Building the National Historical Geographic Information System

CATHERINE A. FITCH
STEVEN RUGGLES
Minnesota Population Center
University of Minnesota

Abstract. The National Historical Geographic Information System (NHGIS) is a new project to make a rich body of aggregate census data accessible within a Geographic Information Systems (GIS) framework for historical population research. The authors are developing a database incorporating all available aggregate census information for the United States between 1790 and 2000, including all surviving machine-readable aggregate data and new data transcribed from printed and manuscript sources. They are also creating new census-tract maps back to 1910, state and county maps back to 1790, and additional maps when feasible. Availability of high-quality boundaries for key statistical areas will permit the reconciliation of changes in census geography. Census data, documentation, and boundary files will be freely disseminated through an integrated Web-based data access and mapping system.

Keywords: aggregate data, census, Geographic Information Systems (GIS), maps, National Historical Geographic Information System (NHGIS), population

The census is the primary source of statistical information about the growth and change in the U.S. population since 1790. Aggregate data tables, in print or electronic format, are the principal means of describing the characteristics of states, metropolitan areas, cities, counties, minor civil divisions (MCDs), and neighborhoods. Approximately 670 gigabytes of U.S. census summary data covering the period 1790 through 2000 exist or are in preparation, but they are scattered across dozens of archives and are stored in incompatible formats on magnetic tape, CD-ROMs, or on paper.

Even if this massive body of aggregate census data were readily accessible, historical analysis would be complicated by changes in the geographic units they describe. Summary counts of the population characteristics of places are meaningful only if those places are clearly defined. In every census year, the boundaries of the geographic units described by the census have been modified. Modification of boundaries makes the analysis of geographic change in the U.S. population exceedingly difficult. When researchers do assess change, they must develop their own maps at great expense. It is not surprising that the number of such studies is small and their geographic and chronological scope is limited (Duncan 1957; Lieberson 1961; Denton and Massey 1991; Massey and Eggers 1990; Massey and Denton 1988; Alba et al. 1995; Logan et al. 1996). To allow systematic analysis of change over time, we need a compatible set of electronic maps that describe the location of each geographic unit tabulated by the census. At present, however, such maps exist only for the 1990 and 2000 census years.

The National Historical Geographic Information System (NHGIS) is designed to make the rich body of aggregate census data accessible within a Geographic Information Systems (GIS) framework for historical population research. The five-year project began in April 2001 with funding from the National Science Foundation Social Science Infrastructure Program (BCS 0094908). The goals are to gather together all surviving census data from 1790 to 2000, format them consistently, develop comprehensive standardized machine-readable documentation, create high-precision historical electronic boundary files describing census tracts and counties, and develop innovative Web-based tools for disseminating both microdata and metadata over the Internet.

The proposed database is very large. It will include almost the same quantity of data as presently exist in the entire archive of the Inter-university Consortium for Political and Social Research (ICPSR)—the world's largest social science data archive. This scale of infrastructure would have been unthinkable just a few years ago. Five recent technological innovations make the NHGIS feasible:

1. The decline in the cost of data storage during the past five years makes it possible to maintain the entire body of machine-readable census data online.
2. The development of the Internet has slashed the cost of worldwide dissemination.
3. The development of GIS technology has provided the tools to describe and display changes in census geography.

GIS methods and concepts enable us to create consistent machine-readable census geography across time and thereby allow coherent chronological and comparative analysis of small-area census data.

4. The development of the Data Documentation Initiative (DDI) international aggregate metadata standard gives us the essential tool for automatic processing of census documentation. Without the DDI, it would be far more expensive to manage this vast collection of data and documentation.

5. The development of advanced Web-based data extraction tools at the Minnesota Machine Readable Data Center, the Minnesota Population Center (MPC), and elsewhere has made it possible to simplify access to complex data structures. Students, policy analysts, journalists, and academic researchers will therefore not need specialized training to use the data.

Although these innovations make the NHGIS possible, it remains a challenging project: we must acquire and clean approximately 670 gigabytes of data, produce the equivalent of approximately 50,000 pages of DDI-compliant documentation, create 630,000 geographic polygons, and write an estimated 60,000 lines of software code.

Research Applications

The ready availability of aggregate census data in a GIS framework will have important applications for a broad range of population research problems. Among key substantive areas are 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; public health and epidemiological studies; the transformation of electoral politics; geographic criminal justice studies; environmental justice; and multilevel analysis integrating aggregate census data and microlevel data. The potential list of topics that can be addressed with these data is far too long to discuss within the space constraints of this article. The following discussion provides just two examples to illustrate the exciting possibilities the NHGIS has introduced:

Residential segregation studies and urban change. Although social scientists have developed a multitude of segregation indices over the past ten years, nearly all work applies such indices to a single moment in time. Moreovèr, most studies focus on simple indices of segregation rather than assessing changing residence patterns (Taeuber and Taeuber 1965; Massey and Denton 1998; Farley and Frey 1996). Through painstaking reconstruction of census geography, Alba et al. (1995) were able to assess factors that predict changes in specific neighborhoods over a 20-year period (see also Wyly and Hammel 1998). The Alba study required the sort of data for New York City that we are creating for the entire United States. Moreover, by linking microdata to tract-level data, analysts can assess how neighborhood change is associated with individual behavior (South and Crowder 1997). The NHGIS promises to transform the study of urban racial and ethnic change by allowing both long-run analysis of change and comparison of cities within regions or across the nation.

Public health and epidemiological studies. The field of epidemiology is rapidly turning to spatial analysis of census data not only to trace the path of diseases but also to determine their impact on subpopulations (e.g., Becker et al. 1998; Leclere, Rogers, and Peters 1998; Sucoff and Upchurch 1998; Miles-Doan 1998; Latkin, Glass, and Duncan 1998; Liu, Deapen, and Bernstein 1998; Sayegh et al. 1999; Arbes et al. 1999; Nuorti et al. 2000). Pulido (2000) has argued that a historical perspective is critical for analysis of disease incidence. Spatial-temporal epidemiological investigations, however, remain difficult. Historical data on the incidence of disease by county and census tract are already available (e.g., http://www.nci.nih.gov/atlas; Lang and Polansky 1994), but epidemiologists often do not have access to appropriate historical small area data needed to calculate morbidity rates for population subgroups or to control for historical variation in population composition at the local level. The NHGIS database will expand the potential for investigating public health in chronological and geographic context.

Source Data

This project incorporates all surviving machine-readable aggregate census data for the United States and adds new data transcribed from printed and manuscript sources. Currently, even raw machine-readable counts of the population characteristics of local areas are inaccessible except to a few experts. Of the existing aggregate data, much is available only in obsolete nonstandard formats accompanied by paper documentation. Census information for the period before 1940 is especially inaccessible because most of it has never been converted to machine-readable form and exists only in paper form housed at various archives around the country. The 1960 and 1970 tract-level data are distributed only in a special early-1970s compressed format developed by DUALabs, Inc., a private census reseller that has been out of business for 25 years. Many scholars are not even aware that small-area machine-readable census data survive for these years. A valuable set of tract data from the 1940 and 1950 censuses that was laboriously keypunched at the University of Chicago a quarter-century ago was recovered only in late 1999.

Table 1 lists the principal data sets. They describe the characteristics of states and counties, census tracts, cities, MCDs, census blocks, and ZIP codes. The most important source for state and county data before 1950 is "Historical,

TABLE 1. Summary of Principal Data Sets to Be Included in the NHGIS

Date	Data-set description	Size (megabytes)	Variable count	Online availability
1790–1970	County and state (ICPSR 0003 and additional files)	253	3,508	Partial
1790–1990	County and state election returns (ICPSR 0001, 0002, 0013)	182	2,867	None
1790–1960	Municipal data (Haines files; size estimated)	160	950	None
1944–2000	County and city data books	66	966	None
1974–2000	County business patterns	3,344	65	None
1947–2002	Economic censuses	3,400	1,000	None
1949–2002	Agricultural censuses (Gutmann)	2,000	4,000	None
1940–1970	Census tract data (Bogue)	211	2,478	None
1910–1950	Supplemental tract data (Beveridge)	150	2,658	None
1960	Census-tract–level data	246	1,073	None
1970	Census small-area data			
	Count 1	275	447	None
	Count 2	710	7,000	None
	Count 4	3,224	4,300	None
1980	Census summary files			
	Summary tape file 1 A-H	10,193	321	None
	Summary tape file 2 A-C	12,697	2,292	None
	Summary tape file 3 A-D	7,328	1,126	None
	Summary tape file 4 A-C	22,713	1,500	None
	Summary tape file 5	5,000	100,000	None
	Equal opportunity employment file	415	1,100	None
	Journey-to-work file	172	88	None
1990	Census summary files			
	Summary tape file 1 A-D	60,476	1,050	Partial
	Summary tape file 2 A-C	33,966	2,187	None
	Summary tape file 3 A-D	33,625	3,225	All
	Summary tape file 4 A-C	157,826	22,040	None
	Special summary tape files: 1–22	15,689	44,600	None
	File 420 place of work 20 destinations	541	406	None
	File S-5 workers by residence by workplace	22	19	None
	Equal opportunity employment file	909	13,135	None
2000	Census summary files (size estimated)	300,000	80,000	Partial
Total		671,293	304,401	

Note. ICPSR = Inter-university Consortium for Political and Social Research.

Demographic, Economic, and Social Data: The United States, 1790–1970," a data set created 30 years ago by the ICPSR with funding from the National Science Foundation (ICPSR study 0003). This data set includes the bulk of published nineteenth- and early-twentieth-century state- and county-level statistics from the censuses of population, agriculture, manufacturing, and religion. Unfortunately, the file is incomplete and is plagued by numerous data-entry errors. As part of the NHGIS project, Michael Haines, at Colgate University, is correcting the errors and augmenting the data set with additional information from published and machine-readable sources.[1] To allow analysis of political change, we are also including the ICPSR county-level election return studies, which cover the period 1790 through 1990. For the period since 1950, we are supplementing the ICPSR files with machine-readable data from county data books, economic and agricultural censuses, and county business-patterns data files.

The most important statistical unit below the county level is the census tract. The tract system was first applied to 8 cities for the 1910 census. By 1930, 19 cities were tracted, and from 1940 onward most metropolitan areas and other densely populated counties were enumerated and tabulated by tract. For the early period—from 1910 to 1930—census-tract data exist only in manuscript form except for New York City, which was digitized under the direction of Andrew Beveridge, at Queens College-CUNY. Elizabeth Mullen Bogue of the University of Chicago digitized most of the published 1940 tract data and about 60 percent of the 1950 data. As part of the NHGIS project, Beveridge is now working on correcting the Bogue data and converting all remaining tract data from 1910 to 1960 into machine-readable form. Tract data produced by the Census Bureau from 1960 and 1970 survive in machine-readable form but are presently stored in an obsolete format. More extensive tract data for 1980 and 1990 exist in ASCII files on magnetic tape, and some files for 1990 are available on CD-ROM. The 2000 tract files are being distributed on CD-ROM and via the Internet.

In addition to the state, county, and tract data, we will incorporate data on cities, MCDs, census blocks, ZIP codes, and other census-designated places. Machine-readable data on cities and MCDs are available only for the period since 1950; with funding from NHGIS, Haines is extending some series back into the nineteenth century by digitizing published census returns. Block and ZIP code data survive only for the period since 1980.

Data Preservation and Access

The NHGIS database addresses one of the most perplexing problems currently facing depository librarians, that of preserving and maintaining functional access to government data. The National Archives preserves much of the machine-readable census data and documentation, but no institution maintains the software that was designed to provide access.

Many summary data files produced from 1960 through 1980 came without search software. Software tools that did exist, such as the Census Software Package (CENSPAC), were dependent on particular hardware and software that are no longer maintained. Thus, today's researchers actually have no functional access to machine-readable aggregate data for the 1960, 1970, and 1980 censuses. Virtually no small-area data for the period before 1990 are available on the Internet.

There is every reason to believe that the same basic problems of hardware and software dependence, obsolescence, and loss of functional access will arise with the 1990 and 2000 census data released on CD-ROM. Indeed, maintaining functional access to data products is in some respects a more serious problem for the 1990 and 2000 censuses than for earlier census years because the bureau has abandoned printed publication of small-area data. The National Archives is preserving raw census tables in machine-readable form, but it cannot maintain the operating-system-dependent software needed to process, extract, and analyze the information. The majority of data users lack the expertise to look up particular items in the raw data files, so long-term access is endangered.

For the time being, selected 1990 data are accessible from multiple Internet sites with basic look-up and subsetting functions, but the great bulk of the 1990 census data is still unavailable. Even within the Census Bureau, access to some of the most useful 1990 tract data is a cumbersome procedure involving specialized software, auxiliary files, templates, and highly skilled personnel.

Chronological analysis multiplies the complexity of data access. Because software and file formats differ in every census year, it is nearly impossible to assess change over time. A generic search and extraction engine, however, could present the contents of each data file intelligently without the need for customization. To create such an engine, we need a method for encoding documentation in a form that allows automatic processing of the data. In other words, we need comprehensive machine-understandable documentation.

A solution is within reach. As described by Block and Thomas (in Part Two of this issue), the DDI provides a standardized structure for social science data documentation. First published in March 2000, the DDI is a document-type definition for microdata in the eXtensible Markup Language (XML). In June 2001, a beta version of the extensions for aggregate data was approved by the DDI Committee. The machine-understandable structure of the DDI allows for automated processing by data access software.

Because the DDI is being adopted by most of the world's leading data archives and statistical agencies, we can be confident that it will not become obsolete in the foreseeable future. Eventually, of course, new metadata standards will emerge. The enormous base of DDI-compliant documentation that will soon exist ensures that software will be available to migrate the entire documentation system smoothly with minimal human intervention.

Geographic Information Systems

The DDI has a critical limitation: it cannot fully describe the basic geographic units used by the Census Bureau. The census data describe small geographic areas, and the boundaries of those areas were modified in each census year. In the past, researchers used small-area census data in conjunction with paper census-tract maps so they could visually identify the places concerned. This approach is unwieldy for analyzing change across time or across many geographic areas.[2] To open the body of census data to chronological and spatial analysis, we must provide machine-readable descriptions of the places covered by the statistics.

In recent years, technological change has revolutionized the field of population geography. We now have powerful computer-based methods for the acquisition, storage, analysis, and display of spatial data. This GIS technology has significantly broadened the scope of questions that can be answered with geospatial data and has popularized the use of mapping techniques for displaying this information.

At the core of GIS technology are electronic boundary files that describe the spatial dimensions of each geographic entity. High-quality machine-readable census boundary files for small areas exist only for the 1990 and 2000 censuses. Low-definition boundary files for 1980 census tracts also exist. We are creating high-quality census-tract maps back to 1910 and state and county maps back to 1790. The availability of high-quality boundaries for the key statistical areas will allow us to reconcile changes in census geography. This in turn will make it possible to provide researchers with estimates of changing population characteristics and distribution under constant definitions of census geography.

Web-Based Dissemination

The key to making census summary files broadly accessible is the development of a powerful and flexible Web-based data access system. The MPC has been working on methods of electronic dissemination for social science data and documentation since 1993 and has developed the most powerful and widely used tool for access to census microdata (http://www.ipums.org). We plan to develop the NHGIS data access system according to the same principles: ease of use will be paramount, and full documentation will be seamlessly integrated with the data extraction functions.

The system will consist of a set of tools for navigating the mass of documentation, selecting variables, and creating formatted tables and thematic maps. In addition, users will be able to extract downloadable data sets and boundary files in a variety of formats. Users will create customized subsets of both data and documentation tailored to their particular research questions. The system will be intelligent enough to allow use by novices but flexible enough to meet the needs of advanced analysts. For example, users will have three options for geographic case selection: a clickable map interface, a scrollable structured list, and a search utility that will return a list of geographic units matching any part of a given place name. We will provide a variety of tools to allow users to locate appropriate variables and will provide basic statistical and arithmetic functions to allow users to calculate percentages, means, and ratios using any desired denominator.

Methods and Procedures

Creating a spatiotemporal database of this magnitude with a user-friendly data access system requires a knowledgeable team of experts. NHGIS brings together a group of Minnesota researchers who are uniquely equipped to handle the challenge. John S. Adams, a leading population and urban geographer, has directed many large cartographic projects and is responsible for overall project coordination. Also from the Department of Geography, Robert B. McMaster and Mark Lindberg provide GIS expertise. Lindberg, director of the Cartography Laboratory, is responsible for the overall creation of the cartographic database. McMaster uses his expertise in digital and analytical cartography and urban-based geographic information systems to lead work on interpolation and generalization. Steven Ruggles works closely with the programmers building data access tools and with the subcontractors who are preparing the new historical census data. Wendy Thomas brings years of experience with aggregate data and extensive work on the DDI. She oversees the development of machine-understandable documentation. William Block, the director of the Information Technology Core, directs work on the data access system.

We have subdivided the project into three closely interrelated work components. The *data and documentation* component collects the data and documentation and converts them to a form suitable for redistribution. The *mapping* component creates historical electronic boundary files describing census geography. The *data access* component develops extraction, browsing, and mapping software for data and metadata. These work components are interdependent. Without the development of electronic maps, it would be impossible to make consistent statistical comparisons across census years; without the work on metadata and data, there would be no statistics to compare; and without the development of new data access tools, the data, metadata, and electronic maps would remain largely inaccessible. The following sections detail the methods and procedures of each component in turn.

1. Data and Documentation

Data and documentation lie at the heart of the project. Under the direction of Wendy Thomas, the NHGIS is acquiring and cleaning the census data and documentation, harmonizing its format, and creating DDI-compliant codebooks for each file. The central task is to convert into DDI form all metadata pertaining to the aggregate U.S. censuses. This step allows us to rationalize the structure of the data files automatically and to implement a single extraction and browsing tool for the entire collection of data. In addition, the data and metadata component is filling in gaps in the existing machine-readable data series with the assistance of Myron Gutmann, at ICPSR, and the aforementioned Michael Haines and Andrew Beveridge.

Creation of XML DDI codebooks. Source codebooks for aggregate census files exist in three different formats: print, electronic images of print, and ASCII text files of the original print file. In some cases, only the data dictionary portion of the original documentation is available in ASCII format. We are creating scanned images of all source documents for archival and reference purposes and are converting these files to PDF documents for Web distribution.

The PDF codebooks will also be our primary source of information for the creation of DDI-compliant metadata. Each piece of information from the codebooks will be labeled with a "tag" that identifies the particular type of information, such as title, author, variable description, or variable label. This markup is being carried out by a team of nine research assistants using a combination of customized XML authoring tools and standard editors. Some census files, especially those from recent years, have either an ASCII text file of the data dictionary with good structural integrity or a data definition file for SPSS or SAS processing. The information contained in these files can be converted automatically to DDI format. We have created PERL scripts to map data into the appropriate elements wherever possible to reduce the amount of hand entry required, thus cutting back on errors.

We use three verification methods to ensure quality control. One or more of these methods are used for each section of the DDI codebook. First, we carry out blind verification through double entry, particularly in textual areas. Second, we create a template that duplicates the original print layout, and then we visually compare the new documentation with the original paper documentation. Third, we use the marked-up documentation to read the data and calculate totals that can be used to check the accuracy of the metadata. If sums across variable categories and geographic entities are correct, this verifies that all variables within a matrix have been entered and that the start position, end position, and width of each variable are consistent with one another and result in the specified record layout.

Preparation of ancillary documentation. Metadata are not confined to codebooks. We are bringing together supplemental documentation to form a comprehensive collection of reference materials. These materials provide full detail on geographic, occupational, and industrial coding schemes; the construction of poverty indices and other derived variables over time; methodological papers, census questionnaires, 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. Much of this background material already exists as part of the online documentation of the IPUMS project at the MPC, but we need to mark up these files to allow full integration with the NHGIS data access system. We are also preparing new documentation to address data harmonization, the data creation process for new data files, descriptions of incompatibility in variables or definitions among and within data series, and the application of data to specific research issues.

2. *Mapping*

The need for computerized census maps is fundamental. For effective chronological analysis of the aggregate data, we need to know how geographic units were altered from one census to the next. The mapping component of the project is substantial. By the time it is complete, we estimate that we will have scanned approximately 7,000 source maps and created 630,000 cleaned and verified polygons.

The unit of analysis in aggregate census data is a geographic entity: the census block, tract, MCD, county division, county, city, metropolitan area, or state. All these units—even states—can change from one census year to the next. Our primary cartographic emphases are census tracts and counties, the basic building blocks of the Census Bureau's statistical system. The finest level of geography in the electronic maps is the census tract. Census tracts usually have between 2,500 and 8,000 persons and, when first delineated, were designed to be homogeneous with respect to population characteristics, economic status, and living conditions.[3] In addition to the tract maps, we are developing new high-precision county maps for the entire country since 1790. We are also constructing geographic entities that are aggregates of tracts or counties, such as metropolitan areas. We do not have sufficient funds in the current project to create maps of MCDs, MCD-equivalents, and census county divisions, but we hope to address these geographic units in a future project.

Constructing census-tract databases. High-quality tract boundary files produced by the Census Bureau are available for the 1990 and 2000 censuses. Our task is to construct similar files for earlier years. We begin by developing a clean set of tract boundaries for the 1990 and 2000 censuses. The Census Bureau has created geocoded street- and enumeration-unit files—the TIGER files—that allow spatial analysis and mapping. We derive tract boundary files from TIGER files by extracting tract boundaries and generating polygonal topology.

Working backward from census to census, we undo the boundary changes of each census year to represent the tract border for the previous census year. This approach allows us to use existing digital data as a basis for generating all base maps. In addition, it minimizes work needed for generating tracts for an earlier census year. The two most common geometric changes over time are the addition of new tracts to the periphery of an already tracted area and the splitting of existing tracts; the most frequent editing operations, therefore, are removing tract designations and merging tracts by removing their common border. By reusing borders across different years we maximize the geographic correspondence between the different data sets.

The digital files are edited onscreen to minimize the high cost of using a graphics tablet digitizer. The primary data set for each pre-1990 census is a copy of the cleaned tracts for the succeeding census. We scan paper tract maps for the appropriate year and rectify them to the cleaned tract files. These scanned images provide the base for editing the cleaned tract files.

We assemble the TIGER data into single layers, by county, and use them in the production of all pre-1990 tract data sets. The TIGER layer provides two essential functions. First, it is used as a general reference, especially for verifying tract borders by street name. Second, it is the first choice for obtaining additional line work. For example, if a tract border changed over time from a railroad line to a street, and if the railroad still exists in the TIGER file, we copy the appropriate line from the TIGER layer directly into the tract base map.

When we need to add new borders that are not found in the TIGER data, we draw from a variety of additional ancillary data including maps in the John R. Borchert Map Library at the University of Minnesota. In some instances, we are using materials held elsewhere, including other university libraries, the Library of Congress, and the National Archives.

The tract databases contain corresponding attribute data describing each tract's history. For example, we can query a particular 1990 census tract and find out that it existed with identical borders for 1980 and 1970 but that the 1970 tract had a different tract identifier (ID). Likewise, we can query a tract that was newly created in 1950, determine that it existed with identical borders in 1960, that it was split into two tracts for 1970, and obtain its sibling tract IDs for 1970 and subsequent censuses. These attribute data will make up an essential building block of the data access system.

When the county-tract base maps are complete for each census year, we reexamine the data sets for quality assurance. We employ multiple approaches, including checking tract borders with paper maps, using the Census Bureau tract comparability tables to verify changes between census years, and matching tract IDs with identifiers used in census tables. After the county files have passed quality assurance procedures, we export the files and archive the original source materials (coverages and scans). Once the tracts for each census year are constructed, the data will be generalized. We will modify these data sets by simplifying some lines to eliminate unwanted artifacts such as remnants of piers, removing selected water bodies such as large lakes and coastal bays, and eliminating sliver polygons.

The tract base map development uses ArcInfo for all editing work: We store base maps as ArcInfo coverages and maintain scanned maps as either TIFF files or ArcInfo grids, depending on the amount of rectification required to match them to existing tract base files. Work on the tract database is proceeding on schedule. Our team of eight research assistants under the direction of Mark Lindberg has completed work on approximately 200 heavily urban counties. We anticipate that work on the tract database will be complete in 2005.

Constructing county and state databases. We began production of the county database in June 2002. The procedures for counties are similar to the procedures for census tracts. For each state, we start with existing data sets and work backward, census by census, to build boundary compatibility across census years. We establish an initial database through minor editing of the 1990 and 2000 Census Bureau boundary files and then work backward to earlier census years. Unlike tract editing, however, we rely on multiple sources for each state. We are collaborating with the Atlas of Historical County Boundaries project, directed by John Long (1984) of the Newberry Library. Additional references for changes in county boundaries are *Population of States and Counties of the U.S.: 1790–1990* (Forstall 1996), *Map Guide to the U.S. Federal Censuses: 1790–1920* (Thorndale and Dollarhide 1987), and a set of Historical U.S. County Boundary files created at Louisiana State University (http://www.ga.lsu.edu/ga/husco.html). We cannot work directly from the LSU files because they do not precisely match the TIGER files, but they are a useful reference when other sources are not available. In addition, we supplement our primary references with data sources available only for specific states. For example, we used *The Establishment of County Boundaries in Minnesota* (Lewis 1946) as an additional reference for the State of Minnesota; figure 1 illustrates a selection of the resulting boundaries, shown here with county population densities. Once the county data sets are developed, we will be able to extract corresponding state sets simply by merging counties.

Interpolation. Changes in boundaries can lead to spatial incompatibility across census years. For example, census tract 101 in 1980 might be split into 101.01 and 101.02 in 1990. The result is that a direct comparison of population statistics, such as median housing value, will not be meaningful unless the data are reaggregated. To address this problem, the NHGIS will allow users to designate a reference year and estimate population characteristics for all years on the basis of the geography of the reference year. These estimates will be based on spatial interpolation incorporated in look-up tables. We are now exploring the advantages and liabilities of several alternative interpolation methodologies.

3. Data Access Tools

We will distribute data and documentation free of charge on the Internet through an integrated data access system. Users will extract customized subsets of both data and documentation tailored to their particular research questions. This will not, however, be simply a data extraction system. Instead of presenting the data in isolation, this system will provide the user with both the data and a rich informational context. Users will have the option of defining their geographic areas of interest through a map interface and will be able to display their results in the form of a table, map, or downloadable data set.

Our goal is to make the system easier to navigate than current tools for accessing aggregate data on the Internet, and to provide more flexible topical search features. Existing systems usually provide only one method for selecting variables. Having provided little or no contextual documentation, the systems typically require researchers to know technical terminology. Moreover, because existing extraction tools cannot handle the largest and most complex data sets, the assistance of an expert data archivist is still required for many applications. The NHGIS data access tools will take advantage of recent developments in information technology to transcend these limitations.

Software development for the NHGIS data access system has been under way for over a year, and the first iteration of the system will go online in 2003. This preliminary system automates access to aggregate census data from Summary Tape File 4 (STF4) of the 1990 census. STF4 is one of the most daunting but useful data sets in the census repertoire. These files provide highly detailed social and economic information for specific racial and ethnic groups at fine geographic levels. For example, they allow researchers and local planners to examine data regarding Hispanics at the tract level within metropolitan areas or immigrant groups, such as the Hmong population, at the city level. Because of the large number of variables and the corresponding complexity of the data set, only a few published studies have made use of STF4 (Logan and Alba 1993). Because the system will be built on a Web-based expert-systems interview, users will not need to consult a data archivist to make informed use of the data, and we anticipate that this improvement will increase usage of the data.

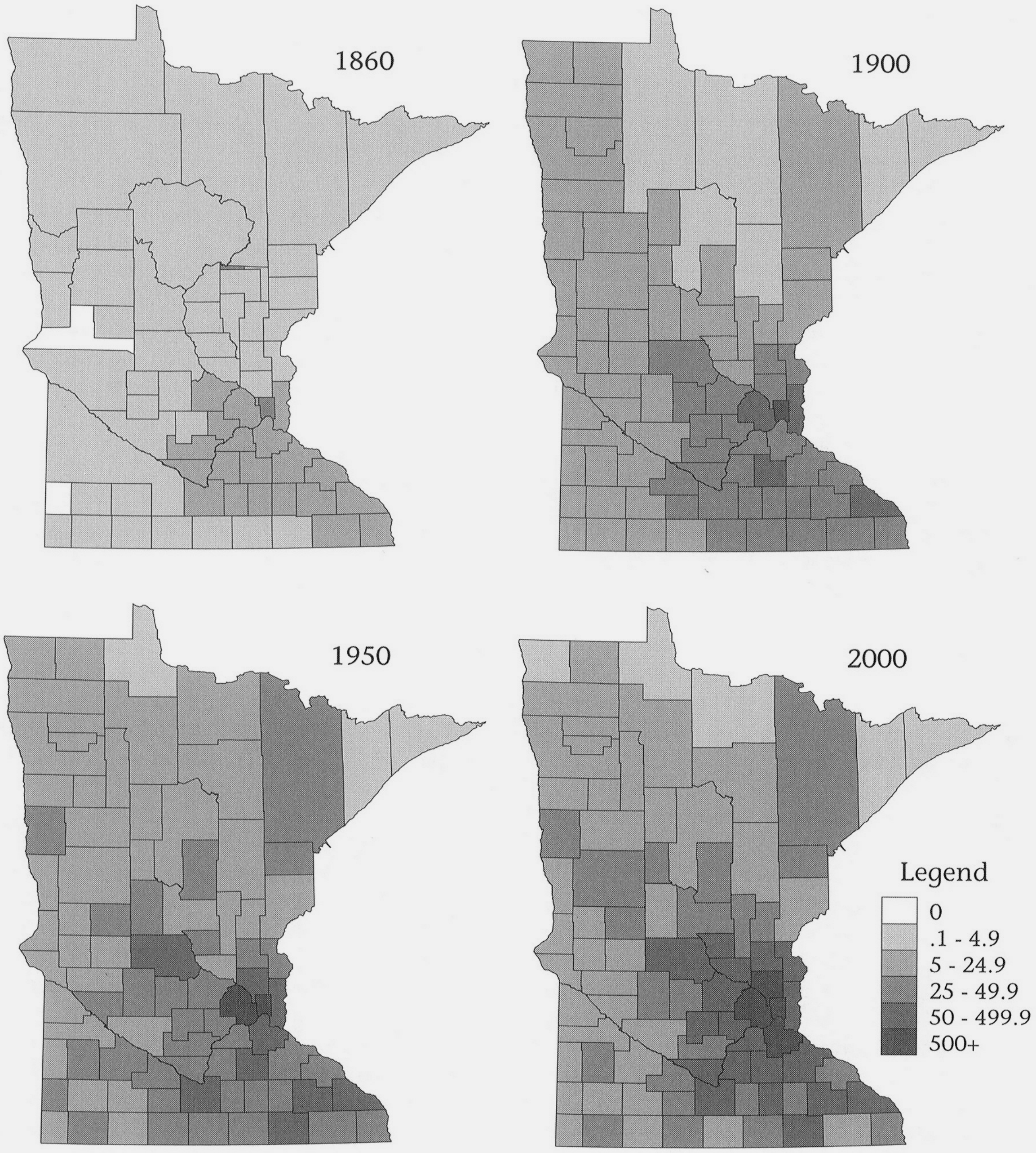

FIGURE 1. Population density of Minnesota counties, 1860–2000 (persons per square mile).

The DDI drives the data access system and underlies both access to documentation and the procedures for locating and extracting statistical information. The NHGIS will include hundreds of data files and hundreds of thousands of variables and geographic entities. A system of this complexity would be impractical if each data set and variable required customized software development. The data access system will use the same software code for all data sets and variables, relying on the machine-understandable documentation to control both the user interface and data extraction.

Because it is driven by the DDI, our approach eliminates the need to alter the user interface or the underlying extraction engine as the database grows. Adding new data and documentation will be a simple matter of creating new DDI-compliant metadata. It will also be possible to apply the software to aggregate data from other countries, with only modest adjustments.

The Web server and the extract engine are two distinct processes. The extract engine manages queries from the user interface, searches DDI-compliant XML-tagged codebooks, and then provides feedback to the user interface or the processing system in the format needed. This approach allows the use of a DDI codebook as a nodal hub within a network of documents, including electronic boundary files, statistical data, guides, and ancillary documents.

Variable search and selection. Variable selection poses special challenges for the aggregate data files because the number of variables is large; we estimate the total number to exceed 300,000. The standard method for variable selection in Web-based extraction systems forces users to select variables from a set of static lists. This approach is impractical for the NHGIS. Access to aggregate data is further complicated because a different array of subject variables is available at each level of geography in each census year. For example, in most census years, income and education are available at the tract level but not at the block level. Some variables are available in all census years, and others appear only once. Therefore, the development of tools to allow users to locate information they need is of paramount importance.

Although our design has not been finalized, we are working on several alternative methods for locating variables. Novice users will be offered an expert-systems interview, which will pose a series of questions to identify their areas of substantive interest, the level of geographic detail they need, and the chronological scope of their investigation. More advanced users will be able to search all documentation by variable name, keyword, and variable description, and to drill down from broad subject classifications to specific groups of variables. Users will be able to specify the level of geographic detail and census years at the outset, thereby restricting the search universe to variables that meet these requirements. Alternatively, they will be able to explore the tradeoffs between geography and variable availability dynamically. Once users have narrowed their search to a manageable group of variables, the system will highlight the subset of variables available for any given combination of geographic level and census year. When users modify the geographical and chronological specifications, the highlighted group of variables will change accordingly. Users will then be able to add variables to a "data basket" for later analysis. At any time during the selection process, users will be able to click on a variable to get a full description and analysis of comparability problems across census years, enumerator instructions, and other supporting documentation.

We also plan to support several methods of geographic case selection. Most users will specify geography through a clickable map interface. They will first define one or more states, metropolitan areas, or regions by clicking on a national map. They will then specify finer geographic level of interest—counties, cities, or tracts. This will yield a clickable map with the appropriate boundaries. Users will add geographic selections to a data basket and at any time will be able to shift to a different geographic level or region. Unlike existing geographic selection methods, users will be able to mix and match different geographies. For example, they will be able to select a combination of municipalities, MCDs, and census tracts within the same extract. Advanced users who know the names of the geographic units of interest will have the option of selecting geographic units from a structured list or of using a search engine to locate geographic entities by name or Federal Information Processing Standards (FIPS) code. The system will allow users to go back at any time and change their geographic and subject variable selections and will also permit them to define their selections in whatever order they wish.

Retrieving and managing data. The system will provide results in the form of tables or thematic maps displayed on the screen. Onscreen tables will be self-documented; hyperlinks on each variable and value label will lead to information on sources, universe, comparability, enumerator instructions, and other ancillary documentation. We will develop table-formatting software specifically for this project and plan to use the ArcIMS tools developed by the Environmental Systems Research Institute (ESRI) to generate maps. Users will also be able to extract data sets for downloading and further local processing. Downloadable statistical data sets will be created as delimited or column-format files and will be accompanied by customized documentation and data definition files for SAS, SPSS, or Stata. Electronic boundary data sets will be distributed in several formats (e.g., ArcInfo exchange format, ESRI shape files, and MapInfo exchange format).

The procedures for defining a given table, map, or downloadable file can be quite complex, involving variable and geographic selections, statistical function, and table or map formatting. We will allow users to save and retrieve a profile for each analysis they create. This feature will enable users to re-create or modify a table or map made in a previous session.

Conclusion

The census constitutes a fundamental underpinning of social science and policy research. It also represents an exceptional untapped resource for secondary and higher education in the social sciences, statistics, and history. The NHGIS project will allow all users—from high school students to research scientists—to adopt a comparative and historical perspective.

Small-area census data are the primary source for studying such critical issues in social science research as suburbanization, the decline and rebirth of central cities, residential segregation, immigrant settlement patterns, rural depopulation, agricultural consolidation, and population shifts from the Rust Belt to the Sun Belt. These issues cry out for chronological analysis, but because historical small-area data are inaccessible, most studies are static. Researchers who do address change over time in spatial processes must either confine their analyses to local areas because broader studies are simply too expensive, or they must adopt large units of analysis (such as states and metropolitan areas) that preclude the nuanced detail needed for full understanding. The NHGIS database will open a new range of powerful approaches to familiar problems, broadening the scope of local and regional analyses to explore variations across time and space simultaneously.

The database will have even broader application when combined with other sources. The census provides basic denominators for an array of studies across the social sciences, including such diverse fields as political science, criminal justice, and epidemiology. The availability of small-area data and geographic boundary files will allow such analyses to incorporate a chronological as well as a spatial dimension.

Social scientists have become increasingly aware that individuals' life chances, choices, and attitudes are shaped not only by their own characteristics but also by the characteristics of their neighbors and communities. The database will encourage and simplify the use of techniques such as multilevel analysis that draw upon such insights. These aggregate-level census data will therefore dovetail with and complement widely used microlevel data sets, such as the IPUMS.

We hope, however, that the NHGIS database will be used by more than just research scientists and will actually help to democratize access to the census. Our goal is to see the database used for social science training and education at all levels, by the media, for policy research at the state and local levels, and by the private sector.

NOTES

[1] We have obtained approximately 65 state and county files (prepared by eight different scholars) that will supplement the ICPSR data, but several important gaps remain that will have to be entered directly from published sources. The raw data from ICPSR study 0003 are available for viewing online (http://fisher.lib.virginia.edu/census), but they cannot be downloaded and no mapping or statistical facilities are available.

[2] It is now possible—although difficult—to map changes across recent census years in areas where few tract changes exist (Denton and Massey 1991; Alba et al. 1995). Studies with longer chronological scope are so time consuming and tedious that they are seldom undertaken.

[3] Block-level maps will be included for the 1990 and 2000 censuses, but we will not construct them for earlier years because of the considerable expense involved and the fact that no machine-readable pre-1980 census detail is available at the block level.

REFERENCES

Alba, R. D. et al. 1995. Neighborhood change under conditions of mass immigration: The New York City region, 1970–1990. *International Migration Review* 29: 625–56.

Arbes, S. J. Jr. et al. 1999. Factors contributing to the poorer survival of black Americans diagnosed with oral cancer. *Cancer Causes & Control* 10: 513–23.

Becker, K. M. et al. 1998. Geographic epidemiology of gonorrhea in Baltimore, Maryland, using a geographic information system. *American Journal of Epidemiology* 147: 709–16.

Denton, N. A., and D. S. Massey. 1991. Patterns of neighborhood transition in a multiethnic world: U.S. metropolitan areas, 1970–1980. *Demography* 28: 41–63.

Duncan, O. D. 1957. *The Negro population of Chicago: A study of residential succession.* Chicago: University of Chicago Press.

Farley, R., and W. H. Frey. 1996. Latino, Asian, and black segregation in U.S. metropolitan areas: Are multiethnic metros different? *Demography* 33: 35–49.

Forstall, R. L. 1996. *Population of states and counties of the United States: 1790–1990.* Washington, D.C.: U.S. Census Bureau.

Lang, D. M., and M. Polansky. 1994. Patterns of asthma mortality in Philadelphia from 1969 to 1991. *New England Journal of Medicine* 331: 1542–46.

Latkin C., G. E. Glass, and T. Duncan. 1998. Using geographic information systems to assess spatial patterns of drug use, selection bias and attrition among a sample of injection drug users. *Drug & Alcohol Dependence* 50: 167–75.

Leclere, F. B., R. G. Rogers, and K. Peters. 1998. Neighborhood social context and racial differences in women's heart disease mortality. *Journal of Health and Social Behavior* 39: 91–108.

Lewis, M. E. 1946. The establishment of county boundaries in Minnesota. Master's thesis, University of Minnesota.

Lieberson, S. 1961. A societal theory of race and ethnic relations. *American Sociological Review* 26:92–110.

Liu, L., D. Deapen, and L. Bernstein. 1998. Socioeconomic status and cancers of the female breast and reproductive organs: A comparison across racial/ethnic populations in Los Angeles County, California (United States). *Cancer Causes & Control* 9: 369–80.

Logan, J. R., and R. D. Alba. 1993. Locational returns to human capital: Minority access to suburban community resources. *Demography* 30: 243–69.

Logan, J. R. et al. 1996. Making a place in the metropolis: Locational attainment in cities and suburbs. *Demography* 33: 443–53.

Long, J. H., ed. 1984. *Historical atlas and chronology of county boundaries, 1788–1980.* Vol. 5: Minnesota, North Dakota, South Dakota. Boston: G. K. Hall.

Massey, D. S., and N. A. Denton. 1988. Suburbanization and segregation in U.S. metropolitan areas. *American Journal of Sociology* 94: 592–626.

———. 1998. The elusive quest for the perfect index of concentration: Reply to Egan and Weber. *Social Forces* 76: 1123–34.

Massey, D. S., and M. L. Eggers. 1990. The ecology of inequality: Minorities and the concentration of poverty, 1970–1980. *American Journal of Sociology* 95: 1153–88.

Miles-Doan, R. 1998. Violence between spouses and intimates: Does neighborhood context matter? *Social Forces* 77: 623–25.

Nuorti, J. P. et al. 2000. Epidemiologic relation between HIV and invasive pneumococcal disease in San Francisco County, California. *Annals of Internal Medicine* 132: 182–90.

Pulido, L. 2000. Rethinking environmental racism: White privilege and urban development in Southern California. *Annals of the Association of American Geographers* 90: 12–40.

Sayegh, A. J. et al. 1999. Does race or socioeconomic status predict adverse outcome after out-of-hospital cardiac arrest: A multi-center study. *Resuscitation* 40: 141–46.

South, S. J., and K. D. Crowder. 1997. Residential mobility between cities and suburbs: Race, suburbanization, and back-to-the-city moves. *Demography* 34: 525–38.

Sucoff, C. A., and D. M. Upchurch. 1998. Neighborhood context and the risk of childbearing among metropolitan-area black adolescents. *American Sociological Review* 63: 571–86.

Taeuber, K. E.,, and A. F. Taeuber. 1965. *Negroes in cities: Residential segregation and neighborhood change*. Chicago: Aldine Publishing.

Thorndale, W., and W. Dollarhide. 1987. *Map guide to the U.S. federal censuses: 1790–1920*. Baltimore: Genealogical Publishing.

Wyly, E. K., and D. J. Hammel. 1998. Modeling the context and contingency of gentrification. *Journal of Urban Affairs* 20: 303–27.