Published in final edited form as:

Hist Methods. 2011 January 1; 44(2): 61–68. doi:10.1080/01615440.2011.564572.

Big Data:

Large-Scale Historical Infrastructure from the Minnesota Population Center

MATTHEW SOBEK, LARA CLEVELAND, SARAH FLOOD, PATRICIA KELLY HALL, MIRIAM L. KING, STEVEN RUGGLES, and MATTHEW SCHROEDER

Minnesota Population Center, University of Minnesota

Abstract

The Minnesota Population Center (MPC) provides aggregate data and microdata that have been integrated and harmonized to maximize crosstemporal and cross-spatial comparability. All MPC data products are distributed free of charge through an interactive Web interface that enables users to limit the data and metadata being analyzed to samples and variables of interest to their research. In this article, the authors describe the integrated databases available from the MPC, report on recent additions and enhancements to these data sets, and summarize new online tools and resources that help users to analyze the data over time. They conclude with a description of the MPC's newest and largest infrastructure project to date: a global population and environment data network.

Keywords

ATUS; census data; CPS; IHIS; IPUMS; ISTADS; Minnesota Population Center; NAPP; NHGIS; survey microdata

It takes big data to truly understand the big picture, whether across large stretches of time or large expanses of territory. Since its inception in 2000, the Minnesota Population Center (MPC) has become a world leader in the production of large databases that form the infrastructure for research in the social sciences. MPC infrastructure projects have (1) created dozens of new historical census data sets, (2) improved the comparability and documentation of census and survey data across time and space, (3) preserved vast quantities of older microdata at risk of destruction or obsolescence, (4) developed a comprehensive historical geographic information system for the United States, (5) pioneered the development and application of new metadata standards for population data, and (6) produced innovative tools for the electronic dissemination of time series data on the world's population. Thanks to our partners in national statistical offices around the world, we now hold the largest collection of population census and survey microdata. We also have the world's most complete collection of documentation on international decennial censuses, augmented substantially with publications shared from the archives of the United Nations and the U.S. Census Bureau.

This collection of big data—available free to qualified researchers via the Internet—has been growing and is about to get much bigger. The purpose of this article is to describe the enhancements in data availability and online research tools that have been added by the

MPC in the last several years and to outline plans for the exciting new project that will launch later this year.

We provide a brief summary of MPC data projects, with particular emphasis on recent or forthcoming data enhancements. We then address new tools and resources for data users, many of which have been implemented in all or most of our integrated databases. We end with a description of Terra Populus, a new project that will harmonize 150 years of population microdata with all available environmental and climate data for the entire globe, which recently was awarded funding from the National Science Foundation.

MPC Data Projects and Their Recent Enhancements

The Integrated Public Use Microdata Series (IPUMS) data (Ruggles et al. 2010) for decennial censuses was the first MPC data project disseminated over the Internet. Today there is a family of IPUMS data—IPUMS-USA, IPUMS-International (for international census data), and IPUMS-Current Population Survey (IPUMS-CPS) (integrating the monthly current population surveys for the United States)—all based on the same integration and harmonization principles. There are currently four additional databases available for use by social science researchers: the North Atlantic Population Project (NAPP), the National Historic Geographic Information System (NHGIS), Integrated Health Interview Series (IHIS), and the American Time Use Survey Extract System (ATUS-X; which harmonizes the American Time-Use Survey). One new project—the Integrated Spatio-Temporal Aggregate Data Series (ISTADS)—will be available to users later this year as part of NHGIS. Data and metadata for all projects can be accessed at <http://www.ipums.org>.

IPUMS-USA

The first integrated database created at the MPC, IPUMS-USA is a harmonized data set of individual-level samples of the U.S. censuses from 1850 to 2000. The January–March 2011 issue of *Historical Methods* included descriptions of the content and methods used to produce representative samples that link records from the 1880 census complete-count database to 1 percent samples of the 1850–1930 U.S. censuses of the population (Goeken, Huynh, Lynch, and Vick 2011; Vick and Huynh 2011). The scope of the IPUMS-USA census database recently expanded with the creation of a 2.5 percent sample for the 1930 census, containing records for 3 million persons, and will soon include 1930 territorial samples covering Alaska, Hawaii, and Puerto Rico. Multilevel analysis using IPUMS-USA census data became easier with the addition of the NHGISJOIN variable, which can be used to link IPUMS data from 1850 to 1930 with the county-level boundary files and aggregate data from the National Historical Geographic Information System (NHGIS).

Beginning in 2000, the U.S. Census Bureau began to move away from the census long form with the establishment of the American Community Survey (ACS), a nationally representative annual survey that includes questions similar to those on 2000 census long form. The survey is sent to approximately 3 million housing units and group quarters in every U.S. county, as well as to American Indian and Alaska Native areas and Hawaiian homelands. ACS data (and the parallel data from the Puerto Rican Community Survey, or PRCS) are an important part of IPUMS-USA and receive the same codes and variable names as the equivalent census data. Where there are differences between the census and ACS or PRCS data (e.g., the reference period for income variables), these differences are noted in variable descriptions.

In 2009, researchers associated with the MPC identified significant errors in age and sex distributions in many public use data sets released by the Census Bureau (Alexander, Davern, and Stevenson 2010). The ACS data for 2003–8, comprising 14.4 million person

records, were recently corrected to amend these and other errors.¹ Recent substantive additions to the integrated ACS data in IPUMS-USA include health insurance variables beginning in 2008 (with logical edits in the IPUMS version to improve accuracy and comparability across time); historically compatible variables on race designed to address multiple-race responses in the ACS and the 2000 census; a set of harmonized subfamily indicators that stretch back to 1880 and are more consistent than the Census Bureau's; new variables on disability and grandparents' responsibility for grandchildren added by the Census Bureau for 2000 forward; and additional variables on place of work, migration, and geography.

IPUMS-CPS

When an earlier article on IPUMS-CPS was published in *Historical Methods* (King and Tertilt 2003), the database was still under construction. Since that time, all variables in the Current Population's Annual Social and Economic Supplement (ASEC, formerly known as the "March" supplement) from 1962 to 2010 have been added to the integrated database.

Several recent additions to IPUMS-CPS should also be noted. A few variables designed to measure the impact of recent events, such as the value of federal stimulus payments during the 2008 tax year and evacuation resulting from Hurricane Katrina, were added to the survey on a single-year basis. Other recent variable additions span multiple years and will be continued into the future. These include 12 alternative poverty measures (for 2002 forward) produced by the Census Bureau, to better account for personal health care costs and geographic differences in the cost of living; eight additional disability variables (for 2009 forward); and eight summary variables on health insurance coverage produced with the help of the University of Minnesota's State Health Access Data Assistance Center (SHADAC). In addition, FIPS county codes are now identified for almost half of households for 1996 forward; for the remaining households, county identifiers are suppressed to preserve respondent confidentiality.

Over the next five years, the scope of IPUMS-CPS data will increase greatly as MPC staff add data and documentation covering all basic monthly supplement data for 1976 forward (i.e., 408 additional data sets) and all other surviving data from topical supplements, some of which date back to the late 1960s (i.e., 366 additional data sets). At the same time, we will simplify longitudinal analyses of CPS data by linking survey participants across supplements and developing new Web-based longitudinal data access tools.

IPUMS-International

When IPUMS-International was last described in this journal, the first five-year phase of the project was drawing to a close (Ruggles, King, Levison, McCaa, and Sobek 2003). At that point, data available through IPUMS-International consisted of six countries (Colombia, France, Kenya, Mexico, the United States, and Vietnam), 21 data sets, and 48 million person records. The ultimate goals of the project were to make the world's microdata available to researchers and to encourage crossnational research with a temporal dimension.

In the intervening years, IPUMS-International has succeeded beyond our wildest expectations. As of this writing, the database includes material from 55 countries, 159 censuses, and 326 million person records, and the data release planned for early summer 2011 will bring the number of samples to nearly 200. The database will continue to grow as

¹The original ACS variables, without correction for age and sex errors, will also be preserved and available on the IPUMS-USA Web site.

we incorporate new data from 2010 census rounds and from dozens of additional countries in Eurasia, Africa, Latin America, and the Caribbean.

More than 90 countries have signed data dissemination agreements to date, encompassing more than 80 percent of the world's population. The map in figure 1 shows the global coverage of IPUMS-International partnerships with national statistical offices. It identifies which countries have data in distribution through IPUMS, which have signed agreements for dissemination of data that are not yet integrated into IPUMS, and which have agreements that are pending or under consideration. Several countries have no notation, indicating that no census microdata are currently available.

The vastly increased scale of the IPUMS-International project required many innovations in how we *produced* the data. One of the most significant changes—from the data user's perspective—was our decision to make the original unharmonized source variables available to researchers. In its initial incarnation, IPUMS-International consisted only of integrated variables, as is still true of the IPUMS-USA data series. As the project grew, we needed an intermediate stage to collate information and handle data irregularities, as it was too difficult to go directly from the original variables into an international classification system handling all the world's variations. As that intermediate stage, we fully documented and performed basic data standardization on every input—or “unharmonized”—variable.

Although the unharmonized variables were initially conceived as an aid to our internal data integration process, we subsequently realized their value to researchers and made these unharmonized variables available through the IPUMS-International Web site and data extract system. This innovation offered benefits to both the project staff and the end users. Because researchers now had access to the original underlying data, we did not need to integrate every variable and could suppress some detail in the integrated coding structures, to enhance comparability. Researchers could access the source material needed to create their own versions of integrated variables (e.g., for a single country), as desired, or work with the unharmonized variables as given, if their research project required data from only a single country and sample year.

NAPP

The NAPP currently consists of complete-count nineteenth-century census data for six countries (Canada, Great Britain, Norway, Sweden, the United States, and Iceland), plus samples of historical census data from five countries (Canada, Great Britain, Germany, Norway, and the United States). For Norway and the United States, linked samples supporting longitudinal analysis are available.

Over the next five years, the size of the NAPP database will triple with the addition of 40 new data sets covering the period from 1787 to 1930. The map in figure 2 shows which countries currently have data available through the NAPP project and which will be added in the next phase. Planned innovations include creating additional linked national panels for longitudinal analysis and connecting the past to the present by merging NAPP with the IPUMS-International database. The scope of NAPP and its planned incorporation into IPUMS-International were described in detail in the previous issue of this journal (Ruggles, Roberts, Sarkar, and Sobek 2011).

NHGIS

First made available to researchers in 2006, NHGIS provides GIS-compatible boundary files and aggregate census data for the United States between 1790 and 2000. It comprises a comprehensive spatial database of historical census tract and country boundaries and a massive collection of historical and contemporary small-area data on population, housing,

agriculture, manufacturing, business patterns, voting, and other georeferenced statistics. All these statistical data, digital maps, and accompanying documentation are accessible through a fast data-retrieval system in a Web-based interface. The NHGIS project was previously described in detail in *Historical Methods* (Fitch and Ruggles 2003).

We are currently upgrading the NHGIS system to meet new geographic standards and take advantage of recently available statistical infrastructure. In particular, the project is realigning NHGIS boundary files to ensure compatibility with new geographic data from the Census Bureau and is disseminating ACS data.

ISTADS

Building on investments in NHGIS, the new ISTADS project will make it easy for researchers to incorporate aggregate U.S. census data into time-series analysis. NHGIS marked a breakthrough in delivering data on small areas together with the geographic boundary files, thus enabling researchers to do spatial analysis. However, NHGIS did not provide data that were consistent over time.

The ISTADS project will enable researchers to select integrated statistical data from 23 censuses covering the period from 1790 to the present. It will also integrate geographic units to maximize crosstemporal comparability. These data will be complemented by a dynamic metadata system that provides users with information about comparability across the censuses that they select. A new Web-based interface is being designed so users can easily identify and address comparability issues, export multiyear merged data sets, and incorporate the data and shape files into a GIS or statistical package. Users will have access to the ISTADS system in 2011. The ISTADS project is described more fully elsewhere in this issue (Noble, Van Riper, Ruggles, Schroeder, and Hindman 2011).

IHIS

The IHIS project began in 2004, with the goal of integrating, documenting, and disseminating data from the U.S. National Health Interview Survey (NHIS). Since 1957, NHIS has been tasked with monitoring the nation's health, and no other health survey in the world has comparable chronological depth and breadth of subject matter. IHIS currently offers access to more than 7,000 consistently coded variables covering health and its determinants among the U.S. population, from the 1960s to the present. An article on IHIS in this issue describes the rationale, content, methods, and planned new features for this exciting new resource for historical analysis.

ATUS-X

The ATUS-X is one of the newest MPC data integration and dissemination projects and was created in collaboration with researchers at the University of Maryland Population Research Center. Funded by the U.S. Bureau of Labor Statistics and fielded by the U.S. Census Bureau, the American Time Use Survey (ATUS) is an ongoing time-diary survey that collects data on the activities of daily life (e.g., sleep, exercise, work, eating). Data collection began with some 20,000 interviews in 2003; approximately 13,000 responses have been collected each subsequent year. ATUS respondents are drawn from households that have concluded their participation in the Current Population Survey and are a nationally representative sample of persons aged 15 years and older. In addition to the time-diary data, background information about the ATUS respondents and their households is recorded as part of the interview.

ATUS time-diary data are collected by asking respondents to report in their own words all the primary activities that they engaged in over the last 24-hour period (from 4:00 a.m. to

4:00 a.m.). Each primary activity is classified into a three-tier coding scheme representing more than 400 distinct activities, and the activity's starting and ending times are recorded. For each reported activity, respondents are asked whether they were engaged in secondary child care, the location of the activity, and—for most activities—who else was present. A special module fielded in 2006 through 2008 asked respondents whether they were eating or drinking as a secondary activity. The 2010 survey asked respondents whether they were caring for elders while they were engaged in their primary activities and about the emotions that respondents experienced during some activities.

In their original form, time-diary data are inherently complex, requiring substantial and sophisticated statistical package syntax to generate data in a form researchers can analyze. For example, the original ATUS public use data are contained in multiple files per year, with separate files referring to the person, the household, the activities, and the characteristics of individuals with whom activities were performed.

The ATUS-X dissemination system greatly reduces barriers to using these complex time-use data. The extract system produces files suitable for analyzing all the data in single or multiple years. The ATUS-X Web site allows users to build custom-designed time-use variables that sum the amount of time people spend in any combination of activities under various conditions (e.g., time spent watching television at home in the company of one's spouse). Data and documentation for 2003 through 2009 are available now, and new data will be added as they are released.

We are currently developing a feature allowing users to request data files containing CPS supplement data linked to ATUS respondents' characteristics and time-use variables. The linked files and accompanying weights will facilitate valuable analyses that would not otherwise be possible using either of these data sources alone.

New Tools and Resources for Data Users

The scale of the IPUMS-International project necessitated a variety of innovations that only a large project with considerable resources could have managed. Those innovations have subsequently propagated out to other data infrastructure projects at the MPC that face similar challenges. The critical mass of data projects at the center allows a degree of crossfertilization and technical advancement that would not have been possible in a series of individual projects housed at different institutions.

Metadata-centric Data Processing

The renovation of the data-processing and delivery framework now used by all MPC microdata projects was begun by IPUMS-International. A key change was the adoption of a metadata-centric model of data processing and Web design. The software used to recode data and display information on project Web sites is quite complex, but it is driven by a set of relatively simple documents maintained by research staff, including MPC research associates, postdoctoral associates, and graduate research assistants. Most of these documents appear to research staff as ordinary spreadsheets, but the spreadsheets are then converted to the XML format read by our custom software.

This metadata-driven system has a number of advantages. Because the data-processing and Web software are driven by the same set of metadata, the Web site stays in sync with the underlying data it purports to describe. Social scientists working on the project can make changes reflecting their informed understanding of the source material, without becoming enmeshed in the details of the higher level programming handled by our software engineers. Under this system, a simple edit in a control file can change how a variable is recoded,

suppress its presence on the Web site, or associate it with a particular block of text in the original census questionnaire. The system has greatly increased productivity, allowing, for example, IPUMS-International to process hundreds of samples and IHIS to harmonize several thousand variables with a relatively small staff and in a short period of time.

Family Pointer Variables

One of the greatest value-added data features of MPC microdata projects is the construction of pointer variables that indicate the person-number within the household of each person's coresident mother, father, or spouse. These variables were part of the original IPUMS-USA database, were readily constructed for the CPS data, and are currently under construction for IHIS. Once the pointer variables are in place, project Web sites allows users to create new variables based on these links during the data extraction process, so customized variables such as educational attainment of mother or age of spouse become part of an individual's record.

Our years of experience inferring family interrelationships for the U.S. census and survey data could offer only rough guidance for constructing family pointer variables for IPUMS-International and NAPP. The international samples simply had too much variation in the reporting order of the enumerated persons, the categories of the relationship-to-head variable, the types of marital statuses, and the quality of the data.

The strategy employed by IPUMS-International to construct pointer variables in the midst of much crossnational variation is simple in explanation but complicated in execution. We carry out a series of passes over each household, looking first for links based on the strongest possible criteria followed by links using progressively more ambiguous information. The basic logic is consistent across samples, but because of variations in the underlying data, each linking stage must have the capacity for customization. By implementing this approach, IPUMS-International has created parent-child and spouse-spouse links for all samples containing entire households of persons, and these family pointer variables accommodate the details appropriate for each sample (e.g., identifying multiple wives in samples recognizing polygamy).

Web-Accessible Census and Survey Text

Text from census and survey forms that is associated with the appropriate integrated variable and accessible via a single mouse click is another important innovation. This utility is currently available for IPUMS-International, NAPP, and IHIS, is coming soon to IPUMS-CPS, and is available for IPUMS-USA under a slightly different methodology.

Under the methodology developed by IPUMS-International, graduate or undergraduate research assistants first type and format Word documents containing the text from census forms, instructions to enumerators, and survey questionnaires. Research staff then inserts XML tags into these documents, to indicate the beginning and ending points of the specific text associated with each source variable, and our software reads this metadata into a database. When these steps are complete, project Web sites use this information to provide a uniquely powerful documentation tool that displays all relevant census or survey text associated with a variable.

This innovation facilitates the data integration process, by making all relevant text visible to an MPC staff member who is harmonizing or writing documentation for a variable. The same material provides a valuable tool to the research community. For example, a researcher can examine the precise wording of the employment status question across over one hundred censuses, to determine whether variations in wording might affect her analysis. Access to this fundamental information frees researchers from depending completely on the judgment

of MPC staff about comparability issues. At the same time, this feature relieves pressure on us to write variable descriptions that anticipate all the ways in which researchers might wish to use the data—a practical impossibility.

Metadata Filtering

The large quantity of information compiled for MPC data projects—dozens of samples, thousands of variables, and hundreds of thousands of words—could easily overwhelm users of these databases. Our solution was to build into our Web dissemination system the capacity to filter information based on sample selections. If a researcher is interested in only a subset of countries (for IPUMS-International or NAPP) or in only samples from a particular time period (for IPUMS-International, NAPP, IPUMS-CPS, and IHIS), the Web system will display only the variables that exist in the samples selected by the user. More importantly, within each integrated variable description, the system displays only the question text, universe information, and codes and frequencies pertaining to the samples of interest. The sample selections can be changed at any point, so researchers have complete freedom to filter information as they navigate the system.

In the case of IPUMS-International, where users are least likely to be interested in all available samples, filtering also limits the text discussing comparability issues to the samples specified. Under this system, for example, an IPUMS-International user who is only interested in Europe need never consider the ways in which certain topics are handled differently in the developing world, but he or she can readily change focus from the entire continent to a single European country and back again.

Online Data Analysis

Several MPC data projects use high-speed tabulation software developed at University of California, Berkeley's Computer-Assisted Survey Methods Program to offer online data analysis capacity to researchers. With the online tabulator, users can quickly produce summary statistics and tables without downloading a data set. The tabulator supports filtering of cases and samples, recoding of variables, and selection of appropriate weights. Online data analysis is currently available in IPUMS-USA, IPUMS-CPS, NAPP, and IHIS, and will soon be added to IPUMS-International.

Variance Estimation for International Samples

The microdata samples in IPUMS-International employ a variety of sample designs: Most are clustered by household, many are stratified, and some are differentially weighted. Either the IPUMS-International samples are systematically drawn from full count data by project staff (or according to specifications supplied by IPUMS-International), or they are drawn by the statistical offices of the country of origin according to a variety of complex sample designs, including oversampling, clustering, and stratification.

All these sample designs have potential effects on multivariate standard error calculation, and all the data should be weighted to ensure representative estimates. The IPUMS-International project recently added sample design information to its Web site, to help users adjust for sample design characteristics in their estimation procedures. An extensive user note on sampling error and variance estimation describes how sample design affects sample precision, estimates the resulting differences in standard errors across variables and IPUMS-International samples, and discusses strategies for obtaining unbiased and efficient estimates of statistical significance. Elsewhere on the project Web site, a table on sampling error estimation provides a quick reference to sample design characteristics that affect standard error estimation and includes recommendations for improving estimation precision. A

second table, on sample descriptions, provides detailed information about the designs of the microdata samples.

Thanks to improvements in the analytical power of modern statistical software, it is now possible to incorporate information about sample design into estimation procedures. All major statistical software programs, including SAS, Stata, SPSS, and R, allow researchers to specify basic elements of complex sample design. For most analyses using IPUMS-International data, there is little risk of drawing invalid conclusions due to underestimated variance. When examining relationships on the margin of statistical significance, however, it may be wise to adjust for household clustering and weighting. In addition to providing online guidance for dealing with effects of clustering and weighting, the IPUMS-International project is developing a new variable that will help users account for the effects of stratification on sample variance. This new stratification variable will include information on explicit strata whenever such information is available and will also include geographic pseudostrata for the systematic samples following the procedure described by Michael Davern, Steven Ruggles, Tami Swenson, J. Trent Alexander, and J. Michael Oakes (2009).

Terra Populus: A Global Population/Environment Data Network (TerraPop)

With support from the National Science Foundation, Terra-Pop will provide global-scale data on human population characteristics, land use, land cover, and climate change. It will make these data interoperable across time and space, disseminate them to the public and to multiple research communities, and preserve these precious resources for future generations. Building on recent advancements in computer technology, TerraPop represents a level of data integration not previously available to researchers and beyond the realm of possibility just a decade ago.

Our understanding of the interactions of population and environment has been hampered by the dearth of internationally comparable data. Newly available population data that is closely integrated with data on the environment will allow us to describe the unfolding transformation of human and natural systems. By creating a framework for locating, analyzing, and visualizing the world's population and environment in time and space, TerraPop will provide unprecedented opportunities for investigating the agents of change, assessing their implications for human society and the environment, and developing policies to meet future challenges.

The TerraPop framework will provide tools for integrating, analyzing, and visualizing any data that have spatial or temporal dimensions. A plethora of high-quality environmental and population data sets are available, but they are widely dispersed, have incompatible or inadequate metadata, and have incompatible geographic identifiers. The project will enable researchers to identify and merge data from heterogeneous sources to study the relationships between human behavior and the natural world.

TerraPop will focus on four specific kinds of data: (1) census and survey microdata describing the characteristics of individuals and their families and households (including IPUMS-USA/CPS/International, NAPP, and IHIS); (2) aggregate census and survey data describing the characteristics of places, including aggregate population characteristics, land use, and land cover (including NHGIS and ISTADS); (3) remotely sensed data describing land cover and other environmental characteristics; and (4) climate data describing temperature, precipitation, and other climate-related variables. All four data types have an important temporal dimension; most of the data span the past five decades, and many sources reach back to the nineteenth century.

Implementation of the TerraPop framework will build on MPC's past experience of integrating census and survey microdata, documenting comparability and harmonization issues, and delivering both data and metadata in a dynamic, user-friendly, Web-based dissemination system. Lessons learned from constructing all the aforementioned databases and research tools will contribute to this new project.

TerraPop will indeed be big data. By integrating global data collections on land use and climate change with the major global collections of population data, TerraPop will constitute a unique international reference collection for investigating changes in the human-environment system. Moreover, TerraPop will support analysis from local scales to global scales and from historical epochs to contemporary epochs. As with all other MPC data products, TerraPop will be available to qualified researchers over the Internet free of charge.

Acknowledgments

Data infrastructure projects at the Minnesota Population Center, University of Minnesota, are supported by grants from the National Institutes of Health and the National Science Foundation: ATUS-X (University of Maryland, R01 HD053654; University of Minnesota, Z195701), IHIS (5 R01 HD046697), IPUMS-USA (1 R01 HD043392), IPUMS-International (5 R01 HD052110); and NHGIS (BCS 0648045).

References

- Alexander, JT.; Davern, M.; Stevenson, B. NBER Working Paper Series 15703. National Bureau of Economic Research; Cambridge, MA: 2010 January. Inaccurate age and sex data in the census PUMS files: Evidence and implications.
- Davern M, Ruggles S, Swenson T, Alexander JT, Oakes JM. Drawing statistical inferences from historical census data, 1850–1950. *Demography*. 2009; 46:589–603. [PubMed: 19771946]
- Fitch CA, Ruggles S. Building the National Historical Geographic Information System. *Historical Methods*. 2003; 36:41–51.
- Goeken R, Huynh L, Lynch TA, Vick R. New methods of census record linkage. *Historical Methods*. 2011; 44:7–14. [PubMed: 21566706]
- King ML, Tertilt M. IPUMS-CPS: An integrated version of the March Current Population Survey, 1962–2002. *Historical Methods*. 2003; 36:35–40.
- Noble P, Van Riper D, Ruggles S, Schroeder J, Hindman M. Harmonizing disparate data across time and place: The Integrated Spatio-Temporal Aggregate Data Series. *Historical Methods*. 44:79–85.
- Ruggles, S.; Alexander, JT.; Genadek, K.; Goeken, R.; Schroeder, MB.; Sobek, M. Integrated Public Use Microdata Series: Version 5.0 (Machine-readable database). Minneapolis: University of Minnesota; 2010.
- Ruggles S, King M, Levison D, McCaa R, Sobek M. IPUMS-International. *Historical Methods*. 2003; 36:9–20.
- Ruggles S, Roberts E, Sarkar S, Sobek M. The North Atlantic Population Project: Progress and prospects. *Historical Methods*. 2011; 44:1–6.
- Vick R, Huynh L. The effects of standardizing names for record linkage: Evidence from the United States and Norway. *Historical Methods*. 2011; 44:15–24.

FIGURE 2.

Complete-count census data from the late nineteenth- and early twentieth-century are available from the North Atlantic Population Project (NAPP) for the shaded countries. Recent data-recovery efforts have expanded NAPP beyond the rim of the North Atlantic.