The Influence of Statistical Inputs on Global Gridded Geospatial Datasets

Kytt MacManus
Columbia University, CIESIN, USA
EFGS 2013, Sofia Bulgaria

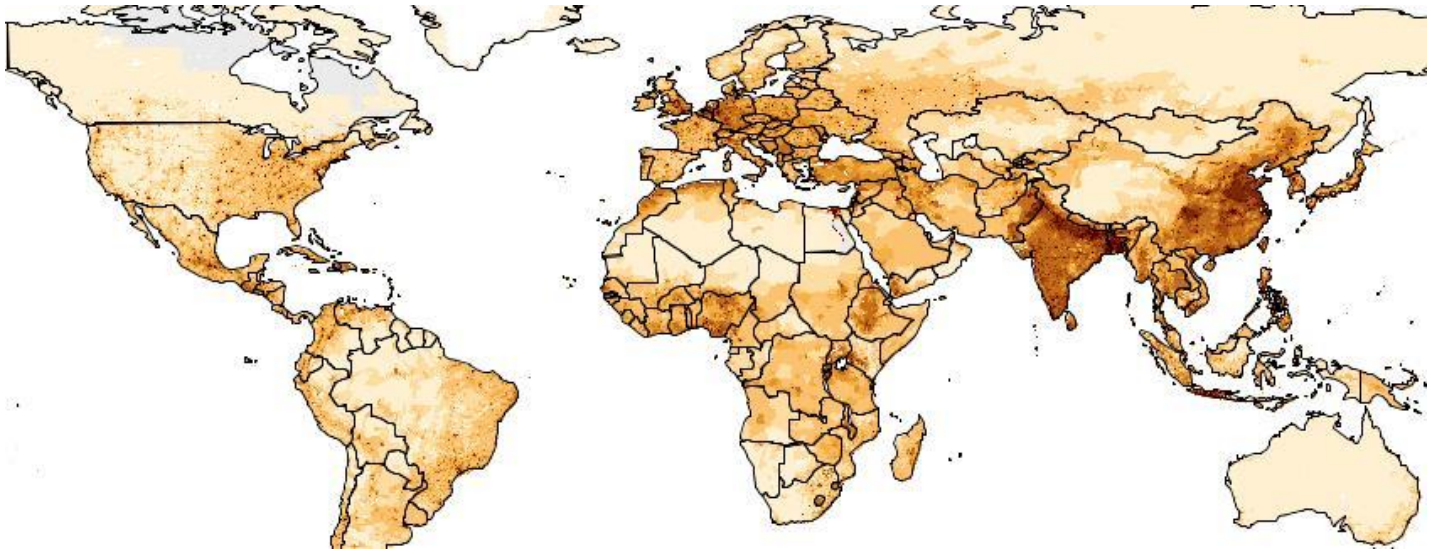
Outline

- Introduce the global population projects: GPW and Terrapop
- Demonstrate the impact of census input resolution on population distribution accuracy
- Discuss the difficulties in linking pop data to boundaries (especially in time series)
- Pleas to CSO's for a better geographic future

Gridded Population of the World

- Gridded (raster) data product developed to provide a **spatially disaggregated population layer** that is compatible with data sets from social, economic, and Earth science fields.

- Population data are transformed from their native spatial units to a global grid of quadrilateral latitude-longitude cells (2.5 arc minutes in GPW3) (Balk et al. 2010)



The basic steps to develop GPW

1. Find tabular population counts
2. Match these to geographic boundaries (census or administrative units)
3. Estimate the population for target years (e.g. 2010)
4. Transform to grids using a uniform distribution algorithm

(Balk, Yetman and de Sherbinin 2010. . Construction of Gridded Population and Poverty Data Sets from Different Data Sources http://sedac.ciesin.columbia.edu/gpw/documents/Balk_etal_GeostatPaper_2010pdf-1.pdf)

GPW v4 highlights

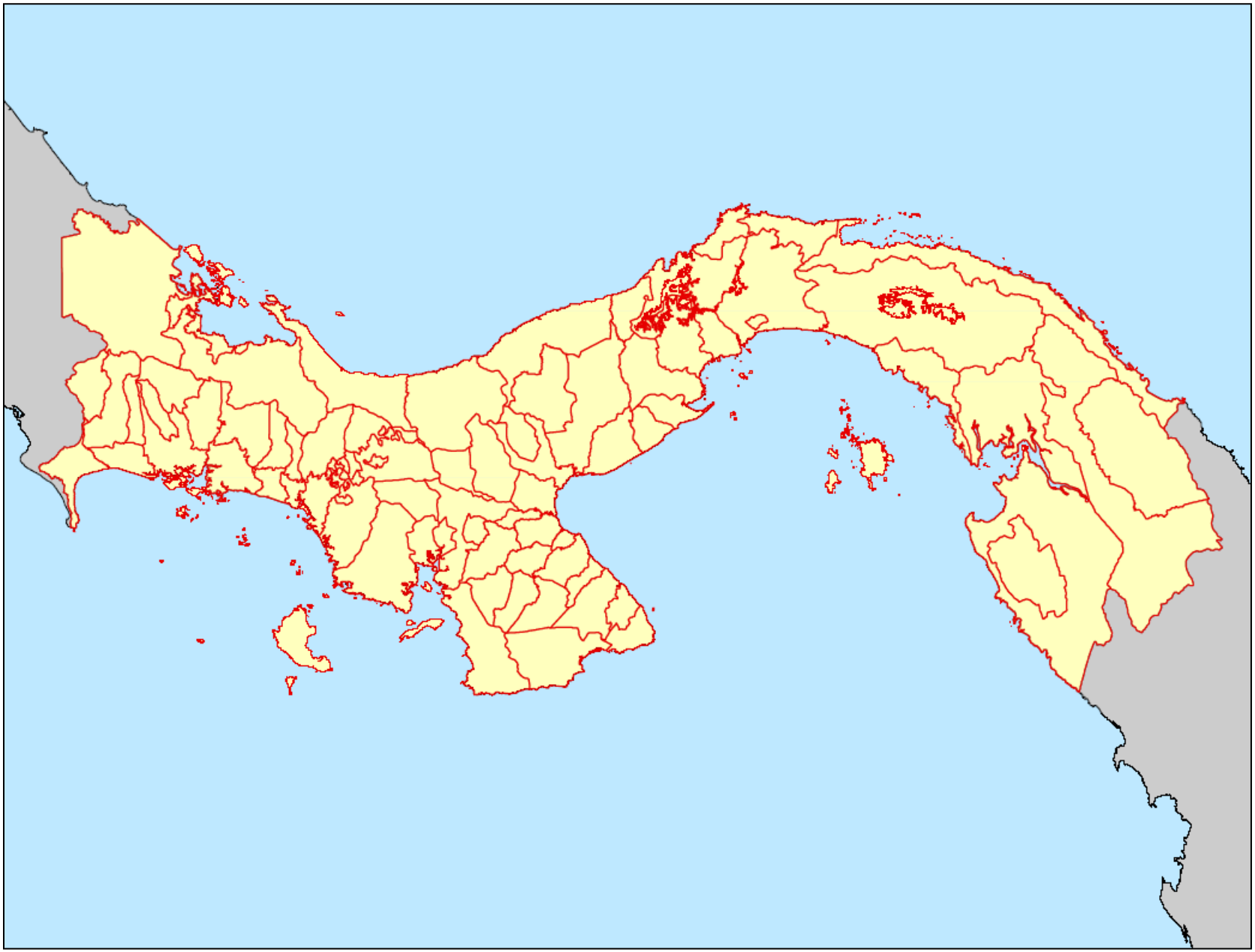
- Basic inputs:
  - 2010 round of population censuses or latest available census data
  - Geographic boundaries matching census cartography

- Variables: population counts, density, **urban/rural status (as defined by the country), age and gender structures**

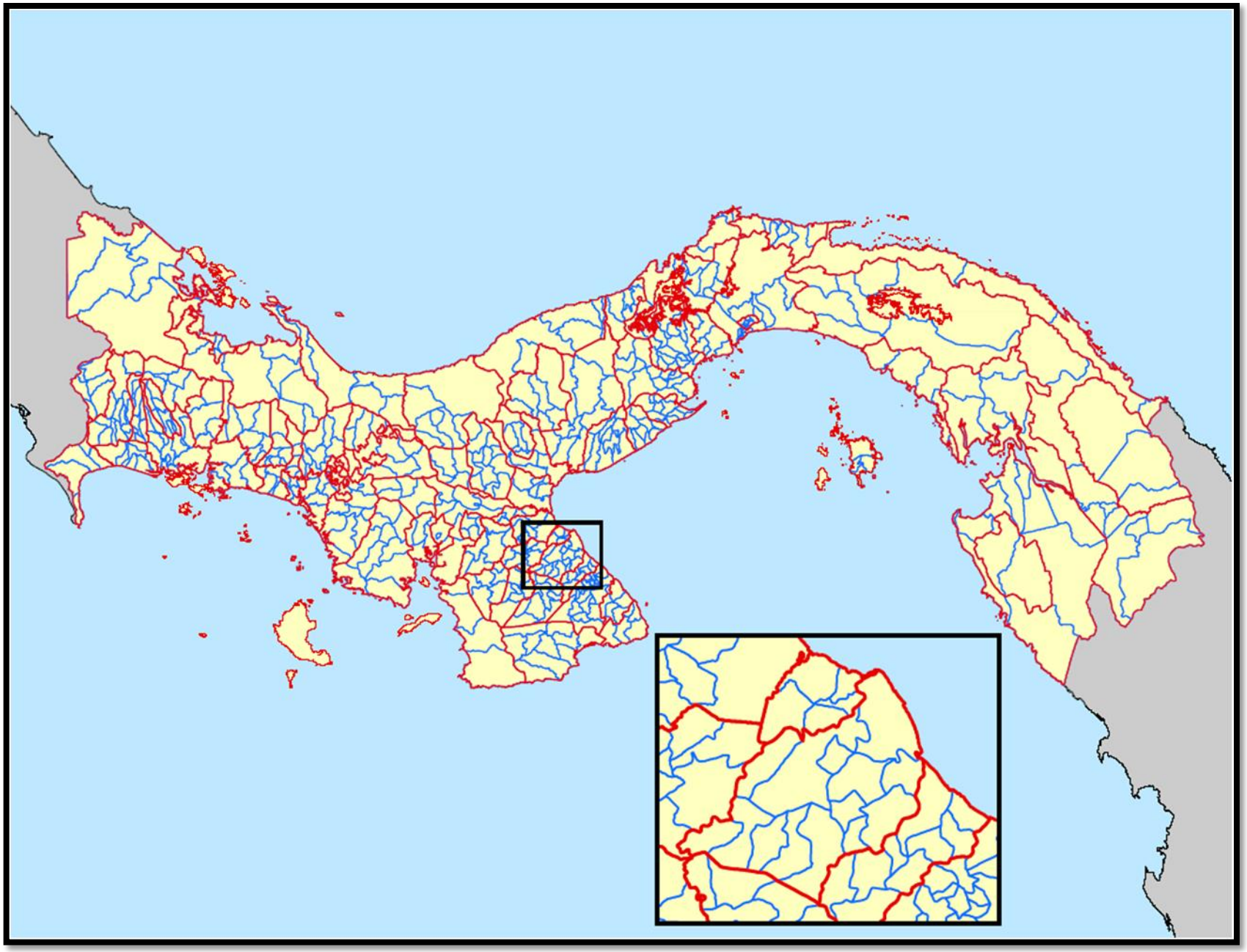
- Higher resolution: **30 arc seconds (approximately 1 km at the equator)**, down from 2.5 arc minutes in GPW v3 (approximately 4km on a side at the equator)

- Expected: changes in the access to the data: from pre-packaged to “on the fly” datasets

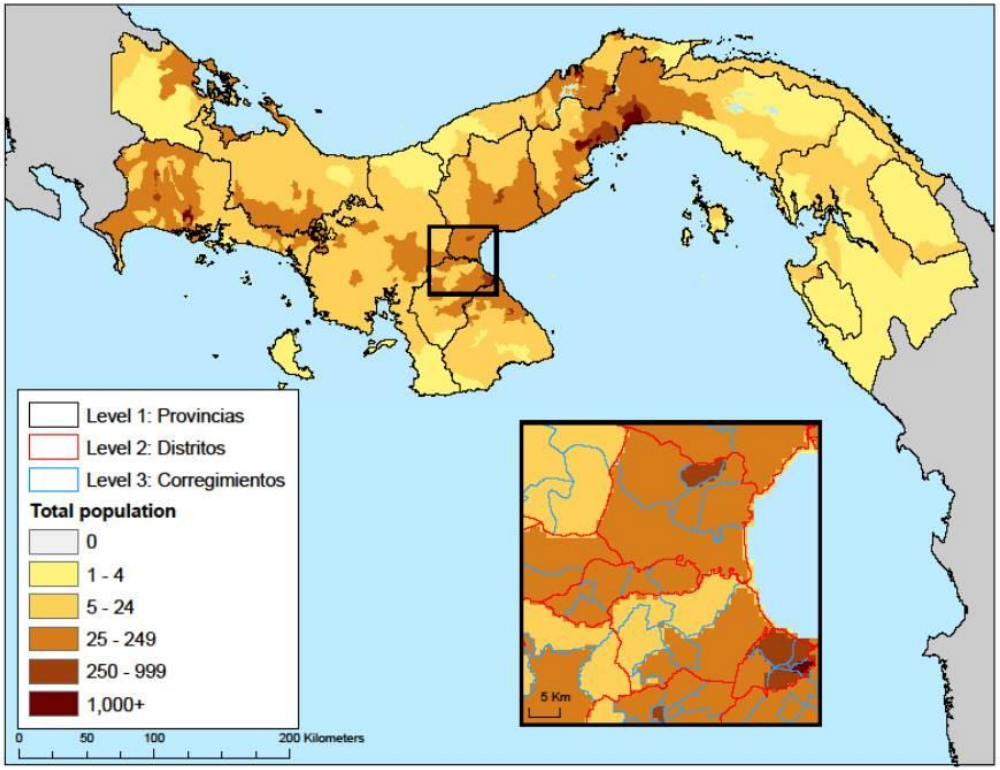
Panama GPWv3 boundaries

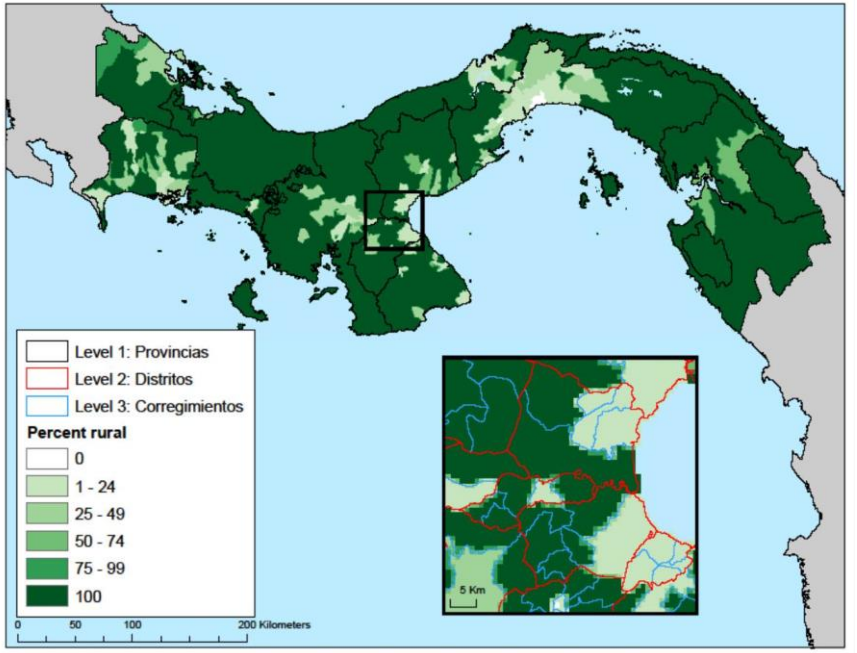


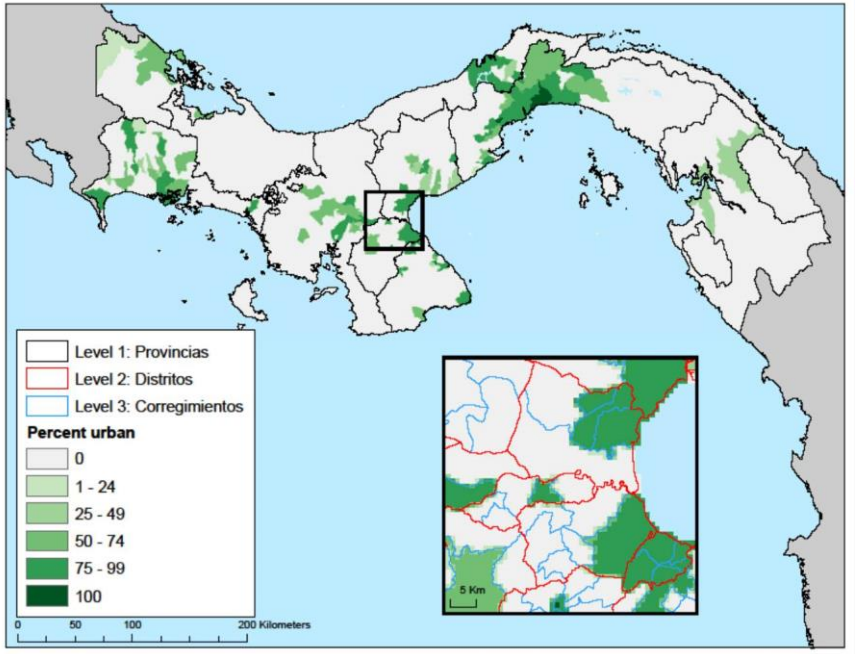
Panama, GPWv4 boundaries



Panama, population distribution grids, 2010







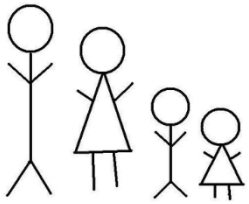
TerraPop Goals

Lower barriers to conducting interdisciplinary human-environment interactions research by making data with different formats from different scientific domains easily interoperable

Provide an organizational and technical framework to preserve, integrate, disseminate, and analyze global-scale spatiotemporal data describing population and the environment.

Terra Populus Data Domains

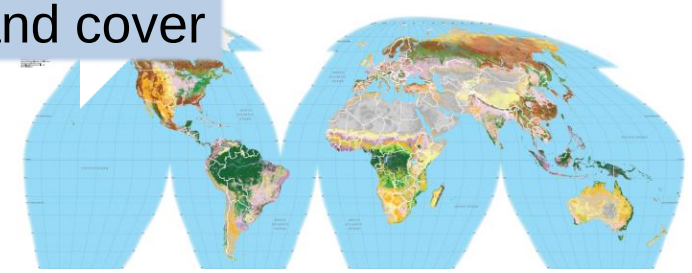
Microdata

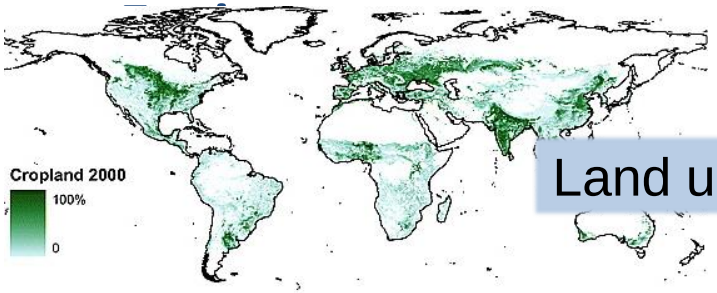


Individuals and households

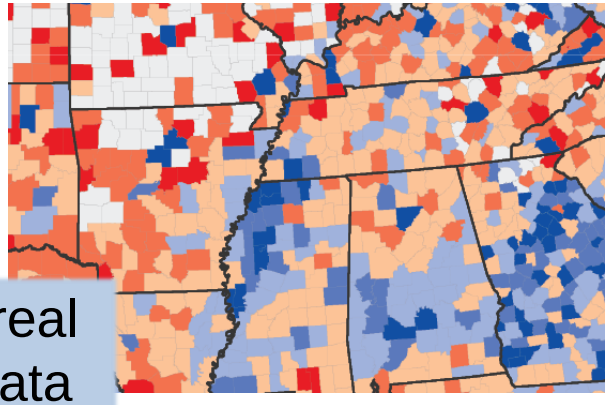
```
H910000240000000088001001000220100
P910000020101032120010010010011504
P910000010201036220010010010011999
P910201000301011220060010010011999
P910201000301009120060010010011999
P910201000301007120060010010011999
P910201000301006120060010010011999
P910201000301004220060010010011999
P910201000301003220060010010011999
P910201000301002220060010010011999
H910000240000000088001001000110100
P910000020101030110010290510511310
P910000010201021210010290290171999
P910201000301001110060010290291999
H910000240000000088001001000220100
P910000020101045120010010010011100
P910000010201025220010010010011820
P910201000301007220060010010011999
H910000240000000088001001000220100
P910000020101049120010010010011100
P910000010201049220010010010011820
P910201000301019220060010010011820
P910201000301015220060010010012820
```

Land cover

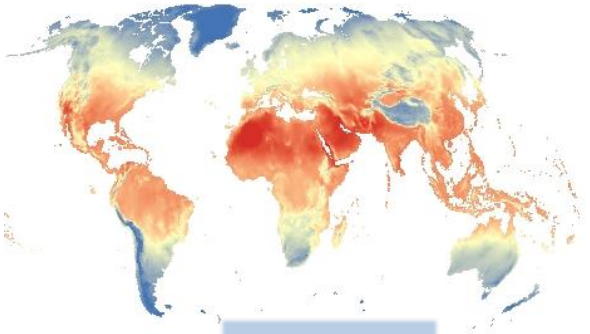




Land use



Areal Data



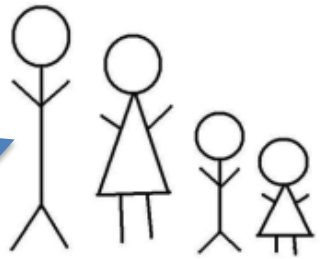
Climate

Area-level Data Sources

- Census tables, especially where microdata is unavailable
- Other types of surveys, data
  - Agricultural surveys
  - Economic surveys, data
  - Election data
  - Disease data
- Legal/policy information

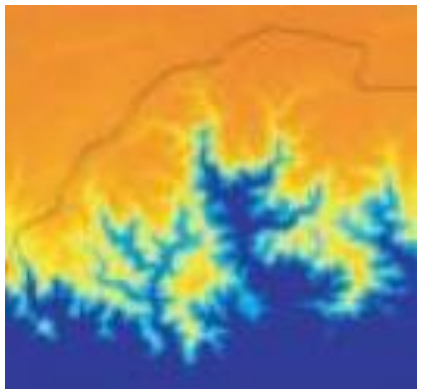
Location-Based Integration

Microdata

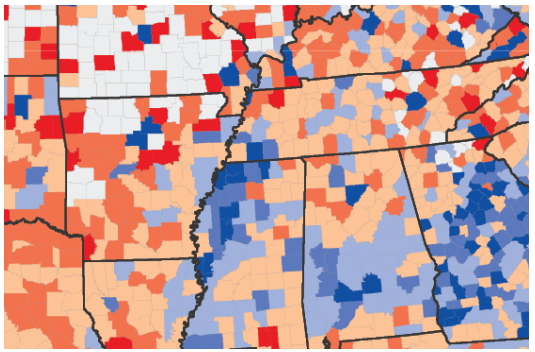


Mix and match variables originating in any of the data structures

Obtain output in the data structure most useful to you



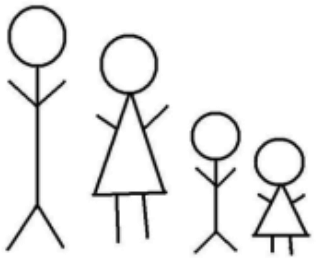
Rasters



Area-level data

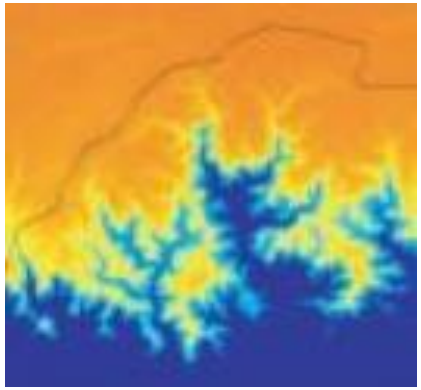
Location-Based Integration

Microdata

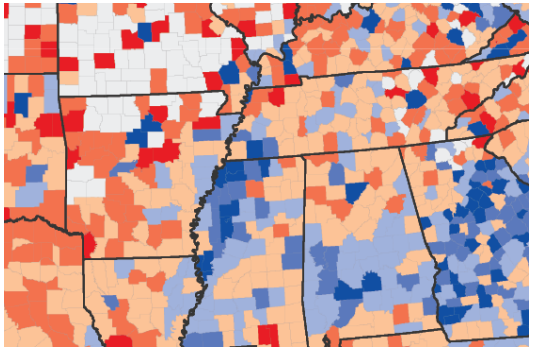


Individuals and households with their environmental and social context

AGE	SEX	LANDCOV	AVGTEMP
10	Male	Forest	21.20
27	Female	Forest	24.30
54	Female	Pasture	24.10
37	Male	Cropped	25.60
37	Female	Cropped	28.10
42	Female	Urban	26.70
20	Female	Forest	24.30
39	Male	Urban	26.80
77	Female	Cropped	27.70
11	Female	Cropped	22.30
31	Female	Pasture	25.10
23	Male	Forest	24.40
24	Female	Urban	21.50
40	Female	Urban	23.40



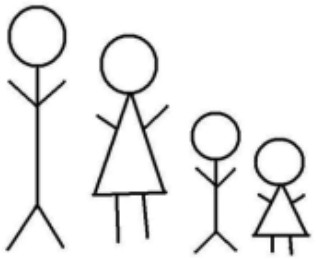
Rasters



Area-level data

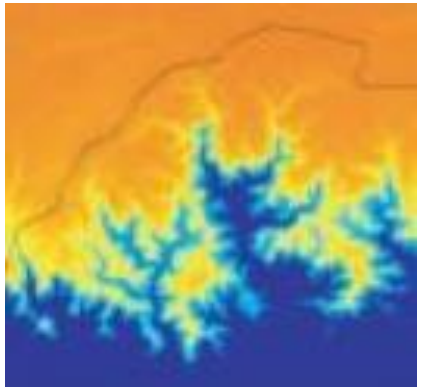
Location-Based Integration

Microdata

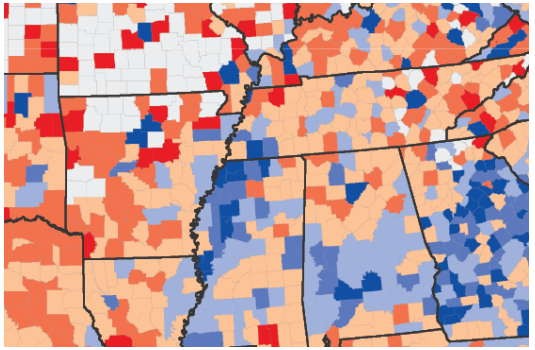


County ID	Mean Ann. Temp.	Max. Ann. Precip.	Rent, Rural	Rent, Urban	Own, Rural	Own, Urban
G17003100001	21.2	768	3129	1063	637	365
G17003100002	23.4	589	2949	1075	1469	717
G17003100003	24.3	867	3418	1589	1108	617
G17003100004	21.5	943	1882	425	202	142
G17003100005	24.1	867	2416	572	426	197
G17003100006	24.4	697	2560	934	950	563
G17003100007	25.6	701	2126	653	321	215

Summarized environmental and population characteristics for administrative districts



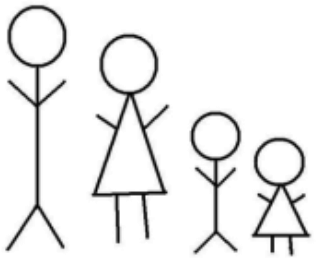
Rasters



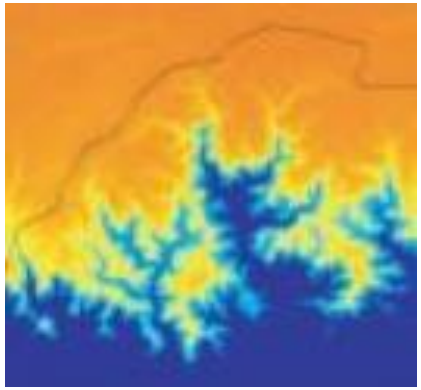
Area-level data

Location-Based Integration

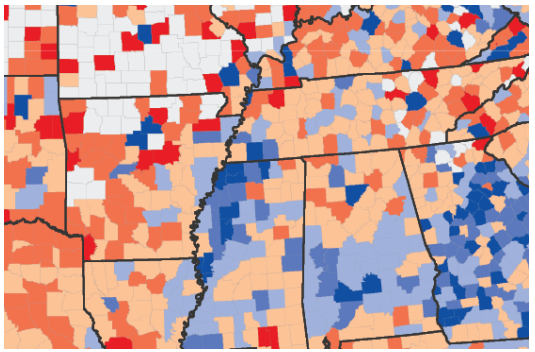
Microdata



Rasters of population and environment data



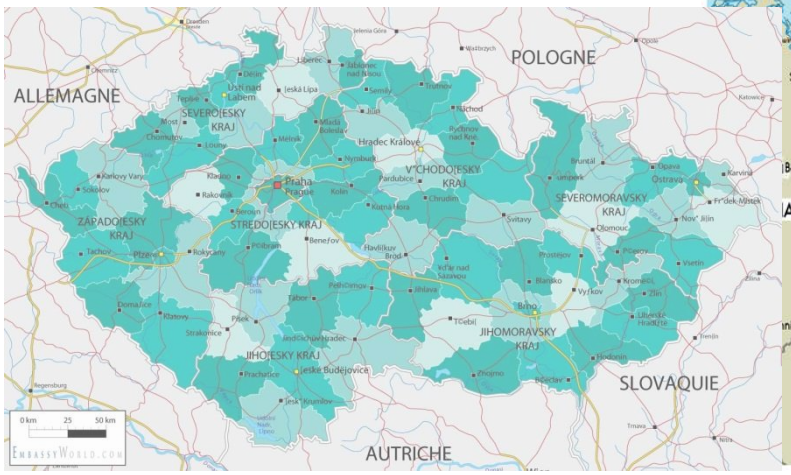
Rasters



Area-level data

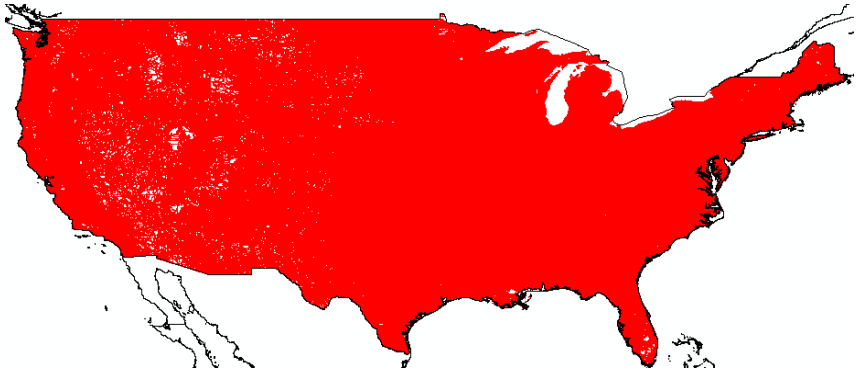
Boundaries are Key

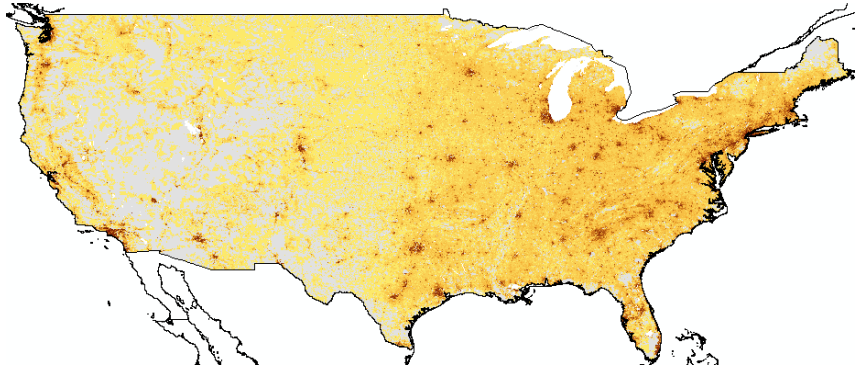
- Linkages across data formats rely on administrative unit boundaries
- Particular needs
  - Higher resolution boundaries
  - Historical boundaries



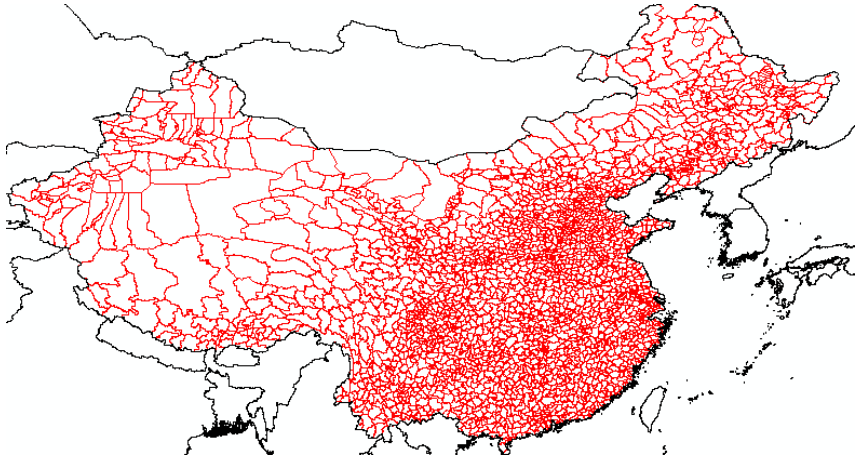
Higher resolution boundaries lead to more accurate population distributions

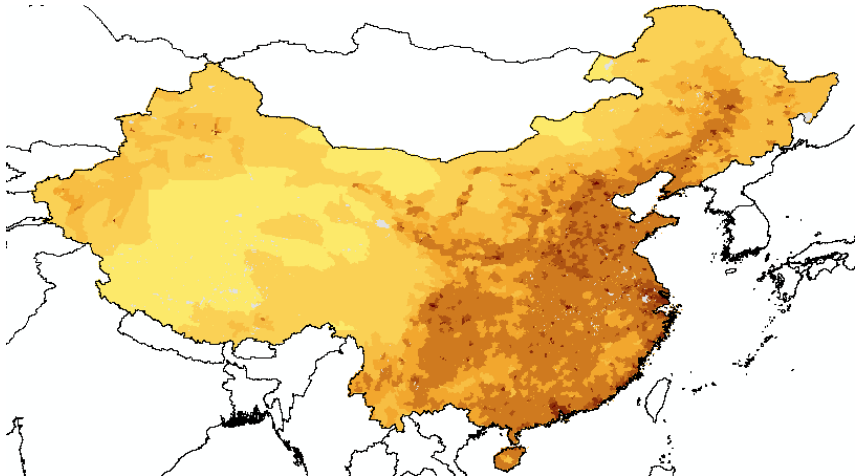
United States: Census Blocks





China: Counties





Linking Census Data to Geographic Data

- Ideally the CSO releases GIS data with Census data with a common identifier for integration
  - More rare than one might think!

- GIS Data from non-Census Source
  - Some common problems
    - Data sources might be from to different points in time
    - Refer to the same unit by different names
    - GIS data might not capture changes in boundaries over time, and therefore requires editing

Linking Census Data to Geographic Data…continued

- GIS Data from non-Census Source
  - Significant effort is needed to reconcile Census areas with those present in the GIS data and to assign common identifiers

- In the worst cases
  - GIS Data at the Census Enumeration Level Cannot be Located
  - Global Dataset is forced to use lower resolution Census Estimates
  - Quality of the product suffers…research in general pays the price

Linking Census Data to Geographic Data…continued

- More complicated when trying to construct time series!
  - *"There is a trade-off between geographical detail and chronological depth so traditionally we can look at how a phenomenon varies over space or how it varies over time but have only very limited capacity to look at both"* (Gregory and Ell 2005)
  - Historical boundaries are often unavailable in digital form
  - Names of units might change over time, with little record of the lineage
  - Significant effort is needed to construct GIS data and to reconcile Census areas
  - At very high resolutions Census units often do not have names, but instead are referred to by numerical codes.

Pleas for a better Geographic Future!

- Many applications (e.g., water resource studies, disaster mitigation and management, food security, assessments of capacity for climate change adaptation) require integrated data for sub-national units **regionally or globally**.

- An effort should be made by national statistical offices to coordinate and cooperate with national geographic data managers to produce integrated products that are both consistent and accurate.

- As changes occur to internal area boundaries, those changes should be tracked and versioned in the geographic data.