Harmonized census geography and spatio-temporal analysis: Gender equality and empowerment of women in Africa

Sula Sarkar†
Minnesota Population Center
University of Minnesota

Lara Cleveland
Minnesota Population Center
University of Minnesota

Majory Silisyene
Minnesota Population Center
Natural Resources Science and Management
University of Minnesota

Matthew Sobek
Minnesota Population Center
University of Minnesota

July 2016

Working Paper No. 2016-3
https://doi.org/10.18128/MPC2016-3

†Correspondence should be directed to:
Sula Sarkar
University of Minnesota, 50 Willey Hall, 225 19th Ave S., Minneapolis, MN 55455
e-mail: sanb0027@umn.edu , phone: 612-624-5818, fax:612-626-8375

Harmonized census geography and spatio-temporal analysis: gender equality and empowerment of women in Africa

Sula Sarkar, Lara Cleveland, Majory Silisyene, and Matthew Sobek — University of Minnesota

Abstract: *Changes in administrative boundaries pose major challenges for spatio-temporal population research. Researchers interested in change over time need to hold space constant to study contextual or spatial effects on behaviors and outcomes. Boundary changes risk polluting their analyses with artifacts that obscure real changes that may have occurred. This paper describes the method by which spatially consistent geographic units have been constructed in the IPUMS-International census data collection for several countries over a fifty year period. We illustrate the utility of spatially consistent units by exploring progress toward UN Millennium Development Goals in a number of African countries at low levels of geography: specifically the goals to "promote gender equality and empower women." The analysis shows progress towards goals, but the pattern of growth differs markedly both across and within countries. We show how the use of harmonized geographic units facilitates comparative metrics.*

INTRODUCTION

Changes in administrative boundaries pose a major challenge for spatio-temporal population research. Researchers interested in change over time need to hold space constant to study contextual or spatial effects on behaviors and outcomes. Boundary changes risk polluting their analyses with artifacts that obscure real changes that may have occurred. This paper describes the method by which spatially consistent geographic units have been constructed in the IPUMS-International census data collection for several countries over a fifty year period. Low-level geographic units are grouped into temporally compatible base units that are spatially consistent across all census years. Regionalization (combining) techniques are applied to create spatio-temporally harmonized units that meet the 20,000 population threshold required for public dissemination of the data. The base units are then disaggregated to create year-specific units that still meet the necessary population threshold requirement. We illustrate the utility of the harmonized units by exploring progress toward UN Millennium Development Goals (MDGs) in a number of African countries at the sub-national level: specifically the goals to "promote gender equality and empower women." The analysis shows generalized increases in the number of women completing secondary education and participating in the labor force, but the pattern of growth differs

markedly both across and within countries. Disaggregation of national trends into regional or local trends highlights areas of change and stasis. The example underscores the need for additional tools that facilitate spatio-temporal comparison. We show how the use of harmonized geographic units facilitates and improves comparative metrics.

THE DATA: SPATIAL AND TEMPORAL CHALLENGES AND LIMITATIONS

Data

The Integrated Public use Microdata Series, International (IPUMS) is the world's largest publicly accessible population database. It currently includes sample data for 258 censuses from 79 countries. The collection grows by approximately 20-25 samples every year by adding data from new partner countries and by extending the collection from existing partners by adding data from the most recent censuses. IPUMS is comprised of microdata, wherein each record represents a person (organized into households) for whom all individual census characteristics are known. The data include variables representing a broad range of population characteristics, including fertility, nuptiality, life-course transitions, migration, disability, labor-force participation, occupational structure, education, ethnicity, and household composition (Ruggles et al. 2003; Sobek et al. 2011). Censuses are taken at fairly regular intervals, commonly every 10 years or so, and data in IPUMS are available for multiple census years for most countries in the collection. Use of the IPUMS data has grown at a dramatic rate as researchers have discovered the value of this easily accessible, user-friendly collection, and as the number of countries in the database has grown.

IPUMS makes a significant contribution to population research by optimizing data for cross-temporal and cross-national comparative analyses. Multiple census years are available for most countries in the database, and variables are harmonized across IPUMS samples so that coding is consistent at all times and in all places. A dissemination system allows users to build custom data extracts that pool data from different countries and across census years. Variable harmonization is a laborious process, requiring hours of research and analysis at the variable, sample, and national level. The work presents numerous interpretive challenges and demands careful documentation about changes in definitions of concepts represented in the coding of the variables. Users of the IPUMS are alerted to changes in meaning, ranging from slight to significant, across time and country through integrated and structured metadata available via the website and in downloadable files (Minnesota Population Center 2014).

Geographic information is typically recorded for place of residence at the household level and for place of birth and place of previous residence (in varying intervals) at the person level. Occasionally, censuses also record place of work or school. In the past, IPUMS performed only rudimentary harmonization of geographic variables, which presented some of the most difficult challenges in the development of the data series. With the most recent data release in summer 2014, IPUMS has initiated a thorough overhaul of sub-national geography using the techniques described in this paper.

Challenges of Space and Time

Geographers are commonly faced with estimation challenges resulting from issues of temporal and spatial scale. A central challenge in dealing with scale is that data measures calculated at different spatial or temporal scales may convey different information. Changes in administrative boundaries over time complicate estimation and analysis in comparative spatio-temporal research. Users of census microdata are limited by the timing of censuses (typically every 5 or 10 years) and by the unit levels identified in the data (typically administrative divisions within country).

The modifiable area unit problem (MAUP) is a classic dilemma in geography and is relevant to analyses of census data where geography is measured only by areas defined by boundaries at a limited number of administrative levels. According to Openshaw (1984), the MAUP is composed of two separate but closely related problems (Openshaw and Taylor 1979; Openshaw 1984). First is the area problem in which analytic results can vary at different levels of aggregation, i.e., when areal units are progressively aggregated into fewer and larger units for analysis. In other words a change in scale of analysis can alter the results. The second aspect of the MAUP is the aggregation problem, referring to variation in results due to the use of alternative aggregation schemes (or calculation methods) at equal or similar scales. This problem arises due to uncertainty about how best to summarize, or aggregate, data across the available identified units (Clark and Avery 2010; MacEachren 2004).

Scale: The area problem presents itself when the appropriate area of study is unclear or under-theorized. In the case of census data, this problem can arise if appropriate units are not identifiable in the data. Census offices record geographic information at the administrative unit level, providing coded data and labels (place names). Each record in the census data includes identifiers (codes) for one or more administrative level units. Administrative levels are often hierarchically coded to preserve the nested logic of the units. In common geographic terminology used by the United Nations and many

other institutions, the country is considered administrative level 0. Within country, administrative level 1 represents the largest sub-national division (e.g., states in the United States, Germany, Brazil or province in Kenya, Pakistan, etc.) that exhaustively partitions the country. The 2nd administrative level (e.g., counties in the United States) exhaustively partitions units of the 1st level. Most countries have progressively lower levels units of geography (3rd, 4th and beyond) (Kugler et al. 2015). The divisions tend to correspond to geopolitical divisions indicating some kind of administrative control. However, some low geographic units identified in census data are purely for statistical or census administrative (rather than political administrative) purposes. The problem of scale is further complicated by confidentiality considerations. In order to preserve confidentiality, and in accordance with National Statistical Office partnership agreements, IPUMS identifies units large enough to meet a 20,000 person threshold in the most recent census samples.

According to Openshaw, a perfect homogeneous zoning system would enable researchers to avoid the MAUP, but such homogeneous spatial units are rare (Openshaw 1984). While such units constitute an impossible ideal for census data, the availability of very low level geographic identifiers in some census samples permits the construction of a set of best available units. The presence of identifying codes for low levels of geography in the microdata makes it possible for researchers to study population characteristics at several geographical scales, thereby providing checks against the area problem. Creating thoroughly documented and verified geographic units and providing the corresponding GIS shapefiles for at least two levels of sub-national geographic units significantly improves the extent to which meaningful geographic research can be conducted. Changes in administrative boundaries over time, however, complicate comparative spatio-temporal research and are discussed below.

Estimation: The second aspect of the MAUP, the aggregation problem, is less problematic for users of census microdata. Census microdata samples are typically comprised of individuals organized into, and sampled at, the household-level. Census microdata provide a great deal of flexibility in the calculation of summary statistics, provided users are familiar with the statistical software techniques to carry out such calculations. Users are also less prone to ecological fallacy when they can customize aggregations or combine geographic units in accordance with the precise requirements dictated by their research questions. Extensive metadata documentation in IPUMS aids researchers in understanding the characteristics in the data, thereby facilitating the use of appropriate methods.

Cross-temporal comparison: Finally, one of the biggest hurdles to cross-temporal spatial analysis using census data is the question of whether, and to what extent, geographic boundaries change across census years. Until now, little has been done to verify the spatial areas corresponding to coded units in the census microdata. Even less has been done to research spatial changes across time.[1] This is not surprising given the limited access researchers have traditionally had to census microdata. The challenges of estimation are compounded by the addition of time to an analysis. Researchers must determine the extent to which consistency of spatial area is essential to their analytic technique. In the study of an identified "place," researchers must decide whether the analysis is relevant to the political unit defined by the name and governing structure of an area regardless of its spatial extent, or whether the analysis depends upon a consistent footprint from one time period to the next. Often, the latter is essential, and spatial consistency must be imposed (Haining 2003). Subnational administrative units are central to spatial demographic analysis because they act as a common denominator for an array of social and demographic analysis.

Geographic harmonization presents many challenges. Geographic units are identified by a code and label (place name). For all but the highest level units, IPUMS may receive only the codes. Codes and labels may or may not change from one census year to the next and changes may or may not reflect spatial changes to the administrative unit. More importantly, consistency of codes and labels is no guarantee of spatial continuity across time. Census offices rarely provide maps corresponding to the census units, making it difficult to determine the extent to which boundaries have changed from one census to the next. IPUMS geographic work over recent years (methods detailed below) has sought to remedy these deficiencies. The IPUMS team has developed a method for creating spatially consistent units in the microdata, starting with the first and second administrative units identified in the census samples. With the summer data release of 2015, the project will add a number of Integrated Statistical Areas (ISA) geographic variables at both administrative levels for about half the countries in the collection. The project will also release updated and more accurate year-specific geographic variables. GIS boundary files corresponding to all geographic variables will also be available for download. Improved geographic variables for most remaining countries will be released in 2016.

[1] Important exceptions such as UNSALB (UN Geographic Information Working Group 2014) and Statoids (Law 2015) exist. IPUMS use of these resources is mentioned in the Methods section.

METHODS

Given the rise in digital mapping capabilities and spatial analytical technologies, social science research increasingly calls for consideration of space (MacEachren 2004). Because of this growing salience, the limited geographic information in the IPUMS census data collection had to be remedied. The work involves extensive metadata acquisition, research, and verification (acquisition and correspondence); the creation of small-area building blocks that cover consistent spatial extent over time (harmonization); the testing and implementation of techniques to group spatial units to meet the 20,000 person threshold (regionalization); and the development of GIS shapefiles and variables (map and variable creation). The most technically and methodologically intense portion of this work involves regionalization. We are especially interested in what Guo (Guo and Wang 2011; Guo 2008) terms the population regionalization problem, which involves regionalizing subnational administrative units while accounting for their attendant attributes. In what follows, we explain our process for creating Integrated Statistical Areas (ISAs) keeping in mind some of the geographic analytic challenges outlined above.

Data-map acquisition and correspondence

The first and most fundamental task involves collecting digital maps from partner countries and statistical agencies, when available, or from open source and online digital sources, when necessary. Three well-known, freely available, and GIS-compatible administrative unit sources include the Global Administrative Unit Layers (GAUL) dataset (Food and Agriculture Organization 2006), the United Nations Second Administrative Level Boundaries (UNSALB) (UN Geographic Information Working Group 2014), and the Global Administrative Areas (GADM) dataset (Robert Hijmna's Laboratory 2014). Available digital maps mostly reflect current political boundaries and seldom historical boundaries corresponding to previous censuses. When digital GIS maps are not available, we scan, catalog, and document paper maps from published census volumes and reports. The paper maps from previous censuses are then georeferenced to modern digital boundary files (U.S. State Department, Office of the Geographer 2014) and digitized to create historical boundary files that match censuses in IPUMS.

Next, digital historical boundaries are matched to the geographical codes from the IPUMS samples. Where codes and maps do not match (which is true more often than we would have expected), we refer to published census volumes for a comprehensive match of digital maps to census codes. Matching map codes to census codes must be implemented for every IPUMS sample, because boundaries of base units and enumerated regions change over time. Some changes are as simple as division of a base unit into

two units; others are more complex, involving shifting boundaries or even the wholescale redrawing of boundaries from one census to another.

Harmonization

Harmonization is the process by which we create consistent units across time using lower level administrative units as building blocks. Where geographic boundaries of modern units do not align with historical census units because of boundary changes, larger aggregated units are created that remain stable over time. We refer to this process as harmonization of geographic boundaries. If units split or merged, the harmonized unit will have the boundaries of the largest version of the unit; if a territory is redistributed between two or more units, the units are combined. In a few cases, particularly in those countries that have experienced significant political turmoil, boundaries have been redrawn to such an extent that harmonization is nearly impossible. In those few cases, we have had to either create sets of consistent units that are available only in limited (pre-transition and post-transition) time spans or provide only year-specific geographic units.

Regionalization

IPUMS distributes integrated microdata about individuals and households only by agreement of collaborating national statistical offices and under the strictest of confidence. Limiting geographic detail is one of the primary means statistical offices employ to ensure confidentiality. If harmonized geographical units have less than 20,000 populations, they are grouped until they exceed that threshold. We refer to this process as regionalization. Regionalization is not required for samples whose total populations at the first and second level of geography are greater than 20,000 persons.

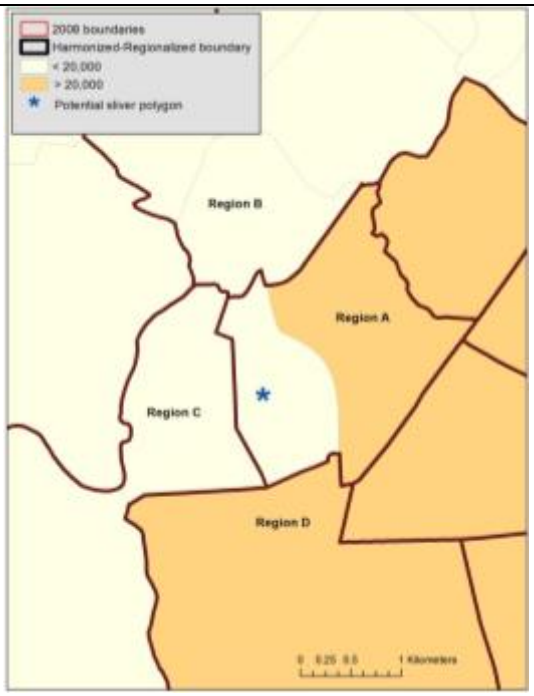
IPUMS uses regionalization (also known as segmentation or aggregation), a subset of cluster analysis, to group census units in a way that minimizes differences within groups and maximizes difference between groups. Spatial regionalization is similar to cluster analysis but it involves classifying spatial units or areas (Martin 2003). It focuses on the problem of grouping spatial entities, such as those defined by administrative boundaries. Spatial regionalization seeks to satisfy inherently spatial conditions, such as ensuring aggregations are spatially contiguous, meeting a minimum area, or maximizing attribute similarity within regions and maximizing dissimilarity between aggregations.

Guo (Guo 2008) describes the many domains that face regionalization problems, ranging from climate research to urbanization to health policy. He goes on to describe how regionalization methods fall into either non-spatial or spatial clustering methods. Non-spatial clustering methods draw on aspatial attributes to group similar base units, such as aggregating census tracts according to average household income or ethnic composition, or using statistical models to determine how attributes can explain differences between base units. Guo's spatial methods go one step further by trying to satisfy a given spatial requirement such as adjacency or contiguity. The computational implementation of aspatial and spatial grouping methods varies a great deal, ranging from statistical and mathematic approaches to geocomputational techniques like artificial neural networks, self-organizing maps, and evolutionary algorithms (Kauko 2004; Martin 2003; Painho 2000).

In addition to the hard constraints of harmonization and regionalization, we seek to optimize additional desired characteristics such as contiguity (where base units in a region should be adjacent to at least one other unit) and compactness (where the harmonized region should be as close to circular as possible as opposed to elongated and irregular) when creating ISAs. We also maintain hierarchical structure in the census units wherever possible. Geographic boundaries represent a system where subunits (second level of geography) are nested within larger units (first level of geography). Spatial and hierarchical ordering also provides flexibility to users with respect to choosing their scale of analysis: analysis at the regional scale, first, or second level of geography through time.

Our processes of harmonization and regionalization proceed in parallel to avoid producing identifiable combinations of units across multiple levels of geography that have populations less than 20,000. Such identifiable combinations of units are referred to as “slivers” where individual households could potentially be identified. Figure 1 helps illustrate the sliver potential. All of the lighter shaded units are in need of combining to meet the population size threshold. The starred unit could have been grouped with any of the regions A through D, but was joined with Region A based on the regionalization algorithm. In releasing a year-specific (non-harmonized) geographic variable, we must account for singleton small areas to ensure that they remain combined within the same region (Region A in Figure 1) for all subsequent years. Releasing the large portion of Region A as a stand-alone unit, would reveal the starred unit. If we combined the starred unit with another adjacent small unit from a different region in a different census year, we would be, in effect, make it possible to uniquely identify the starred unit.

Figure 1: Potential identification of small population areas (slivers) in the harmonization process



We use the Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning (REDCAP) algorithm and accompanying software (Guo and Wang 2011; Guo 2008). Regionalization is conducted using population density, such that the algorithm combines geographic units that have similar population compactness. Population density is used because it is universally available and because many other characteristics of importance are highly correlated with density. REDCAP enforces spatial contiguity and creates regions while optimizing the sum of squared differences.

Both first and second level administrative units are taken into consideration for creation of ISAs. For most countries, regionalization is typically unnecessary at the first administrative level because these units generally have relatively large populations. At the second administrative level, however, regionalization is required for many countries because many of them have populations below 20,000. Regionalization is constrained so that only units within the same higher-level unit may be combined. Units that are both harmonized and regionalized are prevented from crossing the boundaries of higher-level units, thus preserving spatial and hierarchical ordering. All changes in boundaries at the first administrative level are documented in the IPUMS geography variable descriptions. ISAs created by IPUMS are sometimes substantially larger than the places that can be identified in a single census year for a country, but they are stable over time. The main purpose for ISAs is to facilitate research over time.

Map and Variable Creation

IPUMS offers a set of custom-created ISA variables along with their corresponding GIS shapefiles. The GIS shapefiles include an unique identifier, so that users can map IPUMS data summarized at the first or second level of geography. The website provides extensive documentation about how units have been

harmonized and regionalized to accommodate boundary changes over time. Along with spatially consistent boundaries through time (at the first and second level of geography), IPUMS also provides year-specific census geographic variables and boundaries. Users can request ISA geographic variables, year-specific variables, or both when building a data extract. Year-specific variables are ideal for users studying one specific place and time. Year-specific variables provide greater detail than spatially harmonized variables because they do not need to account for changes over time by aggregating units together that otherwise meet the 20,000 population threshold. Year-specific regionalized boundaries are created by relaxing the harmonization constraint. Instead of using first administrative level units as the topmost hierarchy, spatially consistent ISAs are used. This allows us to provide units that were harmonized to be disaggregated based on year-specific boundaries. Producing year-specific geography in this manner prevents the creation of slivers (see Figure 1) across year-specific and harmonized geography, while providing greater geographic detail than the harmonized shapefiles.

IMPLEMENTATION AND DISCUSSION - CASE STUDY

The sections that follow illustrate the utility of ISAs while examining gender inequality at the national and sub-national level for select countries in Africa. We focus on Goal 3 of the Millennium Development Goals (MDGs) as specified by the United Nations (UN) - "Promote gender equality and empower women", specifically targeting changes in gender-based educational equality and women's employment (United Nation 2000). We use harmonized (and year-specific) geographic variables from IPUMS to illustrate how research is shaped by the availability of different geographic units in the microdata. We measure progress on the MDGs at the national and sub-national levels for select countries in Africa. We demonstrate the need for a spatially consistent geographical footprint for some analyses. We also indicate when year-specific census geography should be used in conjunction with the spatially consistent ISA geographies.

Non-geographic variables in IPUMS are coded consistently across time and country. These harmonized data can be used to measure change over time and across space with respect to several of the MDG indicators. In this paper we calculate and map measures of Goals 3.1) gender disparity in primary and secondary education; and 3.2) share of women in wage employment in the non-agricultural sector.

We have categorized the data into three broad time periods according to whether they were collected prior to the implementation of the Millennium Development Goals, within the first 5 years following

implementation, or more than 5 years after implementation. It is reasonable to expect that MDG programs would have had little time to take effect during the middle period but might reasonably be expected to have had an impact during the latest period. Although the use of time in these examples is similar to assessing a treatment effect, we are not striving to establish causality. Rather, assessment of progress toward goals is geared more toward emphasis on improving living conditions for people around the world and ascertaining what work remains at given points in time.

The MDG measures are calculated from the IPUMS microdata only in those countries with at least two censuses containing the requisite variables. Censuses were conducted in different years from one country to another. At the national level, five African countries met the variable and time-period requirements for goal 3.1 and seven for goal 3.2. At the sub-national level, we recalculate measures for Mali and Malawi, mapping the data to demonstrate the ISA regions and GIS shapefiles. We focus on these two countries in order to show how progress at the national level is differentially distributed at sub-national levels.

Gender Equity in Educational Enrollment

We first examine Goal 3.1: "Eliminate gender disparity in primary and secondary education, preferably by 2005, and in all levels of education no later than 2015." (United Nations, 2003). The UN guidelines recommend operationalizing the assessment of this goal as the ratio of girls to boys currently attending primary and secondary education.

Primary school enrollment: Results show improvement in the ratio of primary school enrollment for nearly all countries (Table 1). For both Malawi and Zambia, the ratio of girls to boys who were enrolled in primary school during late MDG implementation either approached or surpassed gender parity. For Mali, the ratio of girls to boys who were enrolled in primary school remained far from the MDG target, while that of Ghana remained constant. Due to the timing of censuses in Senegal, data availability is not temporally ideal for measuring MDG progress. Rather, the data describe change in primary school enrollment from a very early period (1988) to the early post-MDG implementation (2002). We observe gender parity in primary school enrollment by 2002 and must assume that Senegal had implemented changes to facilitate female primary school enrollment prior to the UN goal establishment. Overall, country-level analysis shows that almost every country is moving towards gender parity in primary

school enrollment. While the observed changes in the ratio are not significant for Malawi and Zambia, the ratios for these countries were already close to one before MDGs were implemented.

Table 1: Ratios of girls to boys in primary school

Country	1988 to 2000 Pre-MDG	2001 to 2006 Early MDG	2007 to 2011 Late MDG
Ghana	0.97 (2000)	--	0.95 (2010)
Malawi	0.95 (1998)	--	1.03 (2008)
Mali	0.72 (1998)	--	0.86 (2009)
Senegal	0.70 (1988)	1.00 (2002)	--
Zambia	0.95 (2000)	--	1.00 (2010)

Secondary school enrollment: The ratio of girls to boys enrolled in secondary schools also increased significantly in all countries (Table 2). Disparities in secondary enrollment prior to MDG implementation were far greater than disparities in primarily school enrollment. Gains were greater in all countries, but there was also more room for improvement at the secondary level. Only in Malawi is secondary enrollment approaching parity between girls and boys.

Table 2: Ratios of girls to boys in secondary school

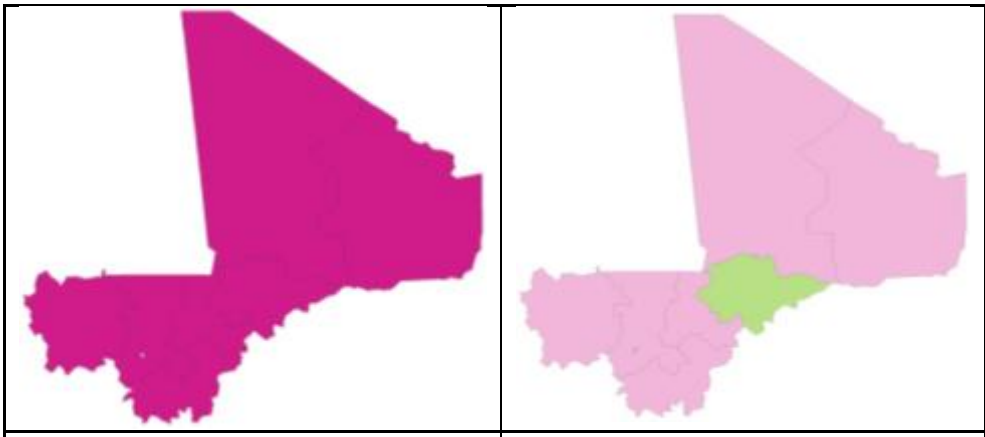
Country	1988 to 2000 Pre-MDG	2001 to 2006 Early MDG	2007 to 2011 Late MDG
Ghana	0.86 (2000)	--	0.89 (2010)
Malawi	0.64 (1998)	--	0.96 (2008)
Mali	0.57 (1998)	--	0.69 (2009)
Senegal	0.58 (1988)	0.76 (2002)	--
Zambia	0.83 (2000)	--	0.90 (2010)

Visualizing Sub-national Educational Enrollment: IPUMS harmonized geographic variables enable us to calculate the same measures at sub-national levels, holding spatial units constant across sample years. In the countries we were able to explore in depth, we found that increases were concentrated in certain sub-national geographic areas. Enrollment ratios in some geographic units increased to one or higher while remained constant or declined in others. Figures 2 and 3 show changes in enrollment ratios at the first and second administrative unit levels for Mali and Malawi respectively.

Holding space constant is critical in measuring progress toward MDG goals at sub-national levels; units that have changed boundaries cannot be compared across time in any meaningful way. In Figure 2, we see that areas in the central region of Mali made the most progress in educational gender equity and may even be favoring female enrollment, while other areas of the country had more modest gains than

the overall country measures imply. As shown in Figure 3, the harmonized second-level geographic units of Malawi (Figure 3, Map B) experienced a moderate increase in secondary school enrollment ratios of girls to boys after MDG implementation, and rates vary across Traditional Authorities.

Figure 2. Gender disparity in primary education, Mali Regions 1998-2009



1998 - Pre-implementation of MDG | 2009 - MDG implemented

Ratio of boys to girls who have some primary education, Mali regions in 1998 and 2009. A value of 1 means gender parity. Note progress towards goals from 1998 to 2009

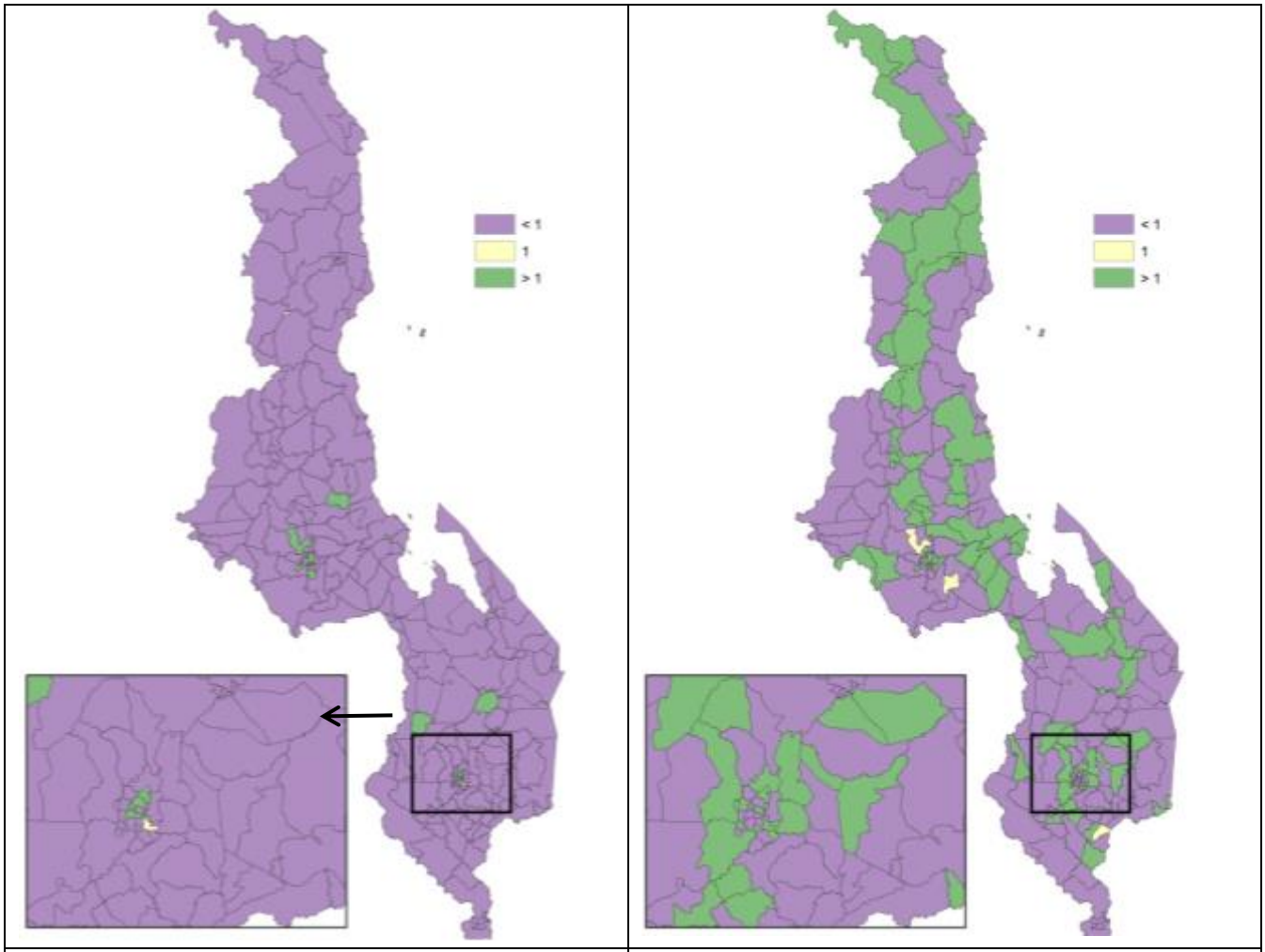
Women in Non-agricultural Wage Employment

National rates: The recommended measure of Goal 3.2 is the share of female workers in wage employment in the non-agricultural sector as a percent of total employment (United Nation 2000). The share of women in the non-agricultural employment sector has increased significantly across several African countries since implementation of the MDGs. Despite this increase, however, the proportion of women in the non-agricultural sector remains far from parity. As presented in Table 3, significant increases have occurred in Egypt, Malawi, Mali, and Zambia. Meanwhile, in Ghana, Morocco, and South Africa, the female employment share has remained almost constant.

Table 3: Percent female in non-agricultural wage employment (MDG goal 3.2)

Country	1988 to 2000 Pre-MDG	2001 to 2006 Early MDG	2007 to 2011 Late MDG
Egypt	18.9 (1996)	21.2 (2006)	--
Ghana	34.4 (2000)	--	33.7 (2010)
Malawi	19.4 (1998)	--	24.6 (2008)
Mali	25.3 (1998)	--	46.2 (2009)
Morocco	23.7 (1994)	24.4 (2004)	--
South Africa	--	44.9 (2001)	43.7 (2007)
Zambia	23.9 (2000)	--	28.0 (2010)

Figure 3. Gender disparity in secondary education, Malawi Traditional Areas 1998-2008



MAP A: 1998 census - Pre-implementation of MDG | **MAP B: 2008 census - MDG implemented**

Note: The boundaries of the Traditional Areas are spatially consistent through the two census time periods. The inset map shows the urban area of Blantyre district.
Note: A value of 1 means gender parity. More green in Map B shows progress towards goals from 1998 to 2008.

Sub-national mapping of female labor force participation : To explore women's employment progress within countries, we map sub-national change for Mali and Malawi, the two countries that indicate greatest progress in achieving MDG indicator 3.2. In both cases we use visual representations of performance toward the gender employment goal at the first geographic level. In Mali, both the national and first-level analyses (Figure 4) show significant progress towards achieving indicator 3.2. However, while all regions show considerable progress, the central area has the highest rates and the west lags behind the rest of the country. When we compare female employment in harmonized versus non-harmonized units of level 1 geography (Figure 4), there is not much difference in the units that split

between the two census years - i.e., between Kidal and Gao. In this case, it may not matter whether a researcher uses recent year-specific units or the harmonized units to measure this MDG indicator. Malawi (Figure 5, Map A and B) presents a more variegated pattern of achievement. Much of the progress was concentrated in the northern districts, which helped drive up the national figures. The far south was largely stagnant.

Figure 4. Female non-agricultural wage employment, Mali Regions 1998-2009

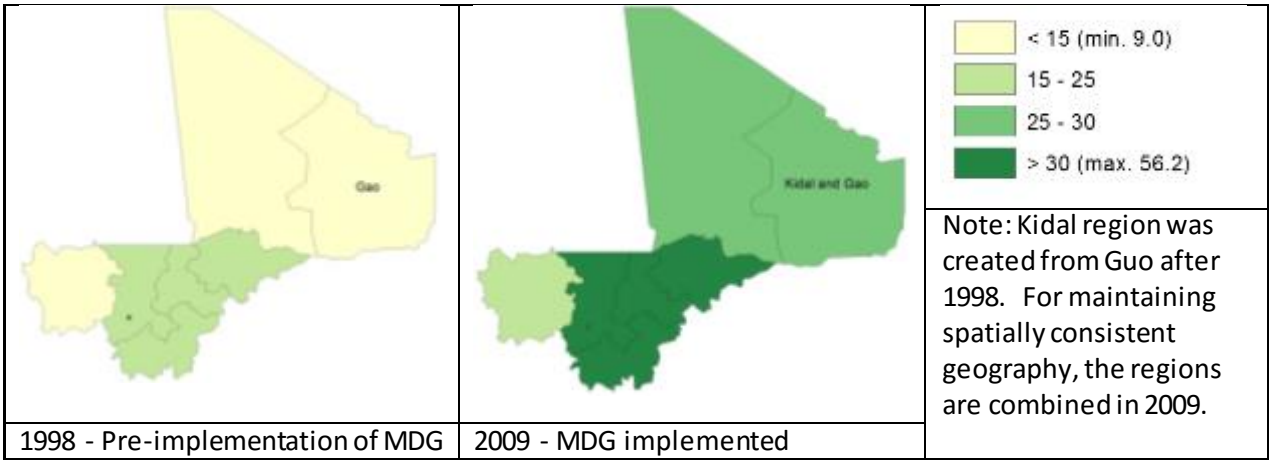
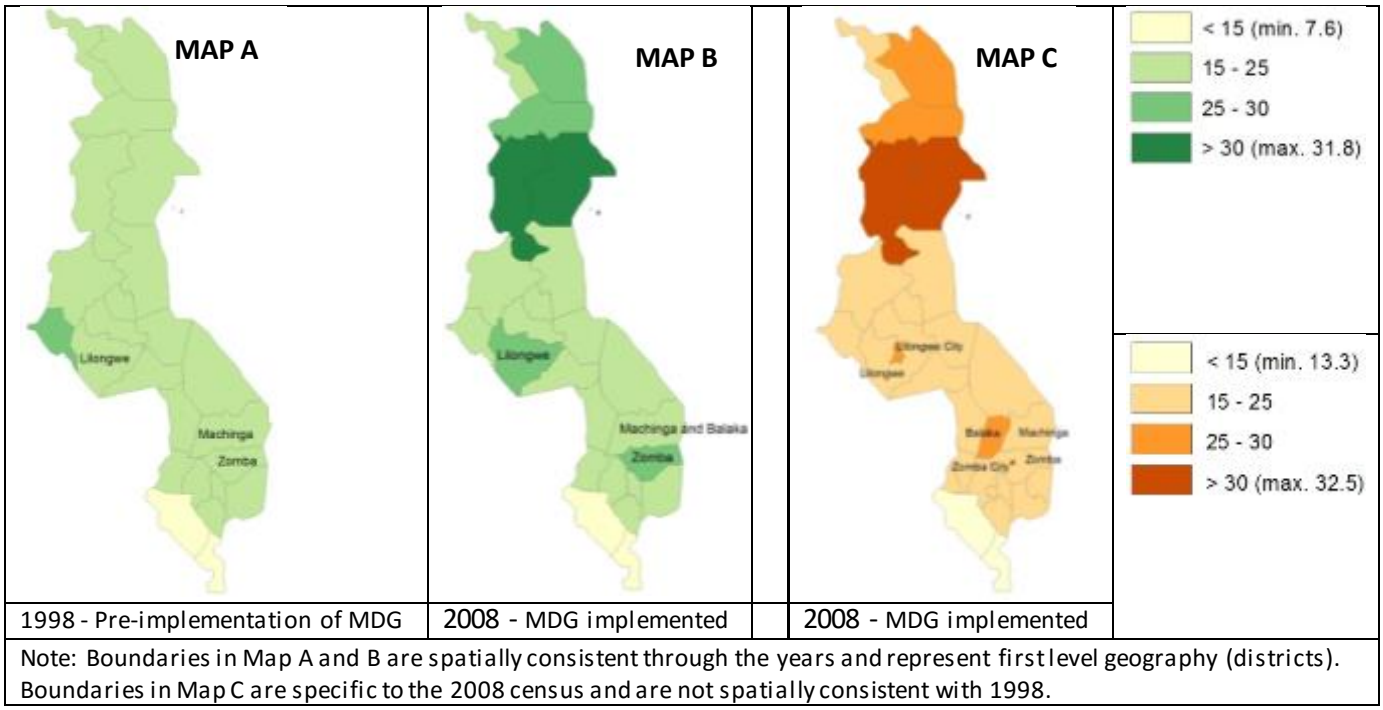
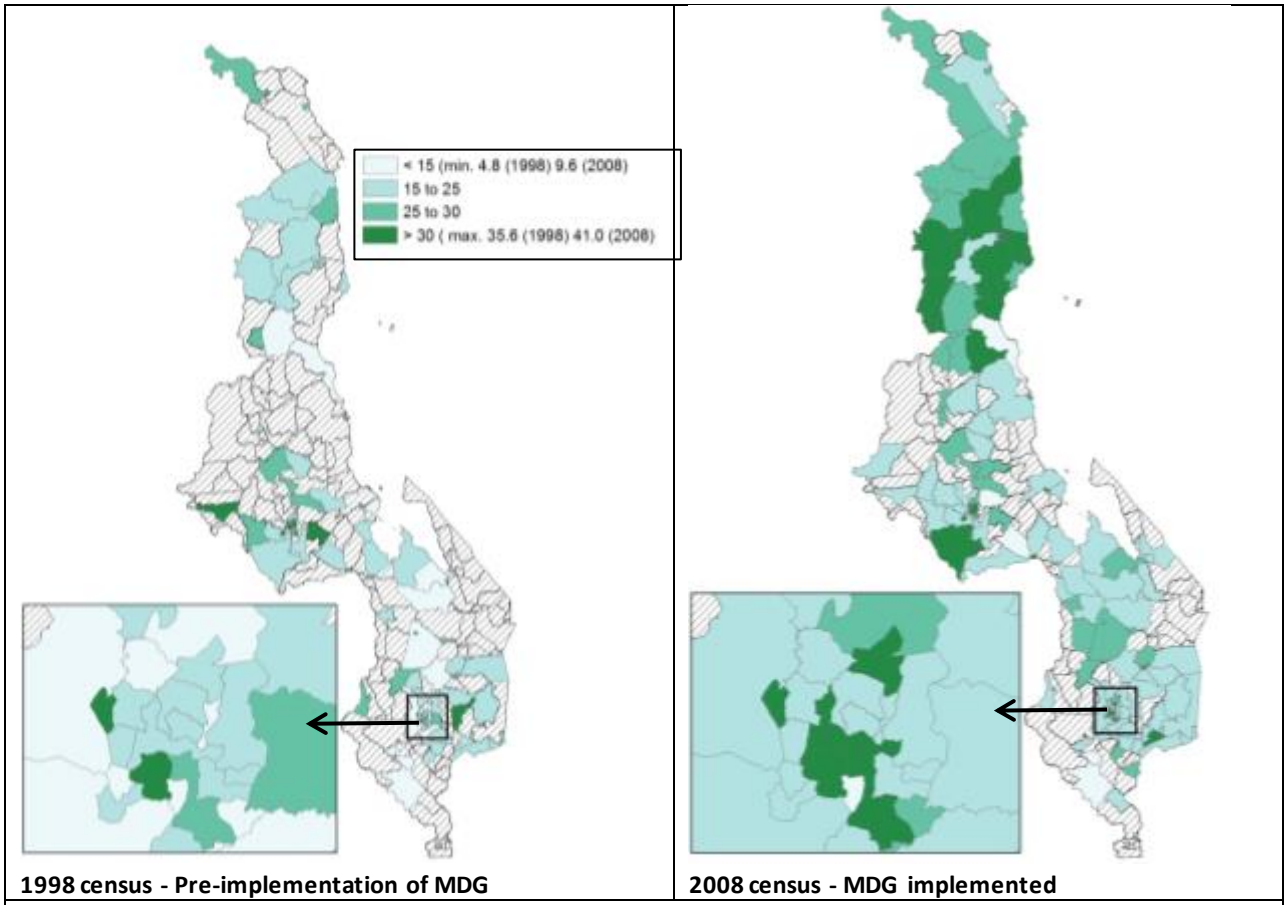


Figure 5. Female non-agricultural wage employment, Malawi Districts 1998-2008



Loss of detail in harmonized units: For the spatial visualization discussed above, we used the ISAs to hold boundaries constant over time. While that enables an apples-to-apples temporal comparison of places, the nature of the ISAs is to merge census units to encapsulate any boundary changes that occurred between censuses. In the process, some detail that might be useful for the analysis gets lost. Figure 5, Map C illustrates this point. In it we map original census units from 2008 Malawi districts. Lilongwe city, Balaka, and Zomba city are new districts in 2008, not observable in the harmonized spatially consistent 1998 and 2008 maps (Figure 5, Maps A and B). All the three units have greater female wage employment rates than their surrounding areas. Figure 5, Map C demonstrates that much of the apparent progress in their regions was more localized in urban places than in the whole area of Lilongwe or Zomba. Year-specific geography provides greater detail and should be used in conjunction with spatially harmonized maps where we hold boundaries constant over time.

Figure 6. Female non-agricultural wage employment, Malawi Traditional Areas 1998-2008



Note: The boundaries of the Traditional Areas are spatially consistent through the two census time periods. The inset map shows the urban area of Blantyre.
Note: The non-colored hatched TA boundaries represent very low (n<20) female non-agricultural wage earners in the sample data.

Size constraints: Figure 6 represents the percent share of female in non-agricultural wage employment in the Traditional Areas (TAs) of Malawi. TAs are the second-level geographic units in Malawi. Figure 6 employs the spatially consistent variant of them to enable direct comparison across censuses. At this scale one gets the benefit of harmonized geography without some of the cost described at the higher geographic level in Figure 5 above. The TAs shows regions that experienced little or no gain in the MDG indicator -- patterns that were not observable at the larger scale. The detailed image of the urban area of Blantyre shows distinctions at a near-neighborhood level, where population densities are sufficient to overcome confidentiality constraints. Even though Figure 5 shows limited progress in the Blantyre area (a district southwest of Zomba City), there is significant progress towards goal 3.2 in some of its constituent parts. The limitations of sample data are evident in Figure 6, however: cases are too sparse to calculate reliable non-agricultural statistics in many Traditional Areas.

CONCLUSION AND ONGOING WORK

Demographers and social scientists are increasingly incorporating spatial elements into their analyses. Until recently, geographic harmonization in census data available through IPUMS International did not account for changing spatial footprints of identified census units. Consistent spatial geographic units are necessary for accurate measures of change over time involving contextual or spatial elements as the examples from Africa illustrate. From our analysis, we have shown that there are several constraints that relate to analysis of outcomes with respect to space and time. These constraints can be experienced by any researcher trying to use both space and time as control variables. While other researchers have tried to find solutions to these challenges, the methods used show no consistence in their approaches. We have demonstrated how IPUMS data collection has rigorously tackled this issue – i.e., through harmonization and regionalization of both spatial and non-spatial variables. Additionally, we have demonstrated the utility of using a combination of year-specific geographic data and harmonized data, rather than either of them, in order to increase accuracy in interpreting observed results. We acknowledge the limitations of harmonized, spatial, and non-spatial variables, especially if the process leads to limited number of units. Additionally, while we argue that the use of lower level sub-national units helps provide a more accurate picture of the outcome variable; this process becomes problematic when units have sparse populations. While we can resolve the problem of small number of units that result from harmonization, by giving year-specific units, we cannot resolve the problem of small number of lower level units that result from regionalization, because of confidentiality issues.

At this time, IPUMS is working on making the second-level geography available for as many countries as possible, releasing the first half in the summer of 2015 and most of the remaining units in the summer of 2016. The project is also developing a protocol of an International Research Data Enclave, a secure data access environment to which researchers can apply for access to confidential data. The application and security requirements would be higher for this environment but will provide access to full-count or higher precision samples and to more detail in variables such as geographic units or occupational classifications. In the long term (resources and raw materials permitting), we would like to continue the harmonization and regionalization work to further subdivide densely populated units to create a variable that divides the country into geographic units of similar population sizes, thereby create something a little bit more like a homogeneous zoning system of the population.

REFERENCES

Clark, W. A. V., and Karen L. Avery. 2010. "The Effects of Data Aggregation in Statistical Analysis." *Geographical Analysis* 8 (4): 428–38. doi:10.1111/j.1538-4632.1976.tb00549.x.

Food and Agriculture Organization, United Nations. 2006. "The Global Administrative Unit Layers (GAUL)." http://www.fao.org/ES/giews/english/shortnews/GAUL1.pdf.

Guo, Diansheng. 2008. "Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning (REDCAP)." *International Journal of Geographical Information Science* 22 (7): 801–23. doi:10.1080/13658810701674970.

Guo, Diansheng, and Hu Wang. 2011. "Automatic Region Building for Spatial Analysis : Automatic Region Building for Spatial Analysis." *Transactions in GIS* 15 (July): 29–45. doi:10.1111/j.1467-9671.2011.01269.x.

Haining, Robert Patrick. 2003. *Spatial Data Analysis*. Cambridge University Press Cambridge.

Kauko, Tom. 2004. "A Comparative Perspective on Urban Spatial Housing Market Structure: Some More Evidence of Local Sub-markets Based on a Neural Network Classification of Amsterdam." *Urban Studies* 41 (13): 2555–79. doi:10.1080/0042098042000294565.

Kugler, Tracy, David Van Riper, Steven Manson, David Haynes II, Joshua Donato, and Katherine Stinebaugh. 2015. "Terra Populus: Workflows for Integrating and Harmonizing Geospatial Population and Environmental Data." *Journal of Map and Geography Libraries*.

Law, Gwillim. 2015. "Administrative Divisions of Countries ('Statoids')." www.statoids.com.

MacEachren, Alan M. 2004. "Relationships in Space and Time." In *How Maps Work: Representation, Visualization, and Design*, Pbk. ed. New York: Guilford Press.

Martin, David. 2003. "Extending the Automated Zoning Procedure to Reconcile Incompatible Zoning Systems." *International Journal of Geographical Information Science* 17 (2): 181–96. doi:10.1080/713811750.

Minnesota Population Center. 2014. "Integrated Public Use Microdata Series, International: Version 6.3 [Machine-Readable Database]." International.ipums.org.

Openshaw, S. 1984. "Ecological Fallacies and the Analysis of Areal Census Data." *Environment and Planning A* 16 (1): 17–31. doi:10.1068/a160017.

Openshaw, S, and P. J. Taylor. 1979. "A Million or so Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem." In *Statistical Applications in the Spatial Sciences*, Wrigley, N, 127–44. London: Pion.

Painho, M. 2000. “Using Genetic Algorithms in Clustering Problems.” In . University of Greenwich, United Kingdom.

Robert Hijmna’s Laboratory. 2014. “GADM Database of Global Administraive Areas.” http://www.gadm.org/.

Ruggles, Steven, Miriam L. King, Deborah Levison, Robert McCaa, and Matthew Sobek. 2003. “IPUMS-International.” *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 36 (2): 60–65. doi:10.1080/01615440309601215.

Sobek, Matthew, Lara Cleveland, Sarah Flood, Patricia Kelly Hall, Miriam L. King, Steven Ruggles, and Matthew Schroeder. 2011. “Big Data: Large-Scale Historical Infrastructure from the Minnesota Population Center.” *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 44 (2): 61–68. doi:10.1080/01615440.2011.564572.

UN Geographic Information Working Group. 2014. “SALB: Second Level Administrative Boundaries.”

United Nation. 2000. “Millenium Development Goals and Beyond 2015.” http://www.un.org/millenniumgoals/.

U.S. State Department, Office of the Geographer. 2014. “Large Scale International Boundaries (LSIB).” https://hiu.state.gov/data/data.aspx.