Problems and Solutions in the Integration of Population Data with Other Disparate Data Sets

Deborah Balk and Gregory Yetman
CIESIN, Columbia University
www.ciesin.columbia.edu

Abstract

When creating databases that cross disciplines, units of analysis are often compartmentalized. This paper examines three different approaches to integration, each of which considers the problems of varying and seemingly incompatible analytic units. We highlight the following issues associated with building, maintaining, and using the database: federally and commercially issued data restrictions, confidentiality, database documentation and metadata, foreign-language translation, and cross-national variable compatibility.

The three approaches differ in scope (national to global), scale (first to fourth-level administrative boundaries), and thematic breadth (single variable to multivariate). These approaches include the creation of (1) a grid-based global database of population (Grided Population of the World), (2) a tool to visualize and export data across contiguous national boundaries (U.S.-Mexico Demographic Data Viewer), and (3) a tool to generate equivalencies between U.S. geographic codes. All three approaches deal with data integration issues at the sub-national level. GPW and GeoCover also facilitate integration of data collected by administrative units with georeferenced biophysical data. The U.S.-Mexico DDViewer contains social, economic, and health behavioral data for three levels of boundaries. The approaches vary in the problems they address, but all are models highly applicable to other themes and scales.

Introduction

Cross-disciplinary analysis is a common activity in the social sciences but some population issues require even broader cross-disciplinary analysis and data integration. Problems such as human-environment interactions and climate-health relationships require multidimensional and multi-disciplinary data integration across different geographic units of analysis and time scales. The needs of any particular study or assessment are framed by the questions at hand:

- The precise combination of data needs may change dynamically and cannot be predetermined
- Pace and requirements of data needs are increasing
- Most existing global data products (e.g., Digital Chart of the World, World Bank indicators) have limitations:
  - small scale or poor resolution (DCW is 1:10,000,000)
  - difficult to keep current
  - sub-national units are rarely available
- "Alternative" (or common demographic) data sources, such as surveys, have selective coverage and confidentiality concerns.

CIESIN has used several approaches for removing barriers to data integration, for example:

1. Start simple! Convert a single variable data set (e.g., population) from a common social science format (census) to a common Earth science format (grid).
2. Move towards open GIS
3. Create tools and methods for converting geographic
   - Create or customize software where existing ones fail and use geographic as a discipline to foster integration.

U.S.-Mexico Demographic Data Viewer (DDViewer)

The U.S. Mexico DDViewer is a free, on-line interactive application that allows thematic mapping of matched variables in the U.S. and Mexico at three levels: region, state and county/municipios. The application is currently being tested and will be released to the public in the spring of 2000. A sample of the U.S. version of the DDViewer is shown below. Matched variable types in the U.S.-Mexico version include:

- Population and age breakdowns
- Birth rates
- Age specific morality data
- Household characteristics
- Education

The viewer allows the user to download the attribute data along with geographic identifiers. In this way, users can perform statistical analysis or combine the attributes with their own data in a GIS.

The integration of micro and multi-level data across national boundaries presents special problems:

- Maintain consistency in or convert thematic variables (e.g., education)
- Select or create a seamless and consistent boundary between countries
- Create variables, tools and documentation in two or more languages
- Obtain adequate and consistent metadata
- Maintain confidentiality (e.g., cause of death data, survey data).

Next Steps:

- Additional data layers
- Environmental data (e.g., pollutant releases)
- Health
- Extended variables to a time series (1980-1995)
- Add functionality to allow spatial analysis.

The U.S. version of DDViewer can be accessed via the web at: http://blue.sedac.ciesin.org/plt/ddviewer/

Once ready, the U.S.-Mexico version will be accessible from CIESIN’s web page at http://www.ciesin.org

Global Population Data

Through gridding of population data or other human variables with the best available data, the integration of population and other variables across national borders can allow for comparability at the sub-national level. Gridded population data also allows for integration with natural science data, which is more often referred to spatial coordinates rather than administrative units. The map shows estimated population densities in 1995 from the preliminary release of GPW version 2.

Conclusion

These examples are only a starting point for data integration efforts. Future integration efforts will focus on leveraging spatial data technologies to provide additional functionality and customized data access. Where traditional integration methods fail, as shown here, geographic can be used to foster dialogue and integration.

Support for this paper was provided by the U.S. National Aeronautics and Space Administration under Contract NAS5-98642. The views expressed are those of the authors and do not necessarily represent the views of CIESIN, Columbia University, or NASA.

GRIDDED POPULATION OF THE WORLD (GPW)

Demographic information is often provided on a national basis, but global environmental studies usually require data that are referenced by latitude and longitude rather than by political or administrative units. To create GPW, sub-national administrative boundary data [levels 1 to 4] and population data from over 200 countries were obtained from government and commercial sources. These data were gridted at the population estimates for each grid cell in GPW. See the figures below for more details on the gridding used for GPW. While there are restrictions on the source data used in creating GPW, the descriptive grids can be freely distributed.

GPW provides the ability to integrate natural or other variables which are stored in a grid or otherwise georeferenced. It has many potential uses, including hazard vulnerability assessment, climate model refinement, and studies of human-induced stress on natural systems.

Next Steps:

- Create new single-variable data sets
- Age distribution
- Income and/or poverty
- Human settlements/urban areas
- Nutrition
- Land use
- Energy Consumption
- Water resources

GLOBAL POPULATION DENSITY

National Level Data, 1990

The simple majority rule depicted above is appropriate for gridding categorical data. Gridding of quantitative data requires that polygon data be distributed across the grid cells that the polygon covers. GPW uses proportional allocation to assign population values to cells where multiple polygons provide input. Diagram adapted from: G.F. Bootham-Carter, 1996 Geographic Information Systems for Conservation.

GLOBAL POPULATION DENSITY

Estimated 1995 Population Density

GRIDDED POPULATION OF THE WORLD


GeoCover is an on-line, interactive service that allows users to select specific geographic layers in the U.S. and generate custom correlation lists. The resulting correlation list, which can be weighed by population, housing units, or land area, gives the proportion of overlap between two geographic layers. "Largographic" capability in GeoCover includes:

- 1990 census geographies (state, county, census tract, place and block, plus 1980 county, place and tract)
- 1990 land use statistics (182* and 105)
- 5-digit ZIP codes
- Public Use Microdata Areas (1% and 5% samples: used for the Census Ling form)
- Metropolitan Statistical Areas (MSAs)
- Hydrologic Unit Codes (watersheds)

GeoCover can enable data integration in many different types of studies. The ability to convert between units associated with socioeconomic data and biophysical units is especially useful. GeoCover can be used to add a tool to maintain confidentiality in survey data, point or small-area data inappropriate for public release can be aggregated to other units with the correlation list provided by GeoCover.

Next Steps:

- Add additional layers in the U.S.
- Add the following additional data layers:
  - Transportation/transportation
  - Climate regions

GeoCover can be accessed on CIESIN’s web site at: http://blue.sedac.ciesin.org/plt/geocover/