GRID3

GEO-REFERENCED INFRASTRUCTURE AND DEMOGRAPHIC DATA FOR DEVELOPMENT

BILL & MELINDA GATES foundation

UKaid from the British people

world pop

FLOWMINDER.ORG

UNFPA

Center for International Earth Science Information Network
EARTH INSTITUTE | COLUMBIA UNIVERSITY

Topics

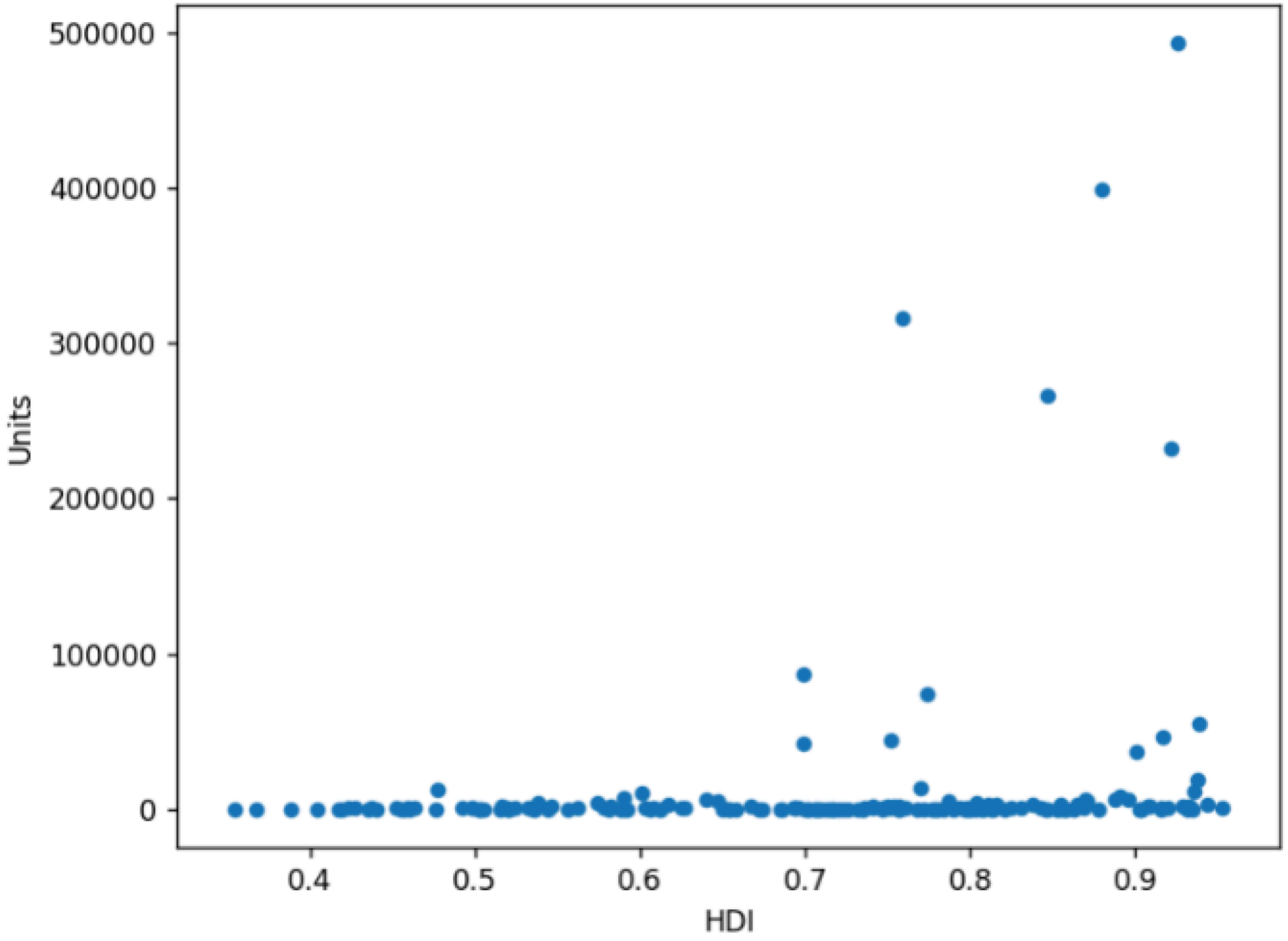
Releasing high-resolution census data for broad use

- Benefits
- Challenges and risks
  - Technical challenges
  - Privacy and confidentiality
- Approaches to manage risk

High Resolution Census Data

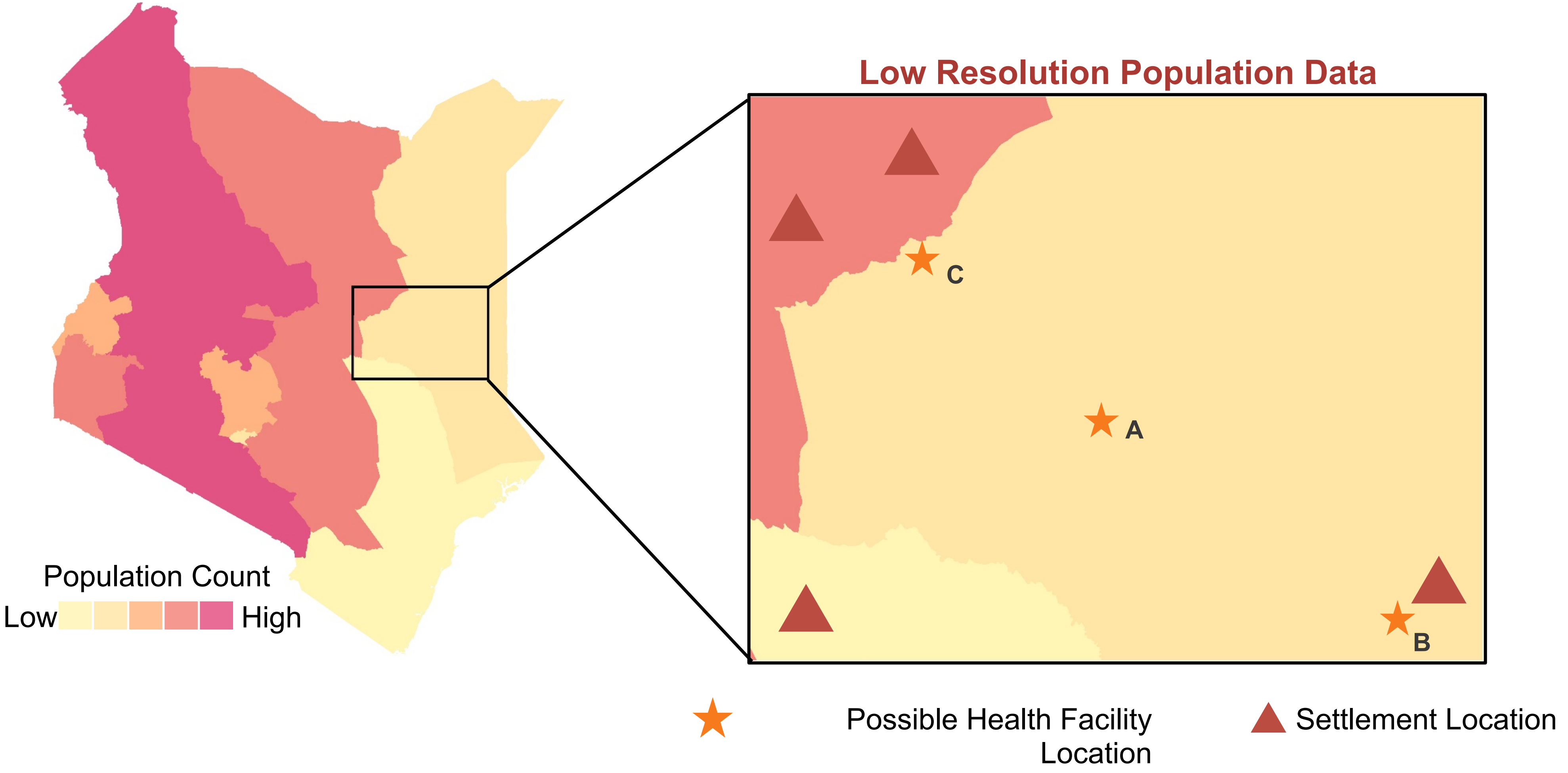
GRID3

Having high-res population data publicly accessible is associated with good development outcomes

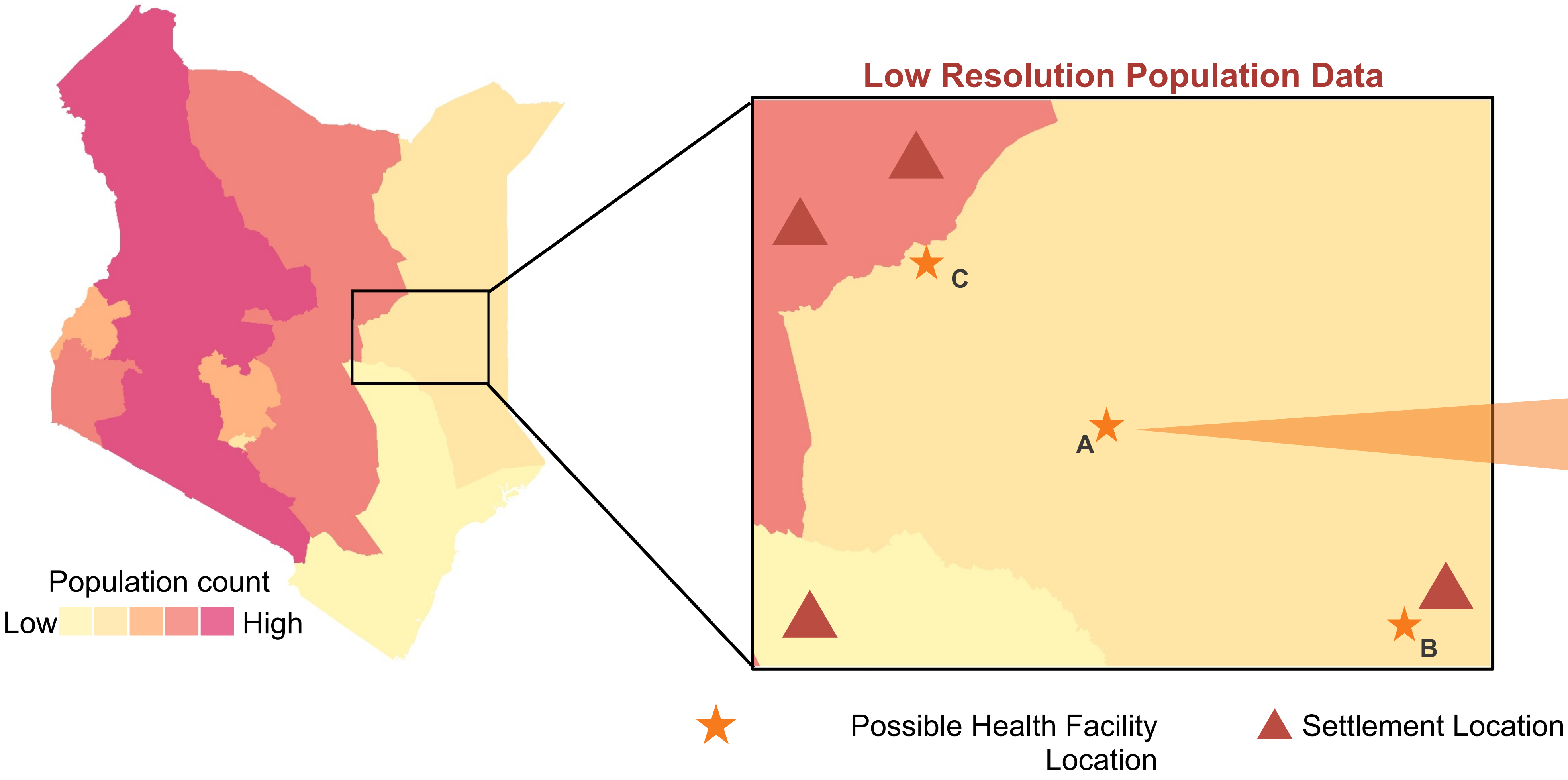


Human Development Index (HDI) vs. number of census units in publicly obtainable census data

The Spatial Data Gap

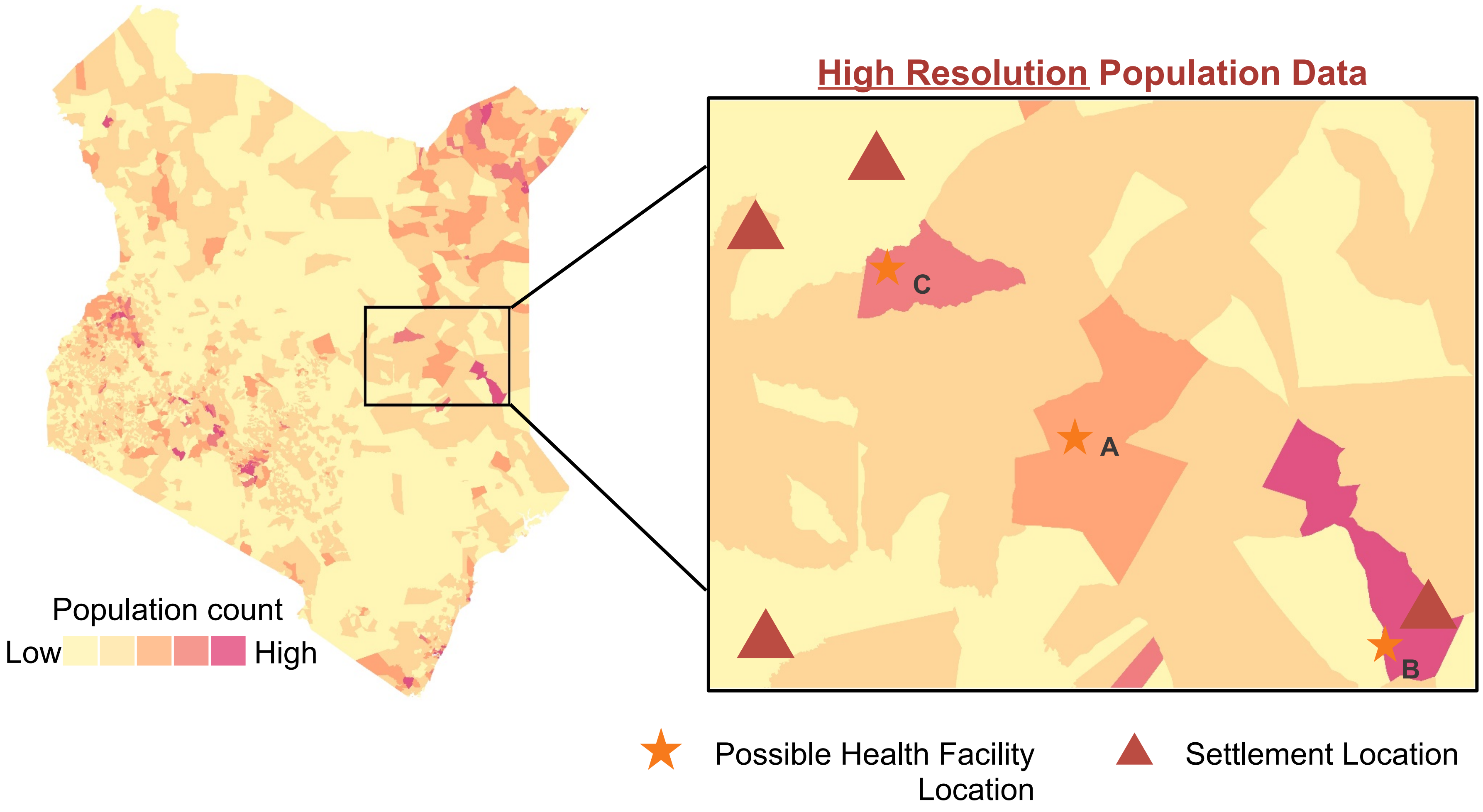

Low Resolution Population Data
C
A
B
Population Count
Low
High
★ Possible Health Facility Location
▲ Settlement Location

The Spatial Data Gap


Low Resolution Population Data
C
A
B
Population count
Low
High
★ Possible Health Facility Location
▲ Settlement Location

At first glace, placing a health facility at location A would have the greatest impact.

The Spatial Data Gap



High Resolution Population Data

But, with high resolution population data, you can see that the population is not evenly distributed across the area but concentrated near one settlement.

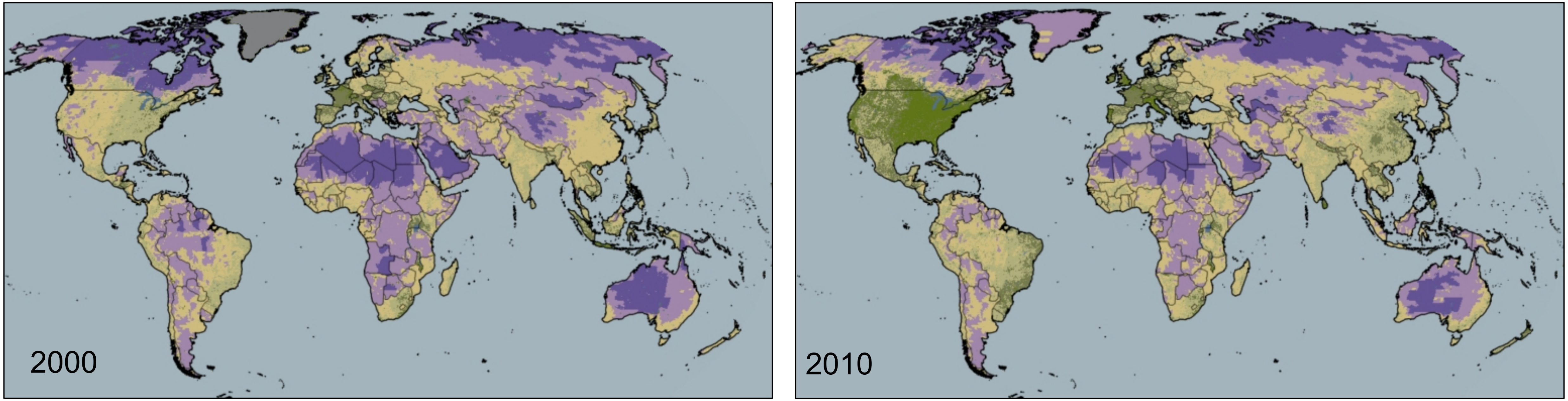
This changes where the health facility should be placed.

Census Data Availability

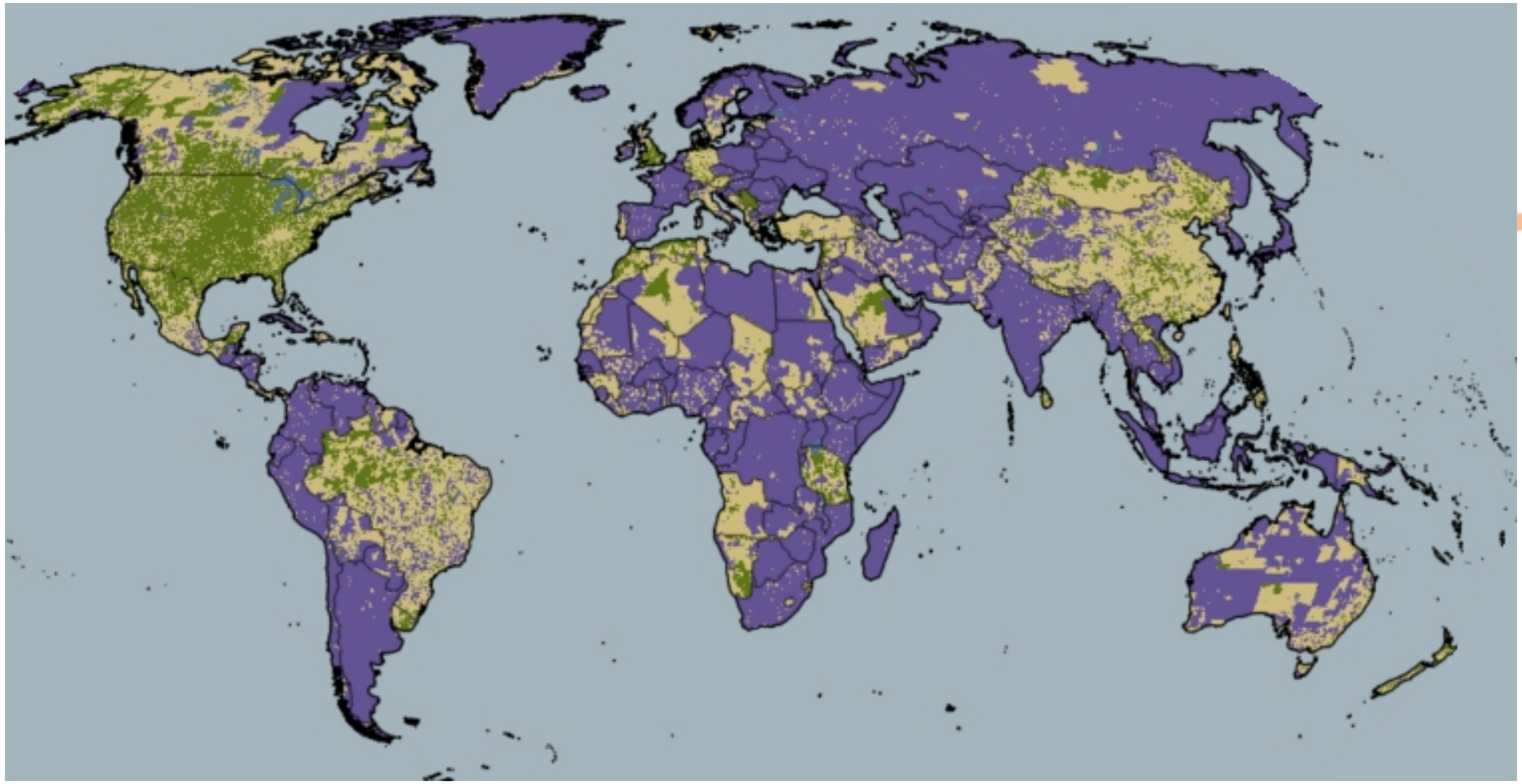
The practice of making available high-resolution population distribution data is now routine in virtually every high-income country, and is spreading to middle-income countries.

- 🟩 < 100 km^2
- ▨ 100 - 1,000 km^2
- ▨ 1,000 - 10,000 km^2
- 🟪 10,000 – 100,000 km^2
- 🟪 < 100,000 km^2

Size of population unit in publicly available data sets



Improvement in spatial resolution of census units, ~2000 round to ~2010 round, in publicly available data sets

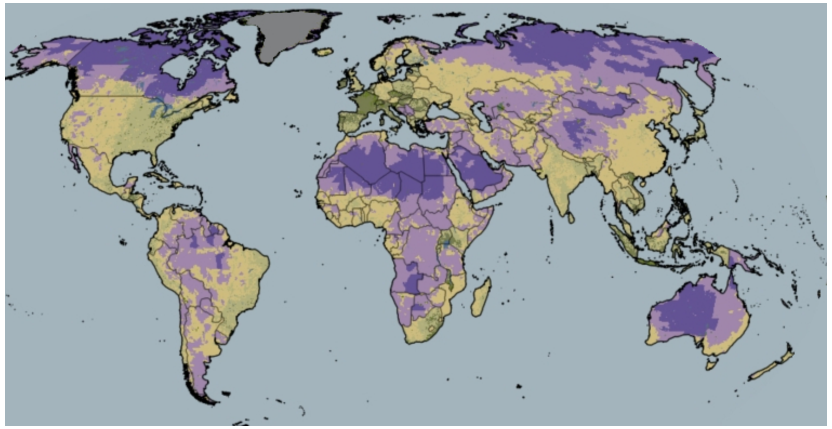


🟩 2 orders of magnitude or better improvement

🟨 1 order of magnitude improvement

🟪 Remained at same order of magnitude

Source: Gridded Population of the World, versions 3 and 4, http://sedac.ciesin.columbia.edu/



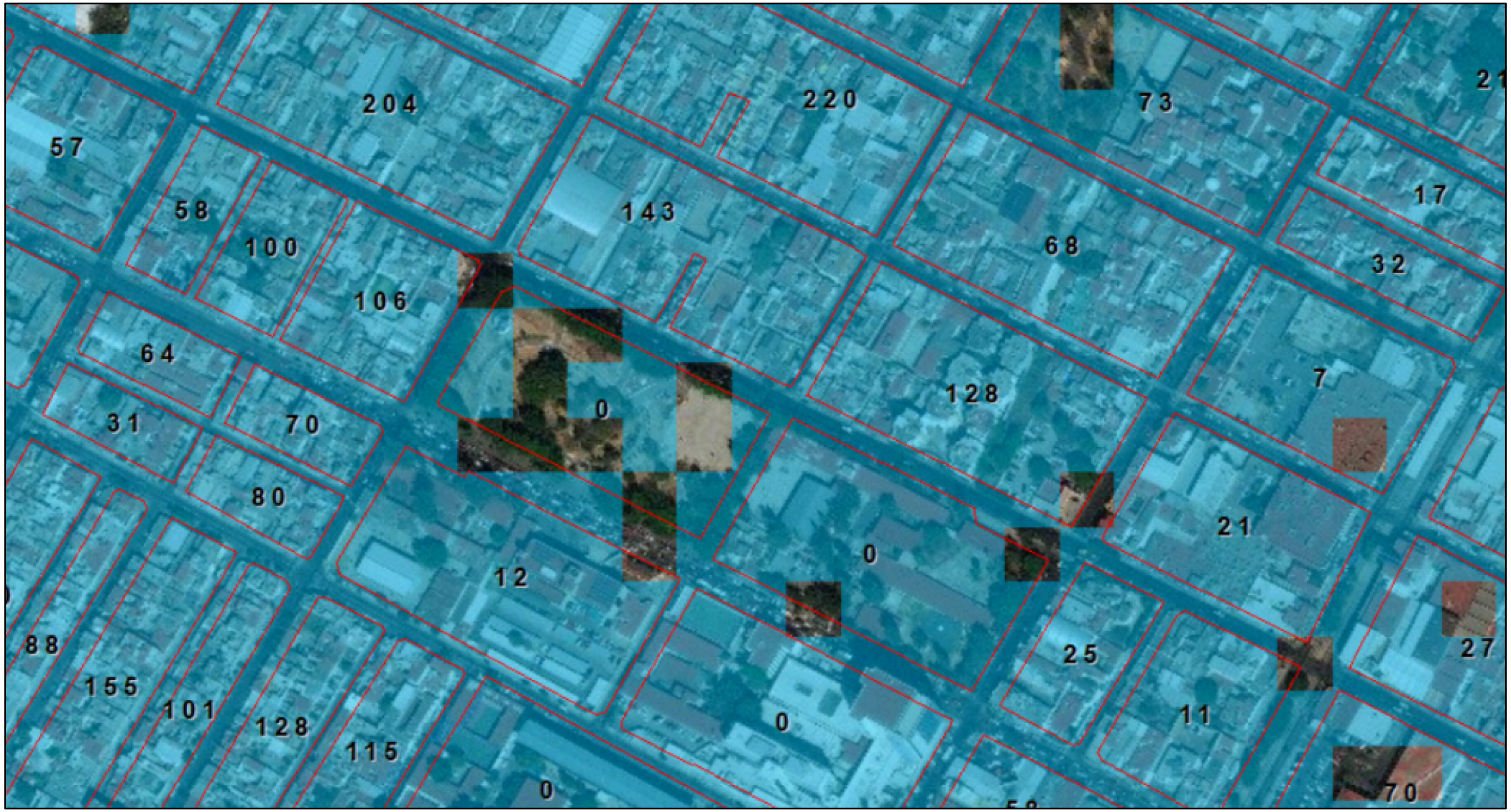
~2000 baseline

However, as of 2010 in most low-income countries population distribution data is only available at highly coarse spatial resolutions.

The countries that could benefit the most from improvements in resolution improved at the slowest rate, between ~2000 and ~2010.

New Data Products

High resolution data aids in developing methods for estimation of population and infrastructure data in areas where detailed data are not available or out of date



INEGI census population counts over settlement classification and high-resolution imagery

Manzanas

🟥

Settlement Extent

🟦

Technical Issues

Improvements in software, hardware, and methods are lowering barriers to the production of high-resolution census data

- Reduced costs in data collection and processing
    - Consumer devices for mobile data collection with GPS
    - Explosion in high resolution remote sensing data acquisition
    - Moore’s law!
- Platforms for near real time data upload and integration
- Tools for data review, quality improvement
- Shortfalls are primarily in technical and economic capacity

Privacy and Confidentiality

Concerns surrounding statistical disclosure prevent the release of detailed census data

- Identity, attribute, and inferential disclosure are legitimate concerns in publicly distributed census and survey data
- Geographic uncertainty in released data is a common practice to reduce the probability of disclosure
  - Aggregation to regions containing thousands of persons (or more) is problematic for detailed mapping & local service planning

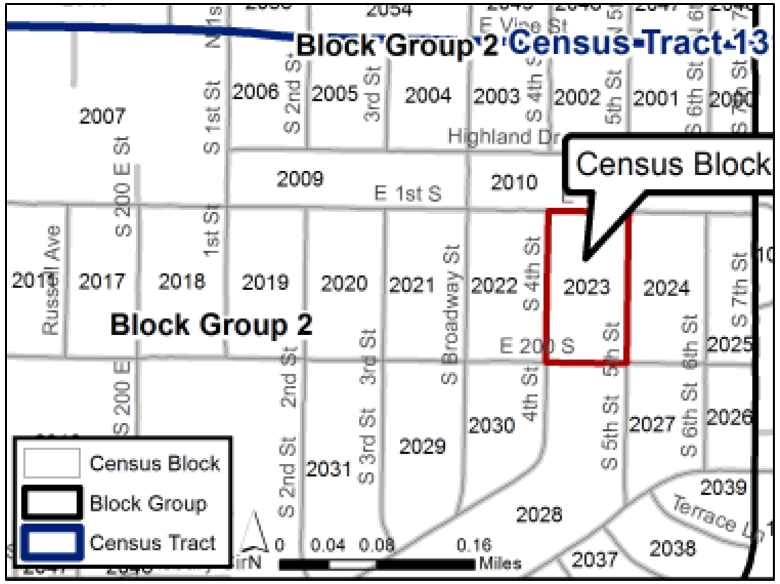
Guarding Against Disclosure

Reducing variables released at the most detailed level reduces disclosure risk

- Basic demographics at a detailed level are very valuable
  - Head counts, age structure, sex are the most "in demand" variables
- U.S. Census data example
  - Basic demographics, race and ethnicity are available at very detailed block level
    - Average population 34
    - Over 11 million geographic units
  - Income, housing, many additional variables are only available at less detailed geographies (block group and larger)
    - Average population 1,200
    - 211,000 geographic units

Census Blocks

Source: U.S. Census Bureau



Protecting Count Data

High-risk records—unique combinations, low counts—can be handled

- Data swapping, where household records at risk of identification are swapped with a different household with similar composition, is used in U.S. block level data
  - Maintains relationships and variance of overall data collection while reducing risk
- Rounding to base
  - Low-count records for detailed geographic units can be rounded to a selected base number while closely preserving overall unit totals
    - Norway uses base 3, with records less than three rounded to zero
    - Canada uses base 5, with records randomly rounded

Protecting Count Data

More detailed spatial databases increases the risk of disclosure

- Additional methods to protect survey data are usually only required for survey and microdata
- Perturbations:
  - Binning
  - Top and bottom coding
  - Suppression of records with higher risk
- Synthetic records can be assembled using software to mask identifiable data
  - Tabulations are calculated from synthetic data
  - Inter-variable relationships are well maintained

Special Considerations

Multiple census years and geographies can compound the problem

- Aggregation of older individual record data to new boundaries can increase the risk of disclosure through boundary and attribute differencing
  - Aggregation approaches that minimize this risk exist
- Publicly available aggregations at multiple detailed geographies (non-nested) can increase the risk of disclosure
  - The lowest-level (most detailed) census units are best when designed to nest within decision making geographies

Conclusion

Detailed census data can be safely released

- Open data and cost-recovery models are both compatible with the needs of national/local planning and the international development community
- With planning and the use of existing technologies, the risk of disclosure can be minimized