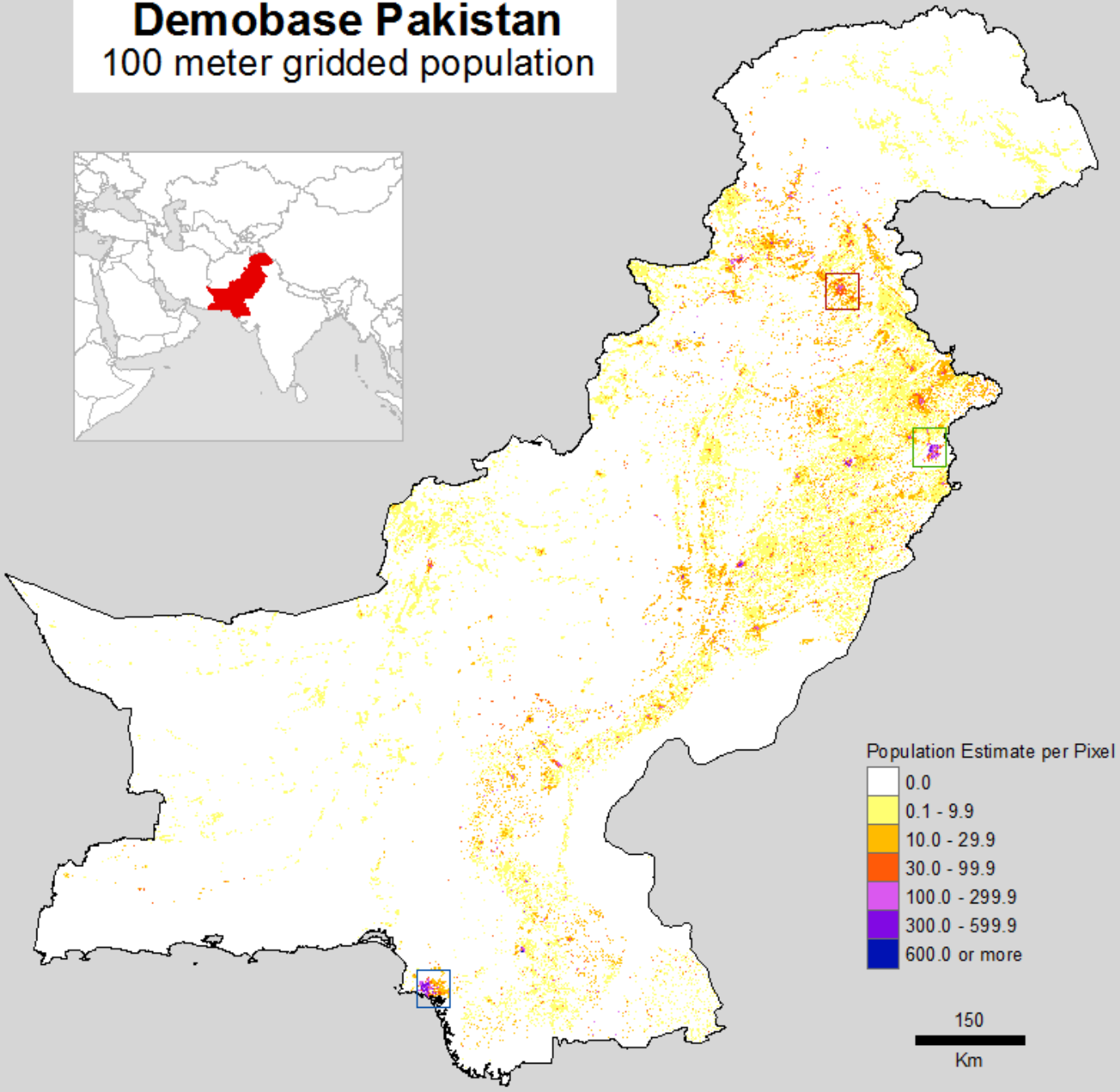
Example datasets

- **Demobase** (U.S. Census Bureau)
- **WorldPop** (University of Southampton)
- **Gridded Population of the World** (CIESIN)
- **High Resolution Settlement Layer** (Facebook/CIESIN)
- **LandScan** (Oak Ridge National Laboratory)
- **Global Human Settlement Layer** (Joint Research Commission)
- **Global Urban Footprint** (German Aerospace Center)

Demobase Pakistan

100 meter gridded population





Source: U.S. Census Bureau, Population Division. For more information, visit census.gov/population/international.

United States Census Bureau

Example methodology: Demobase

- Starting point: Correlation between **population distribution** and **built-up area**.
- **Probability layer:** Combination of built-up area and other layers.
  - Where are humans likely to live?
  - Population assigned to pixels based on fraction of pixel that is built up.
  - Probability layer plays an important role in rural areas.

Example methodology: Demobase

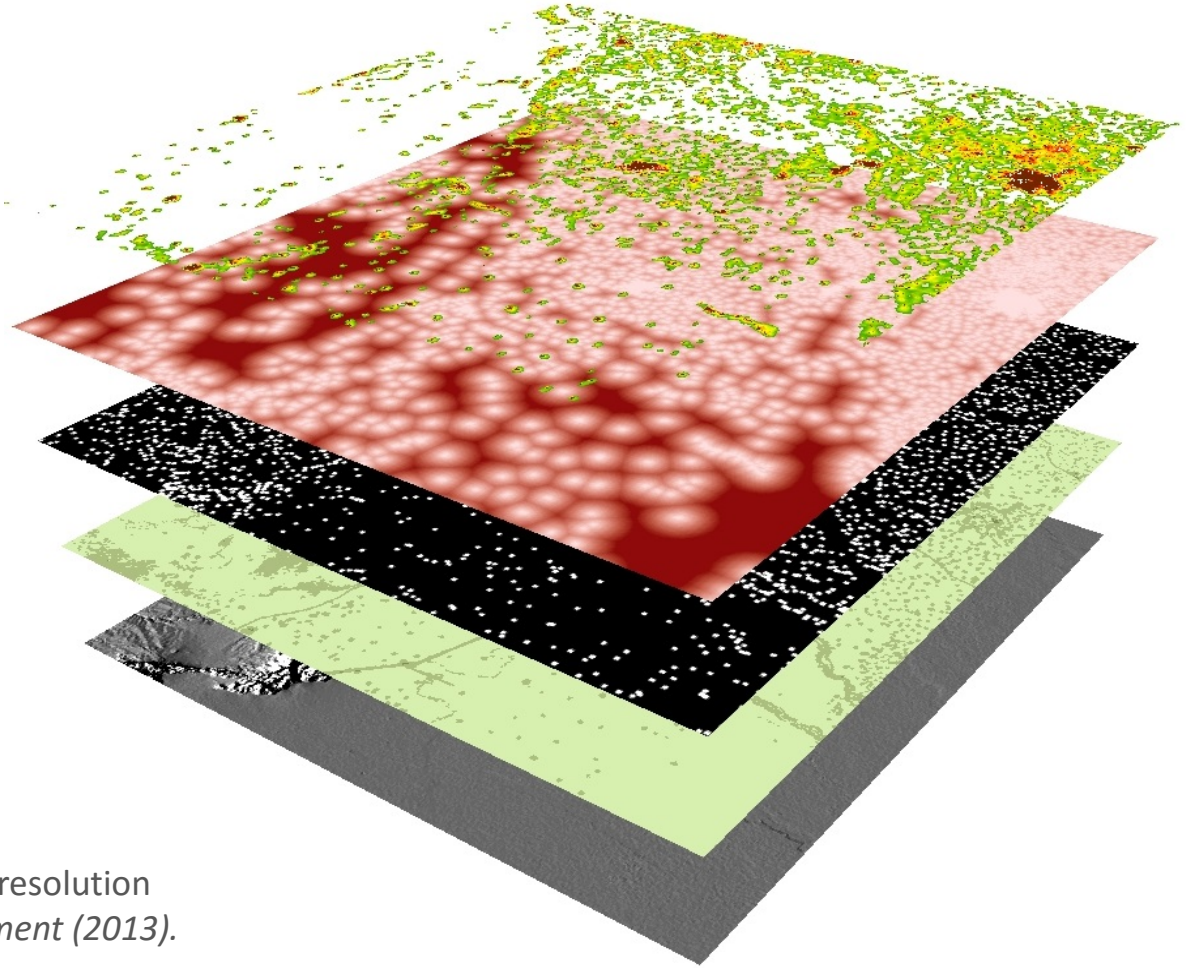
Probability layer combines:

Proximity to built-up

Settlement points

Land cover heterogeneity

Slope and elevation



Detailed methods:

Azar, et al. Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data. *Remote Sensing of Environment (2013).* http://dx.doi.org/10.1016/j.rse.2012.11.022

Thank you!

Tom Fitzwater

U.S. Census Bureau | Population Division

john.t.fitzwater@census.gov