HM
Database Developments
HISTORICAL METHODS, April–June 2011, Volume 44, Number 2


Routledge
Taylor & Francis Group

# Harmonizing Disparate Data across Time and Place

## The Integrated Spatio-Temporal Aggregate Data Series

PETRA NOBLE
DAVID VAN RIPER
STEVEN RUGGLES
JONATHAN SCHROEDER
MONTY HINDMAN
*Minnesota Population Center*
*University of Minnesota*



**Abstract.** In this article, the authors describe a new data infrastructure project being developed at the Minnesota Population Center. The Integrated Spatio-Temporal Aggregate Data Series (ISTADS) will make it easier for researchers to use publicly available aggregate data for the United States over a time span that covers virtually the entire life of the nation: 1790–2012. In addition to facilitating access and ease of use, ISTADS will facilitate the use of these various data sets in mapping and spatial analysis.

**Keywords:** aggregate data, census data, data integration, GIS

The Integrated Spatio-Temporal Aggregate Data Series (ISTADS) will create and freely disseminate integrated statistical and geographic data for the United States covering the years between 1790 and 2010. To reduce barriers to research, the project will build an integrated database that will enable researchers to undertake consistent analyses of spatial and temporal variation across thousands of small geographic areas from the first decades of the republic to the present.

ISTADS will build on the substantial investment in the National Historical Geographic Information System (NHGIS).[1] The NHGIS provides—free of charge—aggregate census data and geographic information system (GIS)–compatible boundary files for the United States between 1790 and 2000. NHGIS collected and systematically formatted more than 1 million data files and accompanying metadata, created comprehensive indexes to the data files, and constructed high-quality electronic boundary files describing census tracts, counties, states, and territories.[2]

*Address correspondence to Steven Ruggles, Minnesota Population Center, University of Minnesota, 225 19th Ave. S, Minneapolis, MN 55455, USA. E-mail: ruggl001@hist.umn.edu*

NHGIS makes it comparatively easy for investigators to explore the availability of census data for a particular geographic level in a particular census year and to download tables and accompanying maps. It makes no attempt, however, to provide data that are *consistent* from one census year to the next. Nevertheless, each of the NHGIS components—maps, statistical data, and data access software—provides essential building blocks for an integrated database.

Using these building blocks as its base, ISTADS will provide a new level of research infrastructure, allowing investigators to see specifically what tables and categories are available for each geography over any span of census years. The new Web-based delivery system will provide users merged data files with full documentation of comparability issues. The ISTADS project includes five complementary components:

1. Expansion of the NHGIS spatiotemporal database to include newly available aggregate census data from several sources;
2. Statistical data integration across censuses;
3. Geographic unit integration to maximize crosstemporal comparability;
4. A flexible, robust metadata system design and data model that facilitate identification of comparable data elements across censuses; and
5. A new Web-based interface to data and metadata so that users can easily identify data that are comparable across time, can export multiyear merged data sets suitable for statistical analysis and visualization, and can then incorporate the statistical data and corresponding shape files into a GIS system or statistical package.

**TABLE 1. Source Data for Interuniversity Consortium for Political and Social Research (ISTADS)**

| Dates | Data set description |
|---|---|
| Existing NHGIS data | |
| 1790–1960 | County and state (Haines/ICPSR) |
| 1790–1960 | Municipal data (Haines) |
| 1944–2000 | County and city data books |
| 1974–2000 | County business patterns |
| 1947–2002 | Economic censuses |
| 1949–2002 | Agricultural censuses |
| 1940–50 | Census-tract data (Bogue/Beveridge) |
| 1910–30 | Supplemental tract data (Beveridge) |
| 1970 | Census small area data |
| 1980 | Census summary files |
| 1990 | Census summary files |
| 2000 | Census summary files |
| Additional files | |
| 1880 | Census summary files |
| 1960 | Census summary files |
| 1900–2012 | Vital statistics |
| 1920–50 | Print conversion of select tables |
| 2005–12 | ACS files |
| 2010 | Census summary files |

*Note*. ACS = American Community Survey; NHGIS = National Historical Geographic Information System.

## Expansion of the NHGIS Spatiotemporal Database

The principal data sources for ISTADS are shown in table 1. Most of the data for ISTADS are already incorporated into NHGIS. We compiled these data from a variety of sources (ranging from census CDs to handwritten manuscripts in municipal archives) and they represent the most comprehensive collection of U.S. aggregate statistics available anywhere. We will supplement NHGIS aggregate data with newly available summary data from several sources, including the 1880, 1960, and 2010 censuses,[3] print sources, and the American Community Survey. We will also incorporate a complete set of county-level vital statistics into the system. These new sources will expand the size of NHGIS by almost 50 percent.

The 1880 and 1960 census summary files leverage microdata projects housed at the Minnesota Population Center and will supplement the existing census data available for these two decades. Other new data will be provided from conversion of printed tables and vital statistics. We will digitize newly recovered data sets to fill significant gaps in topical censuses in the NHGIS, including the agricultural censuses of 1930 and 1940, the censuses of wholesale and retail trade, and the censuses of governments. The inclusion of mortality data from 1900 onward significantly enhances the demographic analytical power of the data set. These data cover both infants and mothers and include variables for cause of death, age, sex, race, nativity, and month of death, for counties, cities, states, and registration areas. ISTADS will also include data on births, infant deaths, and stillbirths by age of mother, race, nativity, month, and birth order for the same geographic units beginning in 1915. Finally, the American Community Survey, the 2010 census, and other updates will be incorporated along with updates of the county and city data book, county business patterns, the economic census, and the agricultural census.

Collecting this vast array of aggregate data is just the first step. To make it feasible for researchers to analyze these data across time and space, the data must be integrated in a manner that minimizes loss of detail information.

## Statistical Data Integration

The integration of statistical data in ISTADS is a multistep process. First, we will construct an integrated aggregate statistical series using category aggregation and estimation techniques to improve crosstemporal comparability. Table-level metadata will be the primary means for defining and describing any transformations needed to make data elements comparable for two or more census years. We will pursue two methods for creating the time series data—aggregation and imputation.

### *Statistical Data Aggregation*

We will produce time series tables that have closely comparable categories in different census years. These time series tables will be accompanied by documentation that explains differences that persist in the final integrated tables. Many tables are directly comparable from year to year. For example, a total population table delineated by "sex"—with values of "male" and "female"—will remain constant and is easily comparable from census year to census year. Many tables, however, will require aggregation of categories to make them comparable. For example, the age distribution for counties is available for 10-year age groups in 1830 through 1860 and 5-year age groups from 1930 onward. ISTADS aggregates the 5-year age categories to produce tables with 10-year age groups for all census years and also provides 5-year age groups for the more recent census years. Thus, users who need a long time series for their analyses may select the 10-year categories, and users who need shorter age categories may select the 5-year categories.

### *Statistical Data Imputation*

In some cases, it will be necessary to impute estimates to construct comparable categories over time. Consider, for example, the age classification for school enrollment shown in table 2. Using the previously described aggregation method, ISTADS will provide the classification of lowest common denominator between 1970 and 1990 (labeled "common" in

**TABLE 2. Age Classifications for School Enrollment**

| 1970 | 1990 | Common | Imputed |
|---|---|---|---|
| 3–4 | 3–4 | 3–4 | 3–4 |
| 5–6 | 5–6 | 5–6 | 5–6 |
| 7–13 | 7–9 | 7–17 | 7–13 |
| 14–15 | 10–14 | | 14–15 |
| 16–17 | 15–17 | | 16–17 |

table 2). However, simple aggregation imposes a heavy cost in terms of lost detail—and lost research capability. In the age example, the final three categories in both years must be combined to show school enrollment for ages 7–17 years; this forecloses the opportunity to compare changes in high-school dropout rates over time. To address this problem, ISTADS will also provide a version of the aggregated table that imputes school attendance for compatible age categories. With school attendance imputed within each geographic area in 1990 for ages 14 and 15 years, data for the two censuses will be comparable. A more detailed description of the ISTADS imputation approach is included in the appendix.

Where possible, we will validate imputation procedures by applying them to cases where we know the answer. For example, we could test the ISTADS imputation procedure by applying it to a census year in which we know actual school attendance for small areas by single years of age.[4] In all cases, we will include standard errors for all imputed values, and provide full details of imputation procedures in the table description. Because we will provide both raw and imputed data, researchers will be able to carry out sensitivity analyses to assess how robust their estimates are under different approaches.

## Geographic Unit Integration

The geographic integration component will identify geographic units that remain consistent across each span of census years, create interpolated estimates of population characteristics to account for boundary changes, and produce generalized boundary files for census tracts and counties that are entirely compatible across census years. Specifically, we propose to produce integrated geographic files for counties, metropolitan areas, places, and census tracts.

ISTADS will pursue two general strategies for geographic integration. First, we will aggregate geographic units to create new geographic levels in which boundaries match across census years. Second, we will use interpolation to develop consistent tract data for 1970–2010. We will also improve the crosstemporal compatibility of the existing boundary files by developing consistent generalized boundaries for the entire 1790–2010 period.

## Geographic Aggregation

Changes to the boundaries of administrative units (e.g., counties, places) and, even more so, statistical units (e.g., census tracts, block groups) inhibit the temporal analysis of census data. The problem is that we cannot know whether changes in some census variables are due to actual changes in the characteristic reported or to changes in the boundary of the geographic unit. To remove this impact of boundary change on census variables, ISTADS will use aggregation techniques to create consistent, time-invariant geographic units across census years.

### *Supertracts*

The census tract was originally conceived by Dr. Walter Laidlaw as a small geographic area that could be used to study neighborhoods in New York City (Bowser, Mann, and Oling 1981). From 1910 to 1930, the number of cities that adopted census tracts grew from 8 to 18. When the U.S. Census Bureau officially adopted the census tract in 1940, 73 counties participated in the program. Tract coverage grew in each successive decade, and by 1990, the entire country was divided into tracts. The idea of a *supertract* was used in New York City in 1940, 1950, and 1960 "for the systematic and continuous consideration of local health needs" (ibid., 19). These supertracts were aggregations of existing tracts, which coincided with New York sanitation districts. ISTADS will provide similar supertracts that will allow consistent comparisons across census years for the entire country.

We will construct supertracts for a variety of time spans, such as 1990–2000, 1980–2000, and 1960–70. Creating a wide variety of time spans provides scholars with data that best fits their research needs. The supertract construction process consists of the following steps: (1) overlay the tracts for the time span in question; (2) identify tracts that are identical; (3) for tracts that changed over time, merge the areas of intersection until consistent boundaries are found; and (4) create a correspondence file between supertracts and their component tracts so that aggregate data may be created.

Many census tracts have remained relatively consistent over time. In many cases, the supertracts will be almost as numerous as the original tracts. For example, St. Paul, Minnesota, had a maximum of 77 tracts from 1930 to 1960, and we can identify 73 supertracts that are completely consistent for that entire period (see figure 1). In Minnesota as a whole, the 2000 census identifies 1,300 tracts, and we can identify 1,004 Minnesota supertracts that are consistent between 1990 and 2000.

The work for aggregation of features will be automated through spatial algorithms. For quality assurance, these algorithms will be tested using manually processed data sets from 1980, 1990, and 2000. In addition, a sample of results will be manually reviewed for quality assurance.

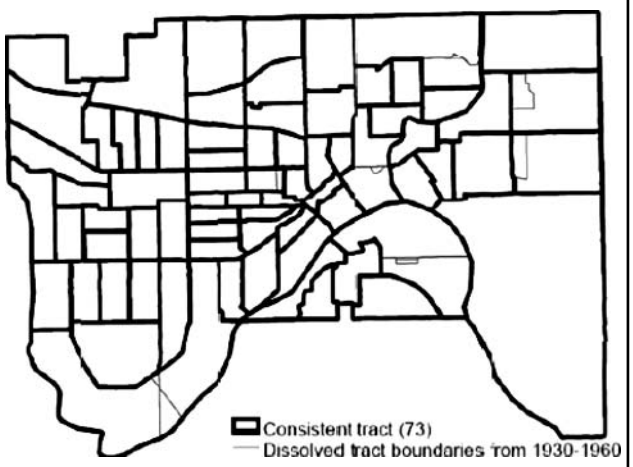


**FIGURE 1. Supertracts in Saint Paul, Minnesota, 1930–60.**

*Aggregated Counties*

The methodology for producing aggregated county boundaries is essentially the same as that for tracts. County boundaries have for the most part remained stable since 1920, so the greatest challenges pertain to the nineteenth century and the early twentieth century. Many recent county boundary changes involve small areas or populations. When a change involves less than 2.5 percent of county area, we will document the change but will not correct it; for larger changes, we will aggregate counties until they meet the threshold.

*Time-Sensitive Metropolitan Areas and Cities*

Consistent counties will serve as the basis for consistent metropolitan areas. Since 1940, the census has defined metropolitan areas as a combination of counties. We will allow users to select metropolitan-area boundaries from a particular period and apply them consistently to all periods. Thus, for example, a user will be able to select 1970 Metropolitan Area definitions and generate statistics showing what the racial and ethnic composition of each metropolitan area would have been between 1950 and 2010 under the 1970 definition. Similarly, we will construct consistent cities by using supertracts as building blocks. This will allow investigators to assess how the demography of cities would be different had there been no annexations.

## Geographic Interpolation

Aggregating areal units into different configurations may introduce statistical bias into analysis through the modifiable areal unit problem (MAUP). MAUP arises when the units of analysis are arbitrary and modifiable and can yield false positives and false negatives (Openshaw 1984). We try to minimize MAUP by aggregating only units that overlap one another. This process creates the maximum possible number of supertracts. In some localities, however, wholesale retracting can yield unacceptably large supertracts. Moreover, even a modest degree of aggregation has costs in terms of lost geographic detail. Supertracts are not a perfect solution to the problem of shifting boundaries.

**TABLE 3. Interpolated Aggregate Data Statistics with Corresponding Census-Tract Geographic Information System (GIS) File**

| | Census-tract GIS files | | | | |
|---|---|---|---|---|---|
| Interpolated aggregate data | 2010 | 2000 | 1990 | 1980 | 1970 |
| 2010 | — | x | x | x | x |
| 2000 | x | — | x | x | x |
| 1990 | x | — | — | x | x |
| 1980 | — | — | — | — | x |

*Note.* x = combination for which interpolated data will be available.

As an alternative to the supertracts for more recent censuses, ISTADS will employ interpolation techniques to estimate the population characteristics of one census for the geographic units of another census. ISTADS will produce five sets of high-quality interpolated data for census tracts (table 3):

1. 2000 and 1990 census-tract aggregate data onto 2010 census-tract geography;
2. 2010 census-tract aggregate data onto 2000 census-tract geography;
3. 2010 and 2000 census-tract aggregate data onto 1990 census-tract geography;
4. 2010, 2000, and 1990 census-tract aggregate data onto 1980 census-tract geography; and
5. 2010, 2000, 1990, and 1980 census-tract aggregate data onto 1970 census-tract geography.

To avoid the assumption of population homogeneity, ISTADS will use characteristics of blocks and block groups, such as race, to estimate tract characteristics. However, there are no consistent digital block and block-group boundaries for censuses before 1990. We therefore do not have sufficient data to estimate the characteristics of the population that resided within current tract boundaries in the 1970 or 1980 censuses with sufficient accuracy for most research purposes. For those years, ISTADS uses land-use data to help guide the interpolation. The USGS provides digital Geographic Information and Analysis System (GIRAS) land-use data from the early 1980s that can help guide the interpolation of data from 1980 to the 1970 census tracts.

In addition to providing more precise estimates of tract characteristics, we will address a second major limitation of

existing interpolated data. We are persuaded by Gregory and Ell (2006), who argued that that error estimates are essential to help users assess the validity of interpolated data. Accordingly, ISTADS will provide indicators of uncertainty for each data value using techniques identified by Jonathan Schroeder (2007).

### A Flexible, Robust Metadata System and Data Model

The core of the entire ISTADS project is the development of a data management system that allows the project to refine, retrieve, enhance, and supplement all existing and future NHGIS metadata files. Metadata are formally structured documentation of digital data. Like NHGIS, the ISTADS metadata format will be interoperable with the Data Documentation Initiative 3.0 (DDI)[5] metadata model, and we will be able to generate DDI-compliant codebooks for data sets on demand. Unlike NHGIS, which created machine-readable HTML files that captured all data set, table, and geographic information within a census year, ISTADS will use a relational database to store all metadata elements.

The HTML format used in the original NHGIS, though it documented all data set elements within one census year well, failed to facilitate easy comparison of data among census years. Without migrating from HTML files to a relational database, it would be impossible to harmonize geographic levels, tables, and categories from multiple years simultaneously. A relational database helps facilitate cleaning, correcting, and establishing relationships among all of the data sets. Also a number of checks for data consistency, missing data, and other data errors can now more easily be performed.

The ISTADS' data model will be robust enough to capture the necessary elements from all of the original source codebooks and documentation; allow for comparison of both the statistical and geographic data; and capture all other ancillary documentation related to the original source tables, research decisions, and other supplemental information. These metadata will ultimately drive all software for data preparation, data conversion, and dissemination.

Metadata are not confined solely to census data sets. ISTADS will bring together supplemental documentation to form a comprehensive collection of reference materials. These materials will provide full detail on geographic, occupational, and industrial coding schemes; the construction of poverty indices and other derived variables over time; methodological papers; census questionnaire and product development reports; sample designs and sampling errors; procedural histories of each data set; full documentation of error correction and other postenumeration processing; and analyses of data quality.[6] All documents will be converted to the appropriate format for representation within the ISTADS data model. We will document our own procedures and data transformations as we proceed.

### Web-Based Interface to Data and Metadata

The new Web-based interface to the ISTADS statistical data, geographic data, and metadata will enable users to easily identify data that are comparable across time and space and create multiyear merged data sets suitable for statistical analysis and visualization. The existing NHGIS system uses a drill-down approach in which users locate tables on a particular topic from a particular census year at a particular geographic level (e.g., tract, county, state). Unfortunately, the existing system provides no mechanism for browsing the data available in multiple censuses; making it difficult to identify variables with consistent definitions through time. To capitalize on the crosstemporal power of ISTADS, an entirely new data dissemination system will incorporate a new user interface and data retrieval system, along with generalized boundary files.

The new user interface will allow a researcher to retrieve data from multiple years, topics, and geography levels at the same time. Once a user has defined a particular combination of data elements, the system extracts the requested data. The ISTADS Web application will communicate through an application-programming interface (API), with the data retrieval system (the so-called "back end") providing access to aggregate statistical data, spatial data, and metadata. By handing the extract request to the data retrieval system, the Web application will stay responsive to users and allow them to browse the vast data holdings with little lag time. Users will be able to download aggregate data in comma-separated value files or fixed width files accompanied by command files for SAS, Stata, and SPSS. All aggregate data downloads will also come with a codebook describing the extract data file contents. GIS-compatible boundary files will be distributed as shape files.

Digital geographic boundary files vary greatly in levels of detail. For example, to create a map showing all tracts in Massachusetts with the detail required to show the small islands in tracts on Boston Harbor would impair the map's clarity and greatly increase the size of the electronic file. To produce maps at smaller scales, we must reduce the detail—that is, *generalize* the data—by simplifying boundaries and eliminating small features. An important benefit of generalization is that it reduces boundary-file sizes significantly, making data downloads and data processing much faster.[7] The generalization routine will maintain shared boundaries between types of geographic units to enforce hierarchy. ISTADS will provide a user with multiple scales for census tracts, counties, metropolitan areas, and states.

### Status of the ISTADS Project

The various project components described are proceeding on separate but coordinated tracks. Much of the effort during the first project year was devoted to designing metadata, developing software, and refining efficient work processes

for statistical and geographic integration. Full-scale production of integrated tables, aggregated geographic units, and interpolated tracts began in June 2009. In 2011, we will release a basic system. This release will include a subset of integrated tables that require only limited manipulation to make them compatible, and the data will be provided for nonintegrated geographies and supertracts. In 2012, we will add integrated tables that do not require extensive manipulation and will provide consistent counties for all data in the two releases. The final release—which will include all the integrated tables and geographies and the full complement of web dissemination features—will occur in June 2013.

The ready availability of integrated aggregate census data in a GIS framework will offer opportunities to address a broad range of social science research problems. Key areas include residential segregation; the decline and renaissance of central cities; immigrant and ethnic settlement patterns; suburbanization and urban sprawl; rural depopulation and agricultural consolidation; the identification of concentrated poverty; causes and levels of change in ecosystems; transportation; the transformation of electoral politics; criminal justice; and environmental justice. The incorporation of consistent, detailed fertility and mortality statistics for the period since 1900 will open unprecedented new opportunities for investigating the long-run spatial correlates of demographic change. Finally, the ISTADS data will—for the first time—enable researchers to easily incorporate both aggregate and microlevel data in multilevel, spatial analysis.



### NOTES

The Integrated Spatio-Temporal Aggregate Data Series project at the Minnesota Population Center, University of Minnesota, is supported by a grant from the National Institutes of Health, 1RO1 HD057929.

1. The NHGIS includes an extraordinary body of statistical information and documentation, including approximately 3 terabytes of statistical data accompanied by 5 million lines of XML-encoded codebooks, 100 gigabytes of indices to the data, and more than 200,000 polygons describing historical census geography.
2. Boundary files for states/territories and counties are available for the entire time series of 1790 to the present. Census-tract boundary files are available from 1910 onward. The NHGIS also provides electronic high-quality boundary files for other units of census geography, such as congressional districts, zip code tabulation areas, voting districts, trust land, standard consolidated statistical areas, places, minor civil divisions, county subdivisions, blocks, and block groups. For the 1990 and 2000 censuses, specialized geography boundary files are provided for American Indian areas, Alaska native areas, and Hawaiian homelands. For metropolitan areas, boundary files are available for urbanized areas, New England county metropolitan areas, and—for the period 1950–2000, metropolitan statistical areas (called "standard metropolitan statistical areas" in pre-1950 data).
3. In the case of 1960 census and American Community Survey, conversion of the data to NHGIS format will be supported by other grants, NIHHD041575 and NSF BCS-0648005, respectively.
4. The necessary data in this instance would include an aggregate file with school attendance data by single years of age paired with a microdata file from the same census that identifies PUMAs. Although such data do not currently exist, we could potentially construct them for 1960 as part of the census recovery project.
5. The DDI is described at http://www.icpsr.umich.edu/DDI/.
6. The IPUMS project (Perlmann 2003) has already compiled most of these materials and is now converting them to marked-up structured documents that make machine-processing comparatively easy. We plan to make this part of the metadata system fully interoperable between the two projects, so we can create and maintain a single set of documents that will serve both the microdata and aggregate data series.
7. The level of detail in the TIGER/Line data (U.S. Census Bureau 2008), which is the basis of NHGIS files, is generally appropriate for maps drawn at scales from about 1:50,000 to 1:100,000. Most visualizations of NHGIS data will occur at smaller scales. For visualization at smaller scales, the level of detail of NHGIS data is too great, yielding undesirable graphical effects and unnecessarily large file sizes.

### REFERENCES

Bowser, B. P., E. S. Mann, and M. Oling. 1981. *Census data with maps for small areas of New York City 1910–1960: A guide to the microfilm*. Ithaca, NY: Cornell University Libraries.

Flowerdew, R., and M. Green. 1994. Areal interpolation and types of data. In *Spatial analysis and GIS*, ed. S. Fotheringham and P. Rogerson, 123–46. London: Taylor & Francis.

Gregory, A., M. Vardigan, S. Ionescu, A. Green, J. Gager, and W. Thomas (eds). 2007. *Data documentation initiative (DDI) technical specification: Part 4A. Field-level documentation (Version 3.0) for public review.* Ann Arbor, MI: DDI Alliance.

Gregory, I. N., and P. S. Ell. 2005. Breaking the boundaries: Integrating 200 years of the Census using GIS. *Journal of the Royal Statistical Society, Series A* 168:419–37.

Langford, M., D. Maguire, and D. J. Unwin. 1991. The areal interpolation problem: Estimating population using remote sensing in a GIS framework. In *Handling geographical information: Methodology and potential applications*, ed. I. Masser and M. Blakemore, 55–77. London: Longman.

Maiti, T. 2001. Robust generalized linear mixed models for small area estimation. *Journal of Statistical Planning and Inference* 98(1–2):225–38.

National Research Council. 2000. *Small area income and poverty estimates: Priorities for 2000 and beyond.* Washington, DC: Committee on National Statistics, National Academy of Science.

Openshaw, S. 1984. *The modifiable areal unit problem.* Norwich, England: GeoBooks.

Perlmann, J. 2003. IPUMS. *Journal of American History* 90:339–40.

Schroeder, J. 2007. Target-density weighting interpolation and uncertainty evaluation for temporal analysis of census data. *Geographical Analysis* 39:311–35.

U.S. Census Bureau. 2008. 2007 TIGER/Line® *Shapefiles [machine-readable data files*]. U.S. Census Bureau: Washington, DC.

## APPENDIX

### The ISTADS Approach to Imputation

The ISTADS approach to imputation builds on prior models of small-area estimation that combine available census microdata with characteristics of the specific geography taken from the existing tabular data (Buskirk and Meza 2003; Chand and Alexander 1999; Chand and Malec 2001; Fisher and Campbell 2002; Fisher and Turner 2003; Jiang 2003; Jiang, Jia, and Chen 2001; Larsen 2003; Legler, Breen, Meissner, Malec, and Coyne 2002; Malec, Davis, and Cao 1997; Malec, Davis, and Cao 1999; Maiti 2001).

Microdata can tell us the relationship of school attendance at particular ages to school attendance in broader age groups, with control for other relevant community characteristics. For example, to estimate school attendance at age 14 years in 1990, we generate data on school attendance and local characteristics for public-use microdata areas (PUMAs). We can then regress percentage of school attendance at age 14 years on school attendance for age groups 10–14 and 15–17 years, as well as the percentage of black, percentage in poverty, median income, employment, or any other ecological characteristics strongly associated with school attendance. Once we

have estimated the model, we can simply plug in the characteristics of each particular tract, county, or other geographic unit to obtain a preliminary estimate of the percentage of 14-year-olds attending school in that geographic area. We then repeat the procedure for 15-year-olds. This strategy makes no assumptions about homogeneity within PUMAs. Rather, we use variation across PUMAs to assess the relationship of aggregate-level characteristics—including school attendance for age groups—with school attendance at a specific age.

By combining the regression estimates of school attendance with the number of persons in each geographic unit who are age 14 years and who are age 15 years, we can estimate the numbers in school and out of school at each age. This will give us sufficient information to impute school attendance for each geographic unit for persons aged 7–13, 14–15, and 16–17 years. In a final step, we will adjust the data to match control totals, in this instance the overall enrollment in the geographic unit from age 7 years to age 17 years. By limiting the estimation to persons aged 14 years and 15 years—ages with minimal geographic variation in school attendance—we minimize the potential for introducing significant error (Smith 1989; National Research Council 1980, 2000; Ghosh and Rao 1994; Pfeffermann 2002).



## REFERENCES


Buskirk, T. D., and J. L. Meza. 2003. A post-stratified ranking-ratio estimator linking national and state survey data for estimating drug use. *Journal of Official Statistics* 19:237–52.

Chand, N., and C. H. Alexander. 1999. Using administrative records for small area estimation in the American Community Survey. Federal Committee on Statistical Methodology, Federal Committee on Statistical Methodology. http://fcsm.gov/01papers/Chand.pdf.

Chand, N., and D. Malec. 2001. Small area estimates from the American Community Survey using a housing unit model. Federal Committee on Statistical Methodology. http://www.fcsm.gov/01_papers/Chand.pdf

Fisher, R., and J. Campbell. 2002. Health insurance estimates for states. Paper presented at the Government Statistics Section of the American Statistical Association Annual Meeting, New York, NY. August 11–15.

Fisher, R., and J. Turner. 2003. Health insurance estimates for counties. Paper presented at the Government Statistics Section of the American Statistical Association Annual Meeting, San Francisco, CA. August 3–7.

Ghosh, M., and J. N. K. Rao. 1994. Mall area estimation: An appraisal. *Statistical Science* 9:55–76.

Jiang, J., H. Jia, and H. Chen. 2001. Maximum posterior estimation of random effects in generalized linear mixed models. *Statistica Sinica* 11:87–120.

Jiang, J. M. 2003. Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of Statistical Planning and Inference* 111:117–27.

Larsen, M. D. 2003. Estimation of small-area proportions using covariates and survey data. *Journal of Statistical Planning and Inference* 112:89–98.

Legler, J., N. Breen, H. Meissner, D. Malec, and C. Coyne. 2002. Predicting patterns of mammography use: A geographic perspective on national needs for intervention research. *Health Services Research* 37:929–47.

Malec, D., W. W. Davis, and X. Cao. 1999. Model-based small area estimates of overweight prevalence using sample selection adjustment. *Statistics in Medicine* 18:189–200.

Malec, D., J. Sedransk, C. Moriarty, and F. B. LeClere. 1997. Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association* 92:825–26.

National Research Council. 1980. *Estimating population and income of small areas.* Washington, DC: National Academy Press.

Pfeffermann, D. 2002. Small area estimation: New developments and directions. *International Statistical Review* 70:125–43.

Smith, T. M. F. 1989. Aggregated analysis: Point estimation and bias. In *Analysis of complex surveys*, ed. C. J. Skinner, D. Holt, and T. M. F. Smith, 135–47. New York: Wiley.