TUTORIAL 1: INTRODUCTION

Introduction to the Problem

Census public-use microdata possess an attribute richness which should make them tremendously useful to researchers interested in demographic small area estimation; however, they are underutilized, largely due to their coarse spatial resolution. Research which focuses on smaller geographic areas mostly relies on a limited number of aggregate population characteristics provided by the Census Bureau in summary tables and cross-tabulations at the census tract or block group level. In order to better exploit the attribute richness of Census microdata at finer spatial scales, spatial allocation methods, which allocate microdata households to small areas and generate summary statistics for these smaller geographic units using the attributes of the allocated microdata households, may be used. The goal of these methods is to provide summary statistics for population attributes, joint distributions of population attributes, or recategorizations of population attributes, which are not provided by the Census Bureau.

The following tutorials offer a step-by-step look at a method which probabilistically reweights and spatially allocates public-use microdata to small enumeration areas (such as Census tracts) based on the known aggregate populations of these areas. The known aggregate populations are identified from decennial Census and American Community Survey (ACS) data. Although the method is founded in the concepts of dasymetric mapping and areal interpolation, no prior knowledge of these concepts is assumed. In this approach, maximum entropy methods impute a set of tract-specific sampling weights for each microdata record, with the initial tract-specific weights derived from the original survey design weight. The imputed tract-specific sampling weights are constrained to match the known tract-level distributions for a given set of population attributes; the weights imputation is thus guided and influenced by this chosen set of attributes (called constraints or constraining variables). Sampling weights for each microdata household sum across all tracts to the original design (or household) weight provided by the Census Bureau. Each imputed sampling weight can thus be interpreted as the number of households of this type that can be expected in the respective tract. The aggregate of these weights within a tract, over all microdata records exhibiting a given population attribute, reflects the revised tract-level estimate of that attribute.

This module consists of five tutorials, each of which presents a detailed discussion of the practical application of a stage in the method. These tutorials are designed to be usable by individuals with no prior knowledge of small area estimation methods; to this end, technical details are largely absent. For users interested in a more technical treatment of these concepts, a current bibliography of published and working papers is provided. Throughout the tutorials, we will illustrate the concepts and computations using two stylized scenarios. These scenarios are described in the next section.

Scenarios

Scenario 1

Existing ACS summary files detail the number of households that receive SNAP benefits within each neighborhood, but this information is not broken down by household size. We are interested in knowing how usage of the Supplemental Nutrition Assistance Program (SNAP) varies by household size, to get a better sense of the reach of the program. In particular, we would like to know how many of the households that receive SNAP benefits are single person households and how many are large households, where large is defined as households with more than five individuals. We will use the maximum entropy weights imputation and allocation to construct tract-level estimates of these populations for the Corvallis, Oregon metropolitan area, based on data from the 2007-2011 ACS.

Scenario 2

We are interested in knowing how usage of SNAP differs between householders who own their own homes and householders who rent, and whether this difference varies by neighborhood. Although ACS summary tables enumerating the number of households that receive SNAP and the number of households that are owner- and renter-occupied are available at the tract-level, there is no joint enumeration of these household attributes. We will use the maximum entropy weights imputation and allocation to create this joint (or cross-tabbed) distribution for a portion of Boulder County, Colorado, based on data from the 2007-2011 ACS.

Brief Introduction to Census Geography

Summary (aggregate) population data from the decennial Census and the ACS are tabulated for a number of different geographies and administrative units, including states, counties, and Census tracts. The base unit for all Census summary tabulations is the Census block, a small area defined by visible features, such as roads, waterways, or rail lines. Census blocks are nested within slightly larger areas known as block groups, and block groups are nested within Census tracts. A Census tract is a semi-permanent statistical division which typically has a population size of 1,200 to 8,000 people, and which is commonly used as a proxy for a neighborhood by social science researchers.

Prior to 2010, data from the decennial Census short form (SF1 or SF2) were reported down to the block level and data from the long form (SF3 or SF4) were reported down to the tract level. The 2010 Census includes only the short form, still reported down to the block level, with the long form replaced by the ACS. In the ACS, the smallest reporting unit is the block group, although only data down to the tract level is available through the American FactFinder website.

The smallest geography reported in public-use microdata from either the decennial Census or the ACS is the Public Use Microdata Area (PUMA). PUMAs have a minimum population size of 100,000 individuals, and are always nested within states. Although PUMAs are built around counties, they often cross county boundaries, as a single county may not meet the population threshold required to have its own PUMA. Densely populated counties may

include many PUMAs; for example, in 2010 Los Angeles County included all or part of 69 different PUMAs. In nearly all cases, Census tracts (and thus block groups and blocks) are nested within PUMAs.

The maximum entropy weights imputation and spatial allocation method used here requires a nested data structure, such as that described above, with smaller geographies with known summary statistics nested within a larger geography from which microdata are drawn. The microdata are then allocated to whichever smaller geography is used. The examples that follow employ summary statistics from, and allocate to, Census tracts, although block groups could be used in place of the tracts. The use of tracts in the following examples is founded in the belief that this is the level of geography most used by researchers. We have not yet assessed differences in allocation performance between a model which uses tracts and a model which uses block groups. In addition, although the allocation could theoretically be performed on Census blocks, we have neither done so nor plan to do so.

For interested users, a more thorough and detailed description of Census geography and geographical concepts can be found at the Census Bureau website: *http://www.census.gov/geo/reference*.

Software

The initial data preprocessing can be accomplished using any statistical or spreadsheet software, including Excel, Stata, or R. The examples given in these tutorials will show sample data files for each of these three programs.

The maximum entropy imputation script runs in R, a free open source software environment for statistical computing. R can be downloaded from *http://www.cran.r-project.org*. RStudio is a GUI (graphical user interface) that runs on top of the R software, and is very helpful in editing scripts and working with datasets inside R. RStudio can be downloaded for no cost from *http://www.rstudio.com*. The *foreign* package in R will read and write datasets from many other statistical programs, including Stata.

The spatial allocation script is currently written in Stata, which is available through most universities or which may be purchased from *http://www.stata.com*. Note that the small version of Stata (Small Stata) is too restricted to handle the amount of data required for the spatial allocation. In the future, the spatial allocation script will be ported to R.

Mapping or GIS software may be used to explore the spatial patterns in the data prior to running the imputation or following the spatial allocation. Commonly used GIS software includes ArcGIS, available for purchase from *http://www.esri.com* (and also available at many universities) and QGIS, a free open source product available from *http://www.qgis.org*.