Integrating Cross-Domain Data in Terra Populus

Peter Clark  pclark@umn.edu  @pclark
Alex Jokela  ajokela@umn.edu  @alexjokela
Minnesota Population Center, U of Mn
Overview of Terra Populus

• Part of NSF DataNet (Sustainable Digital Data Preservation and Access Network Partners)

• Our 5-year mission:
  
  To boldly seek out, integrate, preserve, and disseminate data describing changes in the human population and environment over time, where no one has done it before!

• Focus for this talk is on data available in TerraPop, and how we’re integrating and linking it

• But first some context...
Computers were first used for counting things, mostly about people...
and for throwing things.
Still used for counting things, mostly about people...
and for throwing things.
Terra Populus is about Data

- Data about population & economies
  - ... things that have been counted, mostly about people
- Data about the planet
  - ... from things that have been thrown
- Data which is location-based
Motivation: Demographic

Source data: Population Reference Bureau
Motivation: Environment

Image credit: United Nations Environmental Programme (via http://climate.nasa.gov/state_of_flux)
Data in TerraPop

- Individual-level Demographic
- Area-level Demographic
- Land use, Climate, and Agricultural
- Boundary Files
- Most of it is treated as Variables
- Geospatially linkable
Microdata

- Individual-level survey response data
  - Usually grouped into households
  - Usually has a couple hundred variables
  - Usually has poor geographic locality
- All censuses tend to ask about similar topics:
  - age, sex, occupation, marital status, dwelling characteristics, nativity and immigration...
Area/Aggregate Data

- Describes characteristics of a location
- Nation, state, county, city, metro area
- Some aggregate characteristics are scalar:
  - Total Population, GDP, % literacy
- More often it's a crosstab or table...
- Each cell in the table is a Variable
<table>
<thead>
<tr>
<th>Subject</th>
<th>Estimate (Total)</th>
<th>Estimate (Male)</th>
<th>Estimate (Female)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>506,399</td>
<td>258,167</td>
<td>248,232</td>
</tr>
<tr>
<td>Less than high school</td>
<td>12.9%</td>
<td>14.2%</td>
<td>11.6%</td>
</tr>
<tr>
<td>High school Graduate</td>
<td>25.8%</td>
<td>28.3%</td>
<td>23.1%</td>
</tr>
<tr>
<td>Some College</td>
<td>50.2%</td>
<td>48.5%</td>
<td>51.9%</td>
</tr>
<tr>
<td>Bachelor’s Degree or Higher</td>
<td>11.2%</td>
<td>9.0%</td>
<td>13.4%</td>
</tr>
</tbody>
</table>

Source data: 2011 American Community Survey
Raster Data

• Each pixel is a measurement
  • Geocoded
  • Usually heavily processed
• Type
  • Categorical or Binary (such as land use)
  • Continuous (temperature, elevation)
• Each file is a **Variable**
  • Or serves as raw material for a constructed variable
Global Land Cover 2000

Integration: Same Domain

- Census/Survey
  - Questions differ
  - Categories differ
  - Encodings differ

- Area-level
  - Structures differ
  - Geographies differ
  - Binning differs

- Raster
  - Alignments differ
  - Resolutions differ
  - Processing differs
  - Projections differ
Linking: Cross-domain

- Semantics and definitions vary
- What does “data” mean?
- What does “linking” mean?
  - For Terra Populus, “linking” means geospatial linking.
Microdata Integration

- Source datasets vary across:
  - Questions/Variables
  - Categories
  - Question Universes
- IPUMS gives us a unified coding scheme
- Produced using translation tables
## Microdata Translation Tables

<table>
<thead>
<tr>
<th>CODE</th>
<th>LABEL</th>
<th>China 1982</th>
<th>Columbia</th>
<th>Kenya 1989</th>
<th>Mexico 1970</th>
<th>USA</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>Single/Never Married</td>
<td>1=never</td>
<td>4=single</td>
<td>1=single</td>
<td>9=single</td>
<td>6=never</td>
</tr>
<tr>
<td>200</td>
<td>Married/In Union</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>210</td>
<td>Married (Not Specified)</td>
<td>2=married</td>
<td>2=married</td>
<td>2=monogamous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>211</td>
<td>Civil</td>
<td></td>
<td></td>
<td></td>
<td>3=only civil</td>
<td></td>
</tr>
<tr>
<td>212</td>
<td>Religious</td>
<td></td>
<td></td>
<td></td>
<td>4=only religious</td>
<td></td>
</tr>
<tr>
<td>213</td>
<td>Civil and Religious</td>
<td></td>
<td></td>
<td></td>
<td>2=civil and religious</td>
<td></td>
</tr>
<tr>
<td>214</td>
<td>Polygamous</td>
<td></td>
<td></td>
<td>3=polygamous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>Consensual Union</td>
<td>1=free union</td>
<td></td>
<td></td>
<td>5=free union</td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>Separated/Divorced</td>
<td></td>
<td></td>
<td>3=sep/divorced</td>
<td></td>
<td></td>
</tr>
<tr>
<td>310</td>
<td>Separated</td>
<td></td>
<td></td>
<td>6=separated</td>
<td>8=separated</td>
<td>3=separated</td>
</tr>
<tr>
<td>330</td>
<td>Divorced</td>
<td>4=divorced</td>
<td></td>
<td>5=divorced</td>
<td>7=divorced</td>
<td>4=divorced</td>
</tr>
<tr>
<td>400</td>
<td>Widowed</td>
<td>3=widowed</td>
<td>5=widowed</td>
<td>4=widowed</td>
<td>6=widowed</td>
<td>5=widowed</td>
</tr>
<tr>
<td>999</td>
<td>Missing/Unknown</td>
<td>0=missing</td>
<td>6=unknown</td>
<td></td>
<td>B=Blank</td>
<td>1=unknown</td>
</tr>
</tbody>
</table>
Area-level Integration

• Geography
  • Changes in boundaries, names, codes
• Table structure
  • Survey Structure
• Binning
Area-level Integration

• We’re cheating: making our own tabulations

• Pros:
  • Can enforce same table structure
  • No metadata hassles!

• Cons:
  • Lose out on fine-grain geography
  • Doesn’t eliminate semantic differences
Raster Metadata and Integration

• Best opportunity for basic things to be easy
  • GeoTIFF, ISO-19115
• Spatial integration and linking is straightforward
• Points of integration:
  • Spatial: projection, origin, resolution
  • Semantics?
  • Across datasets and within a dataset over time
• Often heavily processed, with no metadata standards.
VERY HIGH RESOLUTION INTERPOLATED CLIMATE SURFACES FOR GLOBAL LAND AREAS

ROBERT J. HIJMANS, a, b SUSAN E. CAMERON, a, b JUAN L. PARRA, a PETER G. JONES c and ANDY JARVIS, c, d

a Museum of Vertebrate Zoology, University of California, 3101 Valley Life Sciences Building, Berkeley, CA, USA
b Department of Environmental Science and Policy, University of California, Davis, CA, USA; and Rainforest Cooperative Research Centre, University of Queensland, Australia
c International Center for Tropical Agriculture, Cali, Colombia
d International Plant Genetic Resources Institute, Cali, Colombia

Received 18 November 2004 Revised 25 May 2005 Accepted 6 September 2005

ABSTRACT

We developed interpolated climate surfaces for global land areas (excluding Antarctica) at a spatial resolution of 30 arc s (often referred to as 1-km spatial resolution). The climate elements considered were monthly precipitation and mean, minimum, and maximum temperature. Input data were gathered from a variety of sources and, where possible, were restricted to records from the 1950–2000 period. We used the thin-plate smoothing spline algorithm implemented in the ANUSPLIN package for interpolation, using latitude, longitude, and elevation as independent variables. We quantified uncertainty arising from the input data and the interpolation by mapping weather station density, elevation bias in the weather stations, and elevation variation within grid cells and through data partitioning and cross validation. Elevation bias tended to be negative (stations lower than expected) at high latitudes but positive in the tropics. Uncertainty is highest in mountainous and in poorly sampled areas. Data partitioning showed high uncertainty of the surfaces on isolated islands, e.g. in the Pacific. Aggregating the elevation and climate data to 10 arc min resolution showed an enormous variation within grid cells, illustrating the value of high-resolution surfaces. A comparison with an existing data set at 10 arc min resolution showed overall agreement, but with significant variation in some regions. A comparison with two high-resolution data sets for the United States also identified areas with large local differences, particularly in mountainous areas. Compared to previous global climatologies, ours has the following advantages: the data are at a higher spatial resolution (400 times greater or more); more weather station records were used; improved elevation data were used; and more information about spatial patterns of uncertainty in the data is available. Owing to the overall low density of available climate stations, our surfaces do not capture all variation that may occur at a resolution of 1 km, particularly of precipitation in mountainous areas. In future work, such variation might be captured through knowledge-based methods and inclusion of additional covariates, particularly layers obtained through remote sensing. Copyright 2005 Royal Meteorological Society.
Integration in Terra Populus

- Deal with each domain separately
  - but everything is a **Variable**.
- Linking across domains is **geospatial**
  - Tabulate microdata to area-level
  - Summarize raster to area-level
  - Geospatial integration is relatively easy, given good boundary files
Coming this winter!

*(but see me if you want a demo sooner)*

http://www.terrapop.org

http://www.ipums.org

http://www.nhgis.org

http://www.pop.umn.edu