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

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 themes of gender equality from the UN Sustainable Development Goals across different geographic levels available in the census microdata. We show how the use of harmonized geographic units facilitates comparative metrics. We also show how the availability of geographic shapefiles enhances data visualization and communication capabilities.*

Keywords: *geographic harmonization, spatial analysis, data integration, census, administrative boundaries, sustainable development*

†Correspondence should be directed to:
Lara Cleveland
University of Minnesota
50 Willey Hall
225 19th Ave S.
Minneapolis, MN 55455
e-mail: clevelan@umn.edu
phone: 612-624-5818

INTRODUCTION

Changes in administrative boundaries pose a major challenge for spatio-temporal population research. Researchers interested in change overtime need to hold space constant to study contextual orspatial 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 a 20,000 population threshold required for public dissemination of the data. The base units are then disaggregated to also create year-specific units and shapefiles that still meet the necessary population threshold requirement. We illustrate the utility of harmonized units by exploring progress toward UN Sustainable Development Goals (SDGs) in select countries. The analysis illustrates the utility of consistent units for analysis and the way that disaggregation of national trends into regional or local ones can highlight 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 [16, 17]. 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 viathe website and in downloadable files [11].

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 [12], the MAUP is composed of two separate but closely related problems [12,13]. 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 [9,14].

Area problem: The area problem is a question of scale and 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) [7]. 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 area, or 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 [12]. 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.

Aggregation: 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 [5]. 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. As of 2015, the project has added a number of integrated geographic areas as new 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.

METHODS

Given the rise in digital mapping capabilities and spatial analytical technologies, social science research

[1] Important exceptions such as UNSALB [18] and Statoids [8] exist. IPUMS use of these resources is mentioned in the Methods section.

increasingly calls for consideration of space [9]. 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 [3, 4] 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 geographic areas 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 [2], the United Nations Second Administrative Level Boundaries (UNSALB) [18], and the Global Administrative Areas (GADM) dataset [15]. 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 geo-referenced to modern digital boundary files [20] 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 overtime. 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 [10]. 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 [3] 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 a spatial 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 [6, 10, 14].

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. 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 for all subsequent years. Releasing the large portion of a combined area (one that meets the threshold) as a stand-alone unit, would reveal the smaller unit. If we combined a small 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. Our algorithms and procedures for releasing the integrated and year-specific units and maps in tandem prevents the identification of units that are smaller than the threshold.

We use the Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning (REDCAP) algorithm and accompanying software [2, 3]. 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 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. We focus on Sustainable Development Goal 8.6.1: participation in education, employment and training, for both males and females in Cambodia and Vietnam and on women's labor force participation in Malawi. We provide analysis of data from the 2000 and 2010 census rounds for each country. Integrated geographic variables for Malawi are available at the first and second sub-national levels of geography, enabling a clear illustration of the utility of a full range of integrated and year-specific geographic identifiers. 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 geographic units.

The United Nations 2030 Agenda for Sustainable Development proposes 17 goals and 169 targets that aim

to complete what the Millennium Development Goals did not. The goals and targets concentrate on the eradication of poverty, hunger and inequality; access to education and healthcare; gender equality; environmental sustainability; economic, social, and technological progress, and the establishment of new partnerships for the achievement of these goals. The proposed framework for monitoring the Sustainable Development Goals (SDGs) emphasizes the need for disaggregated statistics that measure progress among different demographic and social groups at various levels of sub-national geography. The Sustainable Development Solutions Network recommends spatial disaggregation and stratification by sex, gender, age, income, disability, ethnicity, indigenous status, economic activity, and migrant status for nearly half of the proposed monitoring indicators. While enhanced data collection will almost certainly be necessary to monitor several SDGs, high-precision census microdata samples, like those disseminated by IPUMS-International, represent useful data that are part of the existing statistical infrastructure of most developing countries. These data are highly representative of national populations, are collected at regular intervals, and include measures of the population characteristics required for SDG indicator disaggregation.

Example 1: Employment, education and training in Cambodia and Vietnam: In this example, we use IPUMS integrated microdata to measure SDG 8.6.1 as stated in the SDG indicator framework [19]: the proportion of youths aged 15-24 years not in employment, education or training. For analytical purposes, we favor PEET, 100-NEET, recasting the indicator in a positive form by striking "not" from the definition. Two IPUMS integrated variables are used in defining PEET:

> 1. SCHOOL – attending. The IPUMS integrated variable, SCHOOL, indicates "whether or not the person attended school at the time of the census or within some specified period of time prior to the census." The question is roughly comparable in all census samples, although some censuses limit the designation to attendance at schools with a normal program of study leading to a succession of grades or levels.
> https://international.ipums.org /international-action/variables/SCHOOL#description_section
>
> 2. EMPSTAT (Economic activity status) - cmployed: EMPSTAT indicates whether or not the respondent was part of the labor force -- working or seeking work -- over a specified period of time. Depending on the sample, EMPSTAT can also convey further information. As noted above, the first digit of EMPSTAT is fully comparable and classifies the population into three groups: employed, unemployed, and inactive. Comparability issues arise from the reference period. For most samples, the reference period is current at the time of the census or the last 7 days. Last month, last year or undefined time interval characterizes barely a dozen of the 277 samples currently in the database. Unpaid family laborers are counted as employed and are rarely identified.
> https://international.ipums.org/international-action/variables/EMPSTAT#description _section

By cross-classifying the IPUMS integrated variables SCHOOL and EMPSTAT we arrive at four categories for youth ages 15-24: attending school, working, both attending school and working, and none of the above.

Figures 1 and 2 examine trends in PEET for males and females at the first administrative level for Cambodia and Vietnam. Patterns are clearly observable spatially as well as temporarily. Cambodia and the southern area of Vietnam show great diversity in PEET scores for both males and females in the 2000-round censuses (Figures 1 and 1). The substantial improvement registered in the 2010-round census is notable. In the 2000-round, large swaths of Cambodia and several areas of southern Vietnam showed PEET scores below 70% for females and scores below 80% for males. Ten years later, levels were above 80 percent for females in most areas and above 90 percent for males. The availability of consistent spatial units are essential for this type of comparison over time.

To understand these trends, census microdata may be used to distinguish between schooling and work to determine the main element of change. Although such analysis is beyond the scope of this paper, it does point to the power of census microdata to illuminate intensity and direction of social progress at local as well as national levels. Disaggregation to lower geographic levels will be possible for more countries with the next IPUMS International data release, including for Cambodia and Vietnam. In general, urban and developed areas tend to show more improvement over time than remote areas, but in a few places these trends are changing with recent immigration patterns. Further work could be done to assess the levels for migrant and traditionally disadvantaged groups within countries.

Example 2, Women in non-agricultural wage employment: In a second illustration, female non-agricultural employment in Malawi, we show how further spatial disaggregation lends provides important information insight about local differences. To explore women's employment, we map sub-national change for Malawi, a country that experienced large increases in women's employment between the 2000 and 2010 census rounds. Malawi (Figure 3, Map A and B) presents a variegated pattern of women's employment. Much of the increase was concentrated in the northern districts, which helped drive up the national figures. The far south was largely stagnant.

Loss of detail in harmonized units: For the spatial visualization discussed above, we used the consistent geographic units which hold boundaries constant over time. While that enables an apples-to-apples temporal comparison of places, the nature of the integrated geographic variables is to merge census units to encapsulate any boundary changes that occurred between census years. In the process, some detail that might be useful for the analysis gets lost. Figure 3 Map C illustrates this point. In it we map original census units from 2008 Malawi districts. Lilongwe city, Balaka, and Zomba city were new districts identified separately in the 2008 census, not observable in the harmonized spatially consistent 1998 and 2008 maps (Figure 3, Maps A and B). All the three units have greater female wage employment rates than their surrounding areas. Figure 3 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.

Size constraints: Figure 4 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 4 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 3 above. The TAs shows regions that experienced little or no gain in wage employment for women -- 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 increase in women's wage employment in some of its constituent parts. The limitations of sample data are evident in Figure 4, 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 more accurate pictures of the outcome variables; 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

[1] 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.

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

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

[4] 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.

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

[6] Kauko, Tom. 2004. "A Comparative Perspectiveon 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.

[7] 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." *Journalof Map and Geography Libraries*.

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

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

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

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

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

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

[14] Painho, M. 2000. "Using Genetic Algorithms in Clustering Problems." In . University of Greenwich, United Kingdom.

[15] Robert Hijmna's Laboratory. 2014. "GADM Database of Global Administraive Areas." http://www.gadm.org/.

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

[17] 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 Journalof Quantitative and Interdisciplinary History* 44 (2): 61–68. doi:10.1080/01615440.2011.564572.

[18] UN Geographic Information Working Group. 2014. "SALB: Second Level Administrative Boundaries."

[19] United Nations Economic and Social Council. 2016. "Report of the Inter-Agency and Expert Group on Sustainable Development Goal Indicators." Forty-seventh session. (E/CN.3/2016/Rev.1*), Annex IV. February 19.

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

Figure 1. Percent of females in education, employment or training in Cambodia and Vietnam from the 2000 and 2010 round censuses

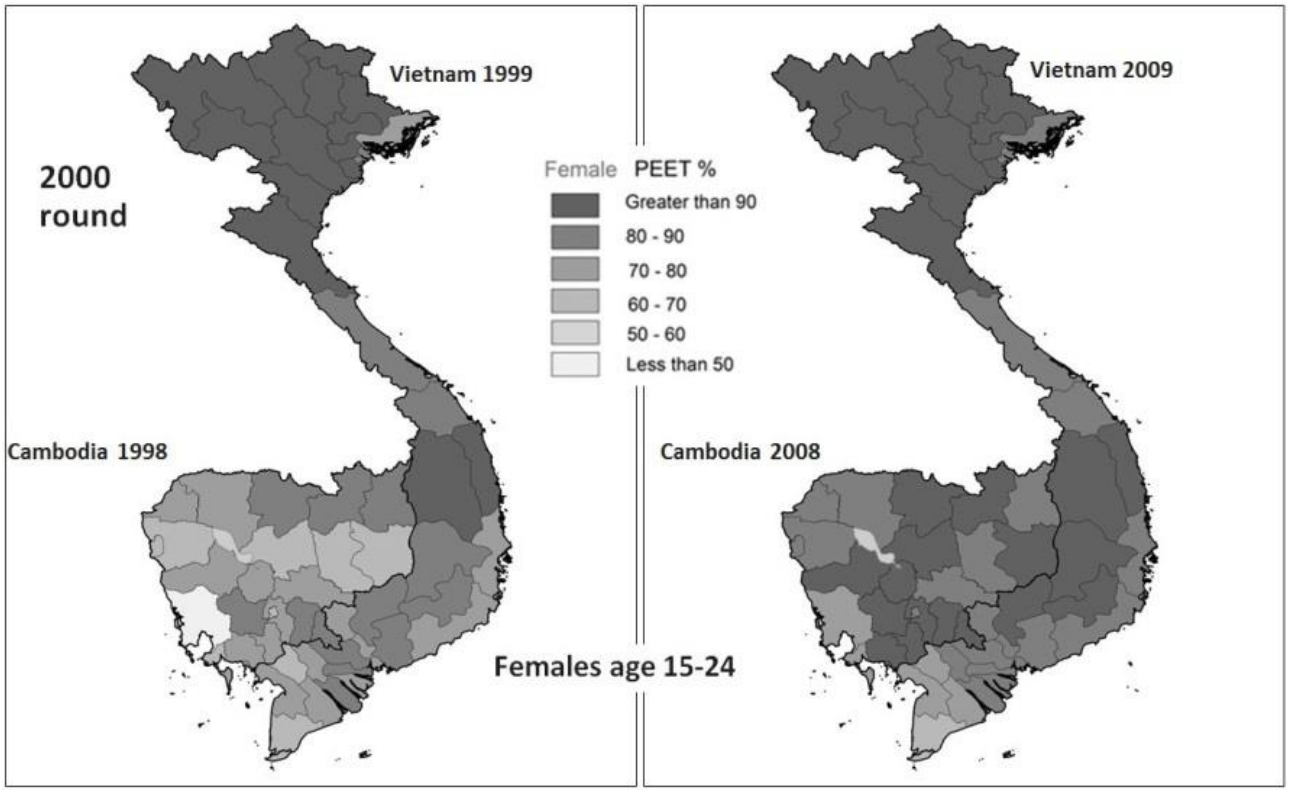


Figure 2. Percent of males in education, employment or training in Cambodia and Vietnam from the 2000 and 2010 round censuses

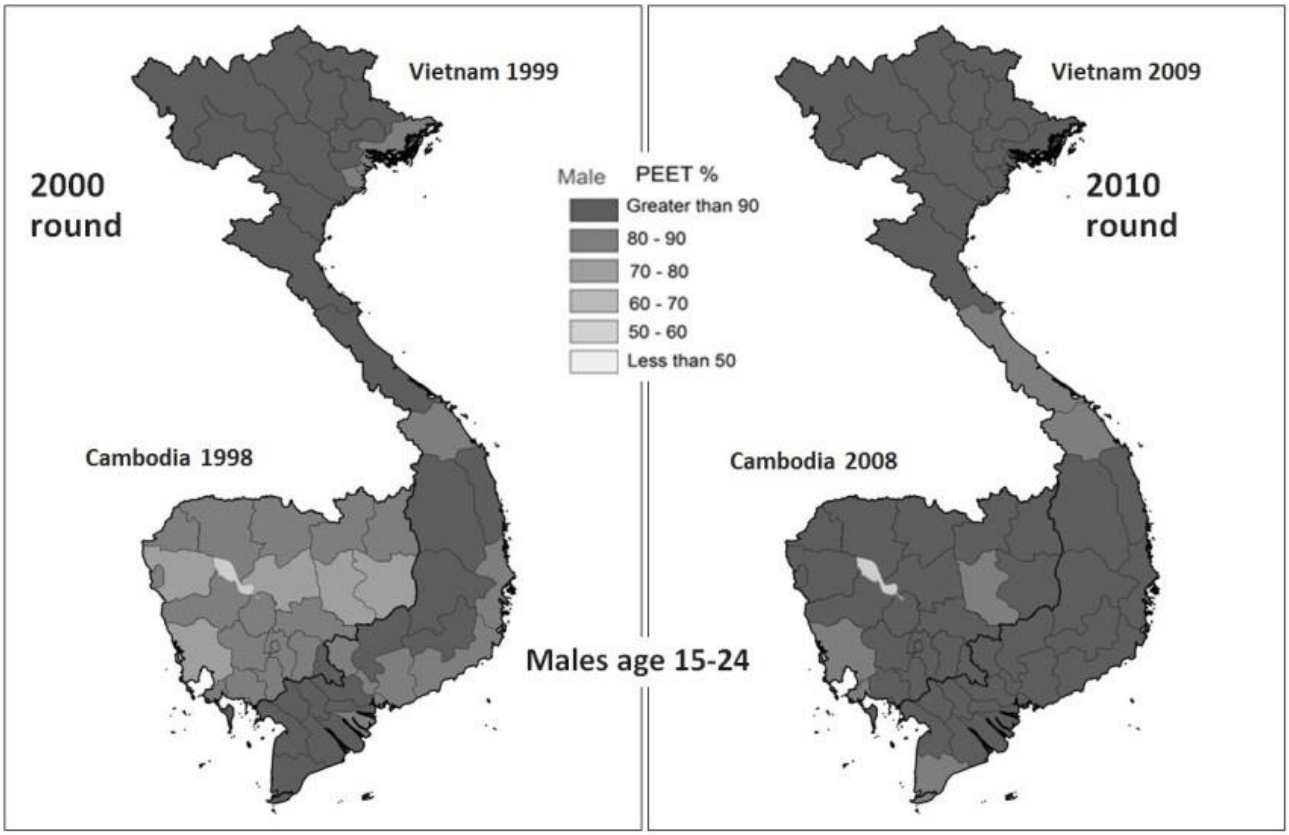
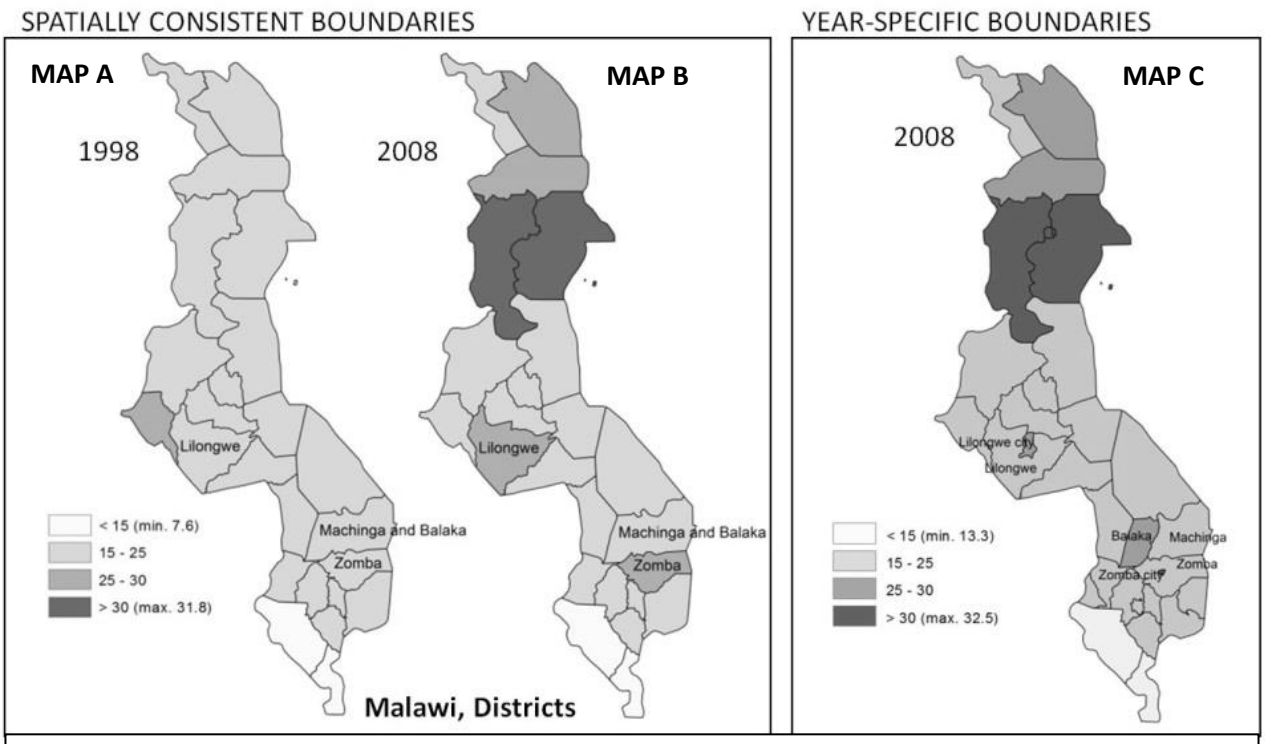
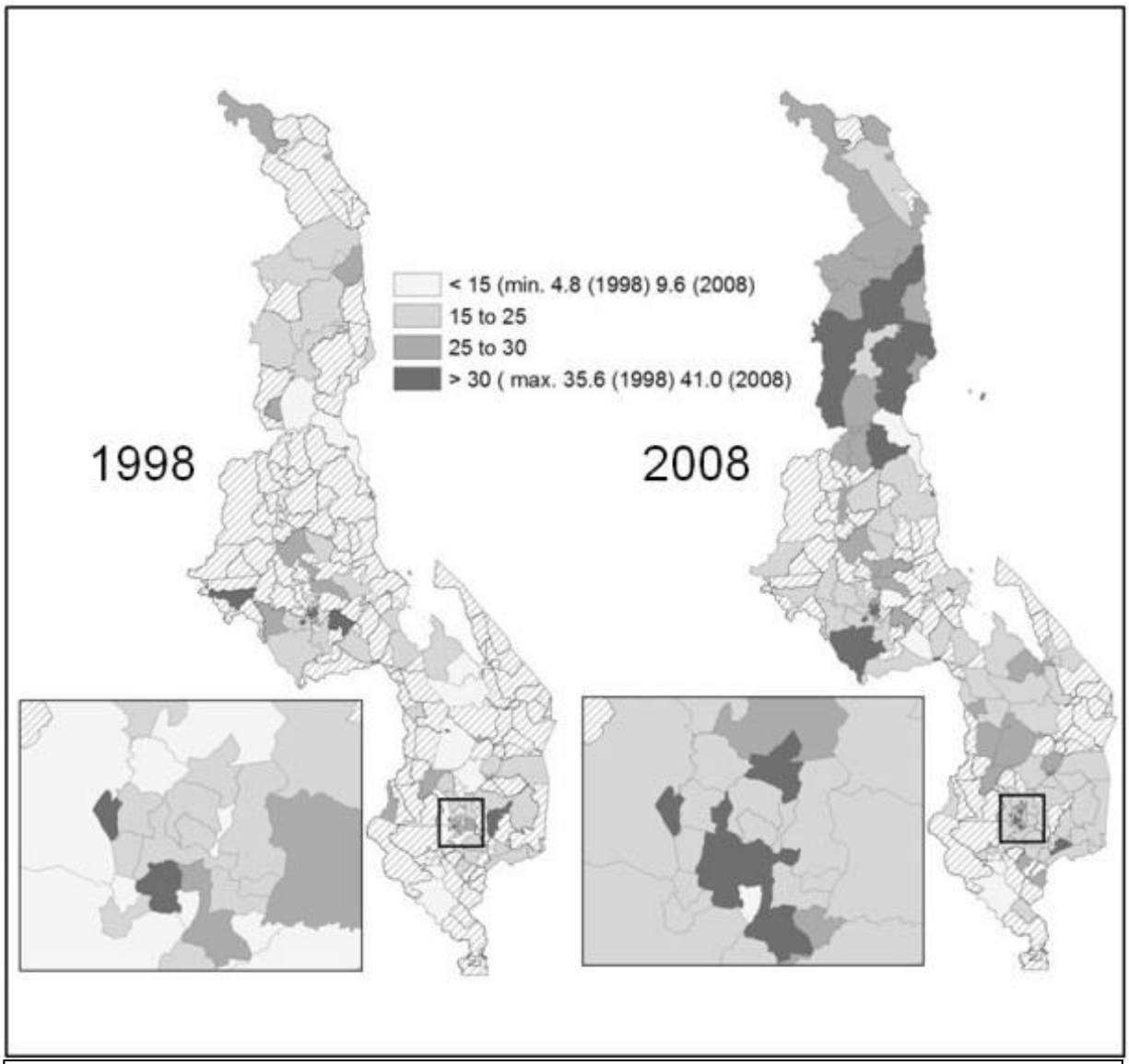


Figure 3. Female non-agricultural wage employment, Malawi Districts, 1998-2008



Note: Boundaries in Maps A and B are spatially consistent across years and represent first-level integrated geographic units (districts). Boundaries in Map C are specific to the 2008 census and contain additional detail.

Figure 4. Female non-agricultural wage employment, Malawi Traditional Areas, 1998-2008



Note: The boundaries of the Traditional areas are spatially consistent across census years. Inset map shows the urban area of Blantyre. Non-shaded hatched areas represent very low (n<20) female non-agricultural wage earners in the sample data and are therefore not shown for statistical purposes.