Computers, Environment and Urban Systems xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Spatiotemporal aggregation for temporally extensive international microdata

Tracy A. Kugler [a,*], Steven M. Manson [b], Joshua R. Donato [a]

[a] Minnesota Population Center, University of Minnesota, 50 Willey Hall, 225 19th Avenue South, Minneapolis, MN 55455, USA
[b] Department of Geography, Environment, and Society, University of Minnesota, 414 Social Sciences, 267 19th Avenue South, Minneapolis, MN 55455, USA

ARTICLE INFO

Article history:
Received 19 October 2015
Received in revised form 8 July 2016
Accepted 22 July 2016
Available online xxxx

Keywords:
Regionalization
Cluster analysis
Census microdata

ABSTRACT

We describe a strategy for regionalizing subnational administrative units in conjunction with harmonizing changes in unit boundaries over time that can be applied to provide small-area geographic identifiers for census microdata. The availability of small-area identifiers blends the flexibility of individual microdata with the spatial specificity of aggregate data. Regionalizing microdata by administrative units poses a number of challenges, such as the need to aggregate individual scale data in a way that ensures confidentiality and issues arising from changing spatial boundaries over time. We describe a regionalization and harmonization strategy that creates units that satisfy spatial and other constraints while maximizing the number of units in a way that supports policy and research use. We describe this regionalization strategy for three test cases of Malawi, Brazil, and the United States. We test different algorithms and develop a semi-automated strategy for regionalization that meets data restrictions, computation, and data demands from end users.

1. Introduction

Researchers working with census data typically work with either aggregate data tables published by the national statistical offices or with microdata samples of individual-level records. Both of these data structures have important advantages and disadvantages. Aggregate data are commonly provided for small geographic units such as blocks and tracts. However, researchers using aggregate data are constrained to the set of tabulations published by the national statistical office, which may not be ideal for addressing their research questions. Microdata enable researchers to perform individual-level analyses incorporating the full richness of characteristics and their interrelationships, or to create customized tabulations of specific populations and characteristics of interest. However, the geographic identifiers available on microdata to date have been limited, primarily as a strategy to preserve confidentiality of individual responses. The lack of spatial specificity available for microdata has hampered researchers' ability to investigate questions related to local context using microdata.

As interest in spatially-explicit methods for demography and other social sciences grows (cf. the new journal *Spatial Demography* and Howell, Porter, and Matthews, 2016), the availability of better geographic identifiers for microdata becomes increasingly important. We present a method of regionalization that enables dissemination of microdata with finer scale geographic identifiers, handles changes in unit boundaries over time, and continues to preserve confidentiality. This regionalization-harmonization strategy is used within the Terra Populus (TerraPop) data access system (Minnesota Population Center, 2015b), further facilitating integration of population and environmental data.

National statistical offices often publish tables of aggregate data for small geographic units, but the detail and flexibility of the population characteristics in such tables is limited. The United States publishes aggregate data tables for units as small as census blocks, but only basic characteristics and few cross-tabulations are available at this level. For census tracts, most characteristics are available, but the specific cross-tabulation required to address a particular research question may not be published in order to preserve confidentiality. Internationally, national statistical offices vary widely in the geographic levels and characteristic details published in aggregate data. Researchers can take advantage of the geographic detail in aggregate data tables to study local neighborhood effects such as changes in diversity or segregation, although the effects of arbitrary zonations of these areal units is subject of much interest (Kramer, Cooper, Drews-Botsch, Waller, and Hogue, 2010; Lee et al., 2008). But in some cases, researchers' ability to study relationships among population characteristics has been hampered by the particular tabulations published. Specifically, using aggregate data to explore the relationships between social variables suffers from two key flaws. One, comparing two or more variables in the aggregate, such as income and education, is subject to the ecological fallacy because we know little to nothing about how these measures vary by individual, although there are ongoing efforts to address this issue (Anselin, 2002; King, 1997). Two, aggregate data do not lend themselves to the creation of many cross-tabulations, meaning that it is difficult to examine whether two variables are related controlling for a third (e.g., a

* Corresponding author.
E-mail addresses: takugler@umn.edu (T.A. Kugler), manson@umn.edu (S.M. Manson), donat050@umn.edu (J.R. Donato).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.07.007
0198-9715/

relationship between education and income that differs by age or gender), a problem of long-standing interest in the social sciences (Duncan and Davis, 1953).

Conversely, census microdata provide great detail and flexibility in terms of population characteristics, but lack geographic specificity. The availability of the IPUMS family of data access systems, including IPUMS-USA (Ruggles, Genadek, Goeken, Grover, and Sobek, 2015) for microdata provided by the U.S. Census Bureau and IPUMS-International (Minnesota Population Center, 2015a) providing census microdata from 82 countries, has greatly increased the accessibility and utilization of these data. Researchers use IPUMS data to study a wide array of topics, including fertility trends, educational inequality, family formation, labor markets, public health, and others (Bailey and Collins, 2011; Bleakley, 2010; McDaniel, DiPrete, Buchmann, and Shwed, 2011; Spijker and Esteve, 2011). However, in order to protect individual respondents from being reidentified based on their characteristics and where they live, fine scale geographic identifiers are generally not included with microdata records. As a result, it is difficult to utilize microdata to study spatial relationships or neighborhood effects. Studies utilizing microdata in a fine-scale spatial context are typically dependent on access to confidential data or other special datasets (e.g., Hamilton and Phaneuf, 2015; Kramer, 2016).

We describe a method that helps to blend the advantages of both aggregate data and microdata by providing more detailed geographic identifiers than had previously been available for microdata, particularly in an international context. Our method takes advantage of automated regionalization to group spatially contiguous sets of small units together to meet population thresholds designed to maintain confidentiality. The regionalization process is embedded within a workflow that retains important characteristics of spatial organization, such as hierarchy. In addition, the workflow enables dissemination of both geographic units that have been harmonized to be stable over time, for researchers studying change, and units tied to a particular census year, for researchers desiring even greater geographic detail.

The TerraPop data access system further enhances the utility of microdata with small-area geographic identifiers by providing location-based integration with other population and environmental datasets (Kugler et al., 2015). TerraPop allows users to attach social or environmental contextual variables to microdata from IPUMS-International based on the identified geographic unit in which the individual lives. Microdata enhanced with contextual variables enables individual- and household-level analysis of population-environment interactions, such as migration and climate change or agriculture and deforestation. The system also allows users to construct customized aggregate data tables, drawing on any combination of characteristics in the microdata. The rows in a customized tabulation consist of the geographic units identified on the microdata records. Small-area data for customized tabulations enables studies focusing on specific populations, such as school-age children, female-headed households, or workers employed in manufacturing, or combinations of characteristics.

2. Background

2.1. Census microdata

Almost all census data have their origin in information collected from individual respondents. Each microdata record includes the full set of an individual's responses, including characteristics such as age, sex, marital status, educational attainment, employment status, occupation, relationship to others in the household, and more. Individuals are grouped into households, and household records typically carry additional information about the housing unit, such as type of water supply, presence of specific utilities and amenities, number of rooms in the house, and, importantly, geographic identifiers.

The geographic identifiers on microdata take the form of codes that identify the administrative or statistical units in which the household is located. Most countries have what can be termed an "inclusive scalar hierarchy" (Gibson, Ostrom, and Ahn, 2000) of administrative units, in that higher level units act as containers for smaller units (e.g., the nesting structure of states, counties, tracts, block groups). These levels may generically be referred to numerically, with level one being the largest set of units (e.g., states or provinces) and subsequent levels being smaller administrative or statistical units (e.g., level two may be counties or districts). Original microdata usually include a series of identifiers, specifying units at each level of the hierarchy. Geographic identifiers offer explanatory power because social regimes and environmental setting often differ from one region to another, as seen in regionally varying election results, tax regimes, cultural identification, or ecological and biophysical characteristics. Location also provides a means to link microdata records to other data, as the TerraPop data access system does.

A key challenge in providing microdata to researchers is preserving confidentiality. The primary means of preserving confidentiality is the use of sampling to select only a fraction of the individual or household records for dissemination. Data providers also employ a number of additional strategies to ensure confidentiality, including randomizing the sequence of dwellings within geographic units, and randomly swapping a small fraction of cases across geographic units (Fienberg, 2005; Frend et al., 2012; McCaa and Esteve, 2005; Zayatz, 2005). One important strategy for protecting confidentiality is limiting the spatial detail conveyed by the geographic identifiers included on each record. Spatial detail is often limited to identifying places that have populations above a specified threshold. The general idea is that a place with at least a minimum number of people would at least plausibly have two individuals with identical characteristics captured by the census questions, making it impossible to positively identify a specific individual.

2.2. Geographic identifiers for microdata

The two major disseminators of census microdata freely available for research are the U.S. Census Bureau (for U.S. census microdata) and the Minnesota Population Center's IPUMS family of projects (for census data from both the U.S. and more than 80 other countries). Both organizations have existing methods for developing geographic identifiers for microdata while preserving confidentiality. However, each approach presents challenges in terms of scalability or geographic properties.

The U.S. Census Bureau disseminates public use microdata with geographic identifiers for public use microdata areas (PUMAs) designed specifically for the purpose. PUMAs are defined by State Data Centers, following requirements and guidelines set forth by the Census Bureau (U.S. Census Bureau, 2011). For the 2010 census, PUMAs were required to be built from counties or census tracts, have a minimum population of 100,000 persons over the period from 2000 to 2010, be contiguous, and not cross state boundaries. In addition, the Census Bureau provided a series of guidelines describing desired, but not required, properties of PUMAs, such as maximizing the number of PUMAs within a state and not splitting American Indian Reservations.

The disadvantage of the U.S. Census Bureau approach in terms of general applicability is that the geographic units are manually defined based on local knowledge. Developing geographic units for dissemination with the entire IPUMS-International microdata collection by hand is simply not feasible, nor would it be possible to assemble local knowledge to effectively inform the process. In addition, geographic units defined manually are subject to inconsistencies, are difficult to document, and are not replicable.

Prior to the advent of TerraPop, IPUMS-International had provided geographic identifiers for existing administrative units meeting a minimum population threshold. The memoranda of agreement with most national statistical offices providing microdata for IPUMS-International stipulate that places with populations of fewer than 20,000 persons will not be identified on the microdata. Nearly all first-level units meet this threshold, as do many second-level units. Therefore, IPUMS-

Please cite this article as: Kugler, T.A., et al., Spatiotemporal aggregation for temporally extensive international microdata, *Computers, Environment and Urban Systems* (2016), http://dx.doi.org/10.1016/j.compenvurbsys.2016.07.007