TUTORIAL 2: DATA DOWNLOAD AND PREPARATION

Determination of Study Area and Required Geographies

Upon selection of a study area, the first step is to determine the PUMAs and the tracts that are a part of this study area. Depending on the size of the study area, it may include many different PUMAs and a correspondingly large number of tracts. One way to determine the PUMAs and tracts that are required for the study area is to obtain boundary files for each of the different geographies, overlay them in a GIS, and select all of the PUMAs and tracts that intersect the study area. Boundary files are available from the National Historical Geographic Information System, run by the Minnesota Population Center (*http://www.nhgis.org*). Another method of accomplishing this same task is to use the Geographic Correspondence Engine available from the Missouri Census Data Center. This application takes a source geography (in this case the study area) and determines the relationship between the source geography and one or more target geographies (in this case PUMAs or tracts). The Geographic Correspondence Engine (called MABLE) works with 2000 or 2010 Census geographies, and is available at the following websites:

2000: *http://mcdc2.missouri.edu/websas/geocorr2k.html*
2010: *http://mcdc.missouri.edu/websas/geocorr12.html*

Note: Due to differences in the implementation dates of updated geographies into Census products, there may be temporal differences in the geographies used by different data sources. For example, tract-level summary tables from the 2006-2010 ACS and 2007-2011 ACS rely on tract definitions from 2010, while public-use microdata from these same two surveys are based on PUMA boundaries from 2000. Fortunately, MABLE will show a geographical correspondence table for 2010 tracts using 2000 PUMA boundaries.

Scenario 1

In this example, our study area is the Corvallis, OR metropolitan area, which consists of Benton County, Oregon. The task is to determine which PUMA (or PUMAs) includes Benton County, and which tracts are in that PUMA (or PUMAs). Going to the 2010 MABLE website, we select the state of Oregon, choose "County" as the source geocode, and choose "PUMA (2000)" as the target geocode. We also check the box for "Generate 2nd allocation factor". After running this request, a sample of the output is shown in data file **scenario 1 mable puma.xlsx**.

Each line in this file is a county/PUMA combination showing how much of the county is in the PUMA ("afact") and how much of the PUMA is encompassed by the county ("afact2"). We can see that Benton County lies almost solely within a single PUMA, 600. There is a tiny part of Benton County that lies within PUMA 1200, but this population (4) is negligible and we will ignore it for now. We can also see that Benton County contains only 42% of PUMA 600;

the rest of the PUMA covers neighboring Linn County. Although this fact is not relevant for the procedure that follows, we will return to it when evaluating our results.

Having determined that the weights imputation and spatial allocation will be based on PUMA 600 in Oregon, we must now identify those tracts which fall within this PUMA. Returning to MABLE, we again select the state of Oregon, choose "PUMA (2000)" as the source geocode, and choose "Census Tract" as the target geocode. We again check the box for "Generate 2nd allocation factor". After running this request, a sample of the output is shown in data file **scenario 1 mable tract.xlsx**.

Each line in this file is a PUMA/tract combination showing how much of the PUMA is in a tract ("afact") and how much of the tract is encompassed by the PUMA ("afact2"). We can see that there are 39 tracts that lie completely within PUMA 600, approximately half within Benton County and the other half within Linn County. Tract 102 in Benton County has a very small portion that lies outside of PUMA 600 (this is the same small portion that was in PUMA 1200 above); we can ignore this rounding error. These 39 tracts fully encompass the set of tracts which we will use in the spatial allocation.

Although we do not show it here, we could also use MABLE to determine the county/tract relationships. We would see that these 39 tracts are ALL of the tracts within Benton and Linn Counties. This point is useful in the data acquisition step.

To summarize, the weights imputation and spatial allocation for this Scenario requires 2007-2011 ACS microdata from PUMA 600 in Oregon and requires summary data for each of the 39 identified tracts, which includes all of the tracts in Benton and Linn Counties.

Scenario 2

In this example, our study area is Boulder County, CO. The task is to determine which PUMA (or PUMAs) includes Boulder County, and which tracts are in that PUMA (or PUMAs). Going to the 2010 MABLE website, we select the state of Colorado, choose "County" as the source geocode, and choose "PUMA (2000)" as the target geocode. We also check the box for "Generate 2nd allocation factor". After running this request, a sample of the output is shown in data file **scenario 2 mable puma.xlsx**.

Each line in this file is a county/PUMA combination showing how much of the county is in the PUMA ("afact") and how much of the PUMA is encompassed by the county ("afact2"). We can see that Boulder County lies within four different PUMAs (801-804). If we want to spatially allocate within the whole county, each of these PUMAs will have to be done separately. However, in this Scenario we are mostly interested in the city of Boulder, which we know from other sources (e.g. mapping software) lies almost entirely within PUMA 803. We will therefore continue this Scenario focusing solely on PUMA 803.

Having determined that the weights imputation and spatial allocation will be based on PUMA 803 in Colorado, we must now identify those tracts which fall within this PUMA. Returning to MABLE, we again select the state of Colorado, choose "PUMA (2000)" as the source geocode, and choose "Census Tract" as the target geocode. We again check the box for

"Generate 2nd allocation factor". After running this request, a sample of the output is shown in data file **scenario 2 mable tract.xlsx**.

Each line in this file is a PUMA/tract combination showing how much of the PUMA is in a tract ("afact") and how much of the tract is encompassed by the PUMA ("afact2"). We can see that there are 26 tracts that lie completely within PUMA 803, and all of these tracts lie within Boulder County. These 26 tracts fully encompass the set of tracts which we will use in the spatial allocation.

To summarize, the weights imputation and spatial allocation for this Scenario requires 2007-2011 ACS microdata from PUMA 803 in Colorado and requires summary data for each of the 26 identified tracts, all of which are in Boulder County, CO.

Download Summary Files

Having identified the Census tracts that are required for the weights imputation and spatial allocation, the next step is the acquisition of the necessary tract-level data. These are the known tract-level aggregates discussed in the Introduction and Choosing Constraints sections above, and are used to constrain the weights imputation. In the next section we discuss possible strategies to use for selecting those variables that may be useful as constraints. It may be helpful to read this section first, prior to downloading the summary data.

Although tract-level Census summary data may be obtained from any number of sources, two are discussed here: The American FactFinder website, managed by the Census Bureau (*http://factfinder2.census.gov*) and the National Historical Geographic Information System (NHGIS), managed by the Minnesota Population Center (*http://www.nhgis.org*). Both sites are free to use and offer Census summary data at a variety of geographic levels. Note that the data on the Census Bureau website only goes back to the 2000 Census; for 1990 and earlier data, NHGIS may be used. NHGIS requires free registration to download data.

Summary data should include variables defined at the household level (e.g. housing tenure or household income). The variables may need to be recategorized or collapsed to increase their utility. For example, we may believe that residential patterns are different for low income households and high income households, relative to middle income households, but the household income variable in the 2007-2011 ACS is divided into 16 categories. The estimates can be summed to create categories that are more aligned with our theoretical expectation. In other cases, the summary variables will already be categorized in a useful way. Such is the case with housing tenure, which is categorized by the Census into owner-occupied and renter-occupied. Clearly, the preparation of the summary data and its use in constraining the maximum entropy weights imputation is subject to some choices interpretation by the researcher.

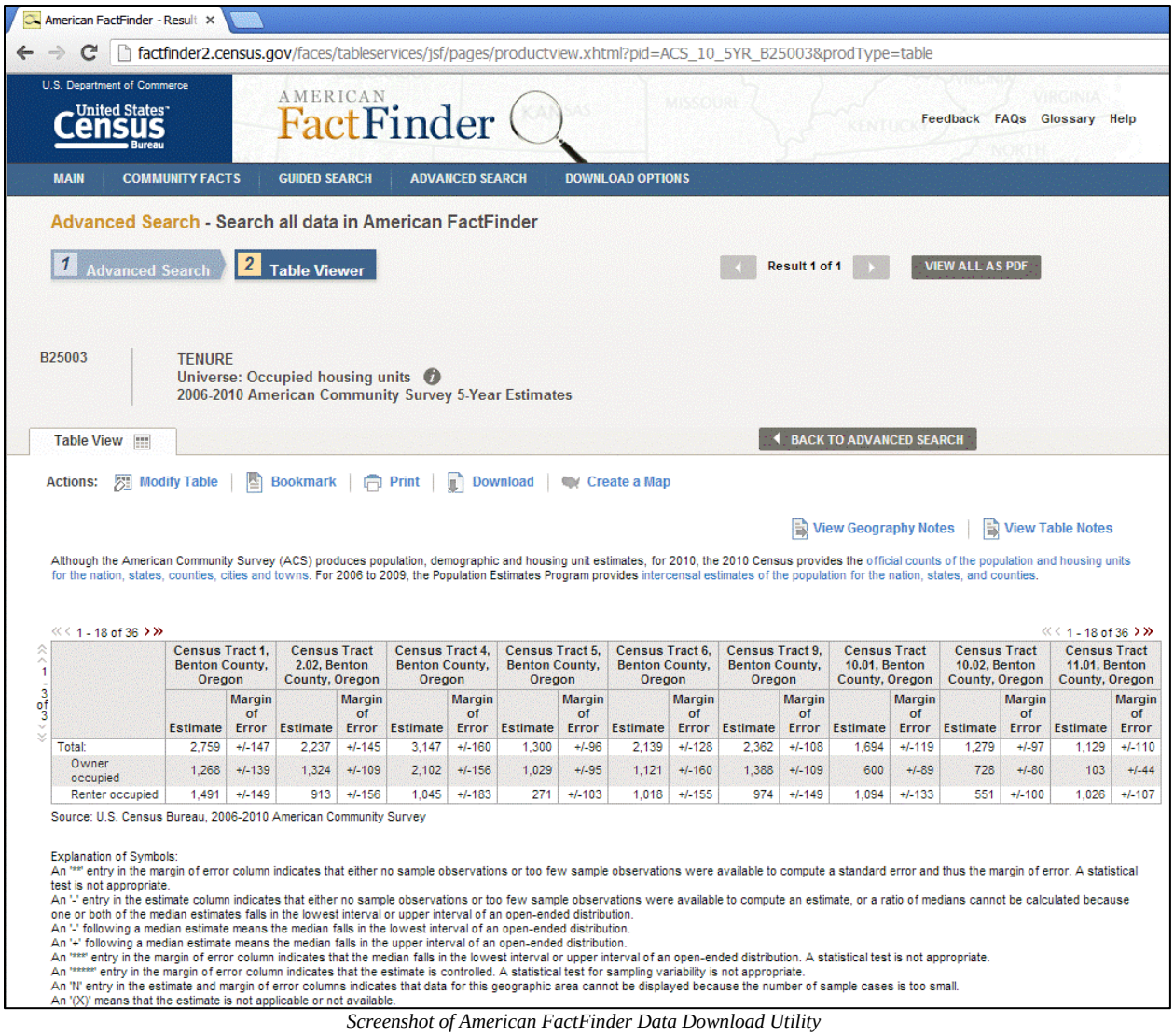
As noted prior, the summary data is used to constrain the maximum entropy imputation. The constraining variables can be used singly or in combination with one another. For example, we can use housing tenure and household income as single constraining variables or we could use the joint distribution of housing tenure and household income as a constraining variable (e.g. owner-occupied households with income less than $25,000, owner-occupied households with

income between $25,000 and $75,000, etc.). The strategy to use depends on the availability of the cross-tabbed variables in the summary files. In the Scenarios presented here, we use each constraining variable singly.

In summary data from the ACS, margins of error are included along with the estimates. In future work, these margins of error will be incorporated into the maximum entropy imputation. They are not used in the Scenarios presented here.

Scenario 1

To select constraining variables for the Corvallis, OR example, we downloaded 2007-2011 ACS summary data for the 39 tracts in the study area for several household variables that we believe may be useful in the weights imputation. These include housing tenure, race and ethnicity of the householder, age of the householder, household income, units in structure, vehicle availability, presence of a married-couple family, and presence of children. Specifically, we obtained tables B19037 (Age of Householder x Household Income), B25044 (Tenure x Vehicle Availability), B11005 (Household Type x Presence of Children), B25006 (Race of Householder), B25024 (Units in Structure), and B25003I (Tenure for Hispanic Householders). Data files from both American FactFinder and the NHGIS come with files describing the structure of the downloaded files. Additional information about the downloaded files from these sources can be found in the FAQs on their websites.



American FactFinder - Result

factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_10_5YR_B25003&prodType=table

U.S. Department of Commerce
United States Census Bureau

AMERICAN FactFinder

Feedback FAQs Glossary Help

MAIN | COMMUNITY FACTS | GUIDED SEARCH | ADVANCED SEARCH | DOWNLOAD OPTIONS

Advanced Search - Search all data in American FactFinder

1 Advanced Search 2 Table Viewer

Result 1 of 1 VIEW ALL AS PDF

B25003

TENURE
Universe: Occupied housing units
2006-2010 American Community Survey 5-Year Estimates

Table View BACK TO ADVANCED SEARCH

Actions: Modify Table | Bookmark | Print | Download | Create a Map

View Geography Notes | View Table Notes

Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, for 2010, the 2010 Census provides the official counts of the population and housing units for the nation, states, counties, cities and towns. For 2006 to 2009, the Population Estimates Program provides intercensal estimates of the population for the nation, states, and counties.

1 - 18 of 36

	Census Tract 1, Benton County, Oregon		Census Tract 2.02, Benton County, Oregon		Census Tract 4, Benton County, Oregon		Census Tract 5, Benton County, Oregon		Census Tract 6, Benton County, Oregon		Census Tract 9, Benton County, Oregon		Census Tract 10.01, Benton County, Oregon		Census Tract 10.02, Benton County, Oregon		Census Tract 11.01, Benton County, Oregon	
	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error
Total:	2,759	+/-147	2,237	+/-145	3,147	+/-160	1,300	+/-96	2,139	+/-128	2,362	+/-108	1,694	+/-119	1,279	+/-97	1,129	+/-110
Owner occupied	1,268	+/-139	1,324	+/-109	2,102	+/-156	1,029	+/-95	1,121	+/-160	1,388	+/-109	600	+/-89	728	+/-80	103	+/-44
Renter occupied	1,491	+/-149	913	+/-156	1,045	+/-183	271	+/-103	1,018	+/-155	974	+/-149	1,094	+/-133	551	+/-100	1,026	+/-107

Source: U.S. Census Bureau, 2006-2010 American Community Survey

Explanation of Symbols:
An '**' entry in the margin of error column indicates that either no sample observations or too few sample observations were available to compute a standard error and thus the margin of error. A statistical test is not appropriate.
An '-' entry in the estimate column indicates that either no sample observations or too few sample observations were available to compute an estimate, or a ratio of medians cannot be calculated because one or both of the median estimates falls in the lowest interval or upper interval of an open-ended distribution.
An '-' following a median estimate means the median falls in the lowest interval of an open-ended distribution.
An '+' following a median estimate means the median falls in the upper interval of an open-ended distribution.
An '***' entry in the margin of error column indicates that the median falls in the lowest interval or upper interval of an open-ended distribution. A statistical test is not appropriate.
An '*****' entry in the margin of error column indicates that the estimate is controlled. A statistical test for sampling variability is not appropriate.
An 'N' entry in the estimate and margin of error columns indicates that data for this geographic area cannot be displayed because the number of sample cases is too small.
An '(X)' means that the estimate is not applicable or not available.

Screenshot of American FactFinder Data Download Utility

The summary data were then reclassified based on theoretical expectations of the explanatory power of the variables:

Housing tenure was divided into owner occupied households and renter occupied households.
Vehicle availability was divided into households with a vehicle and households with no vehicles.
Householder race was divided into households with a black head of household and households with a non-black head of household.
Householder ethnicity was divided into households with a Hispanic head of household and households with a non-Hispanic head of household.

Presence of married family was divided into households with a married family and households without a married family.
Presence of children was divided into households with any children under the age of 18 and households without any children under the age of 18.
Units in structure was collapse into categories representing single family homes (1 unit), apartment buildings (2 or more units), or mobile units (mobile).
Household income was collapsed into categories representing low-income households (less than $25,000), medium-income households ($25,000-$75,000), and high-income households (more than $75,000).
Age of householder was collapsed into categories representing young households (age 24 or younger), middle-age households (age 25-64), and senior households (age 65 or above).

The final edited summary file for Corvallis can be accessed as data file **summary file corvallis.xlsx**. As you can see in this file, the total counts within each regrouping of household attributes (e.g. units in structure or householder race) sum to the total number of households in the tract. This file will first be used in the next section to assess which of these variables to use as constraints. It will then be used in the maximum entropy weights imputation in Tutorial 3.

Scenario 2

To select constraining variables for the Boulder, CO example, we downloaded 2007-2011 ACS summary data for the 26 tracts in the study area for the same variables as described in Scenario 1. The variables were reclassified in the same manner as in Scenario 1. The edited summary file for Boulder can be accessed as data file **summary file boulder.xlsx**.

Choosing Constraints

In order to successfully allocate households to Census tracts, the imputation procedure requires information about the households and information about the tracts. The constraining variables in the maximum entropy weights imputation should ideally delineate different household-level residential patterns; this will increase the variability in the underlying data that can be explained and result in more accurate estimates. Population characteristics (such as gender) that are similarly distributed among tracts are unlikely to produce satisfactory allocation results when used as constraints, since there may be little variation to exploit.

One way to select those variables that might be useful as constraints is to think about how the population is distributed in the area in which you are studying and what characteristics define this distribution. Are there neighborhoods in which certain groups are more likely to reside? For example, are there certain neighborhoods that contain large populations of college students? College students are likely to be young, to be renters, and to have low incomes. If the study area contains many neighborhoods with college students, household income, housing tenure, and age of householder might be good constraining variables to consider. Similarly, if the study area

includes neighborhoods that are divided by racial or ethnic characteristics, then householder race or householder Hispanic origin might be good constraining variables to consider.

A more formal way to select constraining variables might be the use of a segregation index, such as the index of dissimilarity, computed over the range of different household attributes available in the Census summary tables. The index of dissimilarity is a measure of the evenness of the distribution of two groups (Massey and Denton 1988), and may therefore be helpful in determining which variables best differentiate (or segregate) household residential patterns. The dissimilarity index is commonly interpreted as the proportion of individuals of one group who would have to move to reproduce within each neighborhood the distribution of the two groups that exists within the entire area. The formula for the dissimilarity index calculated for 2 mutually exhaustive population attributes, a_1 and a_2, is:

$$\frac{1}{2}\sum_{i=1}^{N}\left|\frac{a_{1i}}{A_1}-\frac{a_{2i}}{A_2}\right|$$

where N is the number of Census tracts in the study area, a_{1i} is the count of households exhibiting attribute a_1 in tract i, a_{2i} is the count of households exhibiting attribute a_2 in tract i, A_1 is the total count of households in the study area exhibiting attribute a_1, and A_2 is the total count of households in the study area exhibiting attribute a_2.

Dissimilarity index values range from 0 to 1, with values tending towards 1 indicative of more highly segregated groups and values tending towards 0 suggesting low levels of segregation among the groups. For the method described here, variables which exhibit (relatively) high values of the dissimilarity index might be useful as constraining variables.

In addition to the use of a segregation index, bivariate correlations of tract-level household attributes can be used to identify those variables that are highly correlated. The use of highly correlated variables as constraining variables may provide little benefit in the imputation procedure, as highly correlated variables will likely be redundant in explaining variation in the underlying population distribution.

A procedure to choose constraining variables might therefore commence with the downloading of summary data for a wide variety of household attributes for all of the tracts within a PUMA. Using the total number of households in a tract and the number of households exhibiting each attribute, the converse of each attribute could be created (e.g. black/not black, household income less than $25,000/household income greater than $25,000). Then the segregation of each attribute could be measured as the dissimilarity between the households that exhibit the attribute and the households that do not, using the formula above or using the *seg* command in Stata. After assessing which attributes are unevenly distributed (and thus possibly useful constraining variables), bivariate correlations between the attributes could be explored to remove redundant variables.

Another consideration in the selection of constraining variables is the size of the population exhibiting the attribute. Attributes that are very rare in the population may not perform well as constraining variables, as there may be a substantial number of tracts which have

0 counts of this attribute. As a result, this variable would contribute nothing in the weights imputation for these tracts.

An obvious question concerns the optimal number of constraining variables to use in the imputation, and this question is not yet answerable. Research on the consequences of additional constraining variables on model performance and allocation uncertainty is ongoing. Overfitting the maximum entropy model with too many constraints could result in non-convergence (no solution) to the problem, and outcome that is obviously best avoided. The Scenarios presented here use four (Scenario 1) and five (Scenario 2) constraining variables, and prior results suggest adequate model performance with these numbers of constraints. Of course, following the imputation and allocation, model performance can be evaluated and, in cases of poor model performance, the selection of constraining variables can be revisited.

Scenario 1

To create segregation indices for the variables downloaded in the previous section, each variable that was not already part of a binary grouping was reclassified into a binary grouping (e.g. householder less than age 25 vs. householder NOT less than age 25, householder age 25-64 vs. householder NOT age 25-64, etc.). The index of dissimilarity was calculated over each of these groupings to assess how well each variable defines residential patterns. These results are shown in data file **segregation indices.xlsx**. This file also displays the total number of households exhibiting each of the tested attributes.

The variables exhibiting the largest dissimilarity indices in Corvallis are the race of the householder, the units in structure, and the householder age. However, there are very few black households (approximately 1% of the total population), and this variable may therefore not serve well as a constraining variable. Housing tenure and vehicle availability show slightly lower segregation indices and are highly correlated (Spearman $\rho = 0.86$). We choose housing tenure as a constraining variable, as it contains a more even division of households between the two groups (owner and renter). Finally, although the total number of Hispanic households is also fairly low, this variable displays a low correlation with many of the other variables; we choose this as a fourth constraining variable.

To summarize, we have selected four variables to use as constraints in this Scenario: The number of units in structure, housing tenure, age of the householder, and Hispanic ethnicity of the householder.

Scenario 2

Each variable was again reclassified into binary groupings, and the index of dissimilarity was calculated over each of these groupings to assess how well each variable defines residential patterns. These results are shown in data file **segregation indices.xlsx**. This file also displays the total number of households exhibiting each of the tested attributes.

The variables exhibiting the largest dissimilarity indices in Boulder are those corresponding to the number of units in structure, housing tenure, and householder age. Similar to the Corvallis example, there are very few black households (approximately 1% of the total

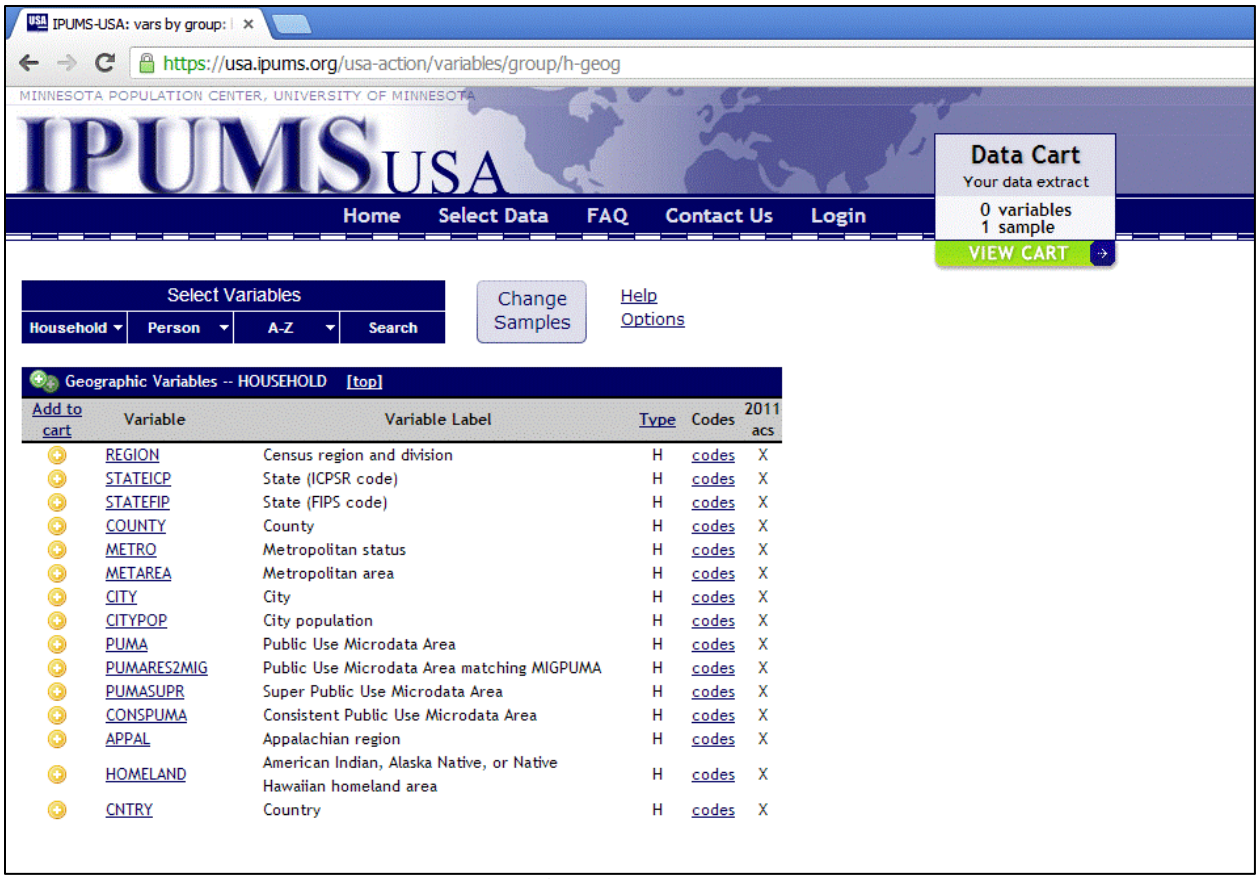
population), so although this variable exhibits a fairly high segregation index, it may not serve well as a constraining variable. Instead, we choose household income, which displays a high dissimilarity index for both low and high levels of income, and Hispanic householder as constraining variables. The choice of Hispanic householder over another variable (for example, married householder), is founded in the fact that Hispanic householder displays little correlation with the other selected constraining variables. Married householder is highly correlated with owner occupied household (Spearman $\rho = 0.95$).

To summarize, we have selected five variables to use as constraints in this Scenario: The number of units in structure, housing tenure, age of the householder, household income, and Hispanic ethnicity of the householder.

Download PUMS Files

Having prepared the summary data and selected the constraining variables, the next step is to download the individual-level microdata to be used in the maximum entropy imputation. This data comes from the public-use micro sample (PUMS) and is available from the IPUMS website, run by the Minnesota Population Center (*https://usa.ipums.org/usa/index.shtml*). This website requires registration (it uses the same username/password as the NHGIS), but is free to use. IPUMS data files are fixed width text files and come with programs to read the data into Stata or other statistical software. In downloading the PUMS data, two things should be kept in mind: We need to obtain variables that relate the microdata to the chosen constraint variables, and we need to obtain any other variables that we want to spatially allocate.

Although IPUMS data cannot presently be limited to certain PUMAs, the selection can be limited by state, using the Select Cases option on the Extract Request screen. This will help to minimize the file size, and the necessary PUMAs can be extracted later in the process.



Screenshot of IPUMS Data Download Utility

Scenario 1

From the IPUMS website, we retrieve the following variables from the 2011 ACS 5-year PUMS sample: SERIAL (household identifier), PERNUM (person number in household), HHWT (household design weight), GQ (group quarters status), STATEFIP (state of residence), PUMA (PUMA of residence), NUMPREC (household size), OWNERSHP (housing tenure), UNITSSTR (units in structure), AGE (age), and HISPAN (Hispanic origin). The raw PUMS files can be seen in data file **raw pums scenario 1 excel.zip** (zipped Excel) or **raw pums scenario 1 stata.zip** (zipped Stata).

Scenario 2

This example proceeds exactly as above, except that we retrieve two additional variables: HHINCOME (household income) and FOODSTMP (SNAP usage). The raw PUMS files can be seen in data file **raw pums scenario 2 excel.zip** (zipped Excel) or **raw pums scenario 2 stata.zip** (zipped Stata).

Data Preprocessing

Although the summary data are ready to be used in the form in which they are shown above, the PUMS data require some preprocessing. This can be done within the R script, or it can be done in Excel or Stata prior to importing the data into R.

First, the PUMS data should be limited to the relevant PUMA that comprises the study area. This is the PUMA that was identified earlier in this tutorial (e.g. PUMA 600 for Corvallis, PUMA 803 for Boulder). If multiple PUMAs are going to be considered, the PUMS data for these other PUMAs can be saved to different files and run subsequently.

PUMS records for individuals within group quarters should be removed. This can be accomplished through the elimination of any PUMS record with a GQ code of "3" or "4". Group quarters are considered distinct from households by the Census Bureau and are therefore not used in the weights imputation and spatial allocation.

The constraining variables selected earlier were based on the characteristics of the head of each household (the householder), and the PUMS data should thus be limited only to these individuals. This can be accomplished by retaining only those PUMS records with a PERNUM code of "1".

PUMS variables that are necessary to relate a PUMS record to the selected constraining variables should be recoded to match the categorizations of the constraining variables. Note that household attributes may be defined by different numbers of constraining variables. In our example, housing tenure is defined by two constraining variables (owner-occupied or renter-occupied), while household income is defined by three constraining variables (householder age 24 or less, householder age 25-64, or householder age 65 or greater). Thus the recoded PUMS variable that defines housing tenure will have 2 levels (0 and 1) and the recoded PUMS variable that defines householder age will have 3 levels (0, 1, and 2, or 1, 2, and 3). In the weights imputation, it is important that these codes are input so that they match the summary data inputs exactly. A table such as the one below, which corresponds to Scenario 1, will be helpful for the weights imputation in Tutorial 3. This table displays the recoded PUMS variable that was created (e.g. UNITS_RECODE), the codes contained in that variable (e.g. 1, 2, and 3), and the summary data variables that correspond to those codes (e.g. 1_UNIT_STRUCTURE, 2_OR_MORE_UNIT_STRUCTURE, and MOBILE_STRUCTURE).

		Recoded PUMS Variable (and Codes)	Corresponding Summary Variable
		UNITS_RECODE	
Units in Structure	1	1	1_UNIT_STRUCTURE
	2 or more	2	2_OR_MORE_UNIT_STRUCTURE
	Mobile	3	MOBILE_STRUCTURE
		TENURE_RECODE	
Housing Tenure	Renter-occupied	0	RENTER_OCCUPIED
	Owner-occupied	1	OWNER_OCCUPIED
		AGE_RECODE	
Age of Householder	24 or less	1	AGE_24_OR_LESS
	25-64	2	AGE_25_64
	65 or greater	3	AGE_65_OR_MORE
		HISPANIC_RECODE	
Hispanic Origin	Non-Hispanic	0	NON_HISPANIC_HOUSEHOLDER
	Hispanic	1	HISPANIC_HOUSEHOLDER

The PUMS data for any variable that is to be allocated should also be recoded in a similar way. These variables should be recoded as binary variables – this is how the allocation is currently performed. At a minimum, the recoded PUMS file should include the household weight, one variable for each constraint, and any variables that are going to be allocated.

Scenario 1

We begin by creating four recoded variables, UNIT_RECODE, TENURE_RECODE, AGE_RECODE, and HISPANIC_RECODE, which correspond to our four constraining variables. These variables are then coded as follows:

UNIT_RECODE: The original variable detailing units in structure (UNITSSTR) has 10 codes. Codes 1 and 2 correspond to mobile housing units; these are coded as 3 in UNIT_RECODE. Codes 3 and 4 in UNITSSTR correspond to single family housing units; these are coded as 1 in UNIT_RECODE. Codes 5-10 in UNITSSTR correspond to multi-family housing units; these are coded as 2 in UNIT_RECODE.

TENURE_RECODE: The original variable detailing housing tenure (OWNERSHP) has 2 codes. Code 1 corresponds to owner-occupied housing units; these are coded as 1 in TENURE_RECODE. Code 2 corresponds to renter-occupied housing units; these are coded as 0 in TENURE_RECODE.

AGE_RECODE: The original variable detailing householder age (AGE) is a continuous variable. AGE_RECODE is coded as 1 if AGE is 24 or lower, coded as 2 if AGE is between 25 and 64, and coded as 3 if AGE is 65 or higher.

HISPANIC_RECODE: The original variable detailing Hispanic origin (HISPAN) has 5 codes. Code 0 corresponds to householders who are non-Hispanic; these are coded as 0 in HISPANIC_RECODE. Codes 1-4 correspond to householders of different Hispanic origins; these are coded as 1 in HISPANIC_RECODE.

Next we want to create binary variables for the household attributes that we want to allocate. In this example, we are interested in knowing how many households that receive SNAP benefits are single person households and how many are large (5+ individuals) households. We create two new variables, SNAP_SINGLE and SNAP_LARGE. These variables are then coded as follows:

SNAP_SINGLE: The original variables detailing SNAP usage and household size are FOODSTMP and NUMPREC, respectively. FOODSTMP is coded as 1 if the household does not receive SNAP benefits and 2 if the household does receive SNAP benefits. NUMPREC is a continuous variable indicating the number of persons in the household. We thus code SNAP_SINGLE as 0 if FOODSTMP is 1, or if FOODSTMP is 2 and NUMPREC is greater than 1. SNAP_SINGLE is coded as 1 if FOODSTMP is 1 and NUMPREC is 1.

SNAP_LARGE: SNAP_LARGE is coded as 0 if FOODSTMP is 1, or if FOODSTMP is 2 and NUMPREC is less than 5. SNAP_LARGE is coded as 1 if FOODSTMP is 1 and NUMPREC is 5 or greater.

A screenshot of the PUMA dataset following recoding is shown below. The full dataset can be accessed as data file **pums scenario 1 recoded.xlsx** (Excel) or **pums scenario 1 recoded.dta** (Stata).

SERIAL	HHWT	UNITSSTR	OWNERSHP	AGE	HISPAN	NUMPREC	FOODSTMP	UNIT_RECODE	TENURE_RECODE	AGE_RECODE	HISPANIC_RECODE	SNAP_SINGLE	SNAP_LARGE
4768331	25	3	1	82	0	2	1	1	1	3	0	0	0
4768358	15	3	1	70	0	2	1	1	1	3	0	0	0
4770980	5	3	2	39	0	5	2	1	0	2	0	0	1
4768387	29	3	1	59	0	2	1	1	1	2	0	0	0
4768411	5	3	1	25	0	2	1	1	1	2	0	0	0
4768420	22	7	2	39	0	4	1	2	0	2	0	0	0
4768427	19	3	1	70	0	1	1	1	1	3	0	0	0
4768430	12	3	1	67	0	2	1	1	1	3	0	0	0
4768453	16	3	1	53	0	2	1	1	1	2	0	0	0
4768465	18	3	1	62	0	2	1	1	1	2	0	0	0
4768466	21	7	2	22	0	2	2	2	0	1	0	0	0
4768467	5	5	2	29	0	4	2	2	0	2	0	0	0
4768473	16	6	2	67	0	1	2	2	0	3	0	1	0
4768481	22	3	1	53	0	2	1	1	1	2	0	0	0
4772567	7	3	2	34	0	5	2	1	0	2	0	0	1
4768496	2	1	1	52	0	3	1	3	1	2	0	0	0
4768505	4	3	1	70	0	2	1	1	1	3	0	0	0
4768534	15	3	1	63	0	2	1	1	1	2	0	0	0
4768568	22	1	1	61	0	2	1	3	1	2	0	0	0

Screenshot of Recoded PUMS Data File for Scenario 1

Scenario 2

Scenario 2 proceeds similarly to Scenario 1, but we have the additional constraining variable of household income:

INCOME_RECODE: The original variable detailing household income (HHINCOME) is a continuous variable. INCOME_RECODE is coded as 1 if HHINCOME is less than 25,000, coded as 2 if HHINCOME is 25,000 or more and less than 75,000, and coded as 3 if HHINCOME is 75,000 or more.

Next we want to create binary variables for the household attributes that we want to allocate. In this example, we are interested in knowing how many households that receive SNAP benefits are owner-occupied households and how many are renter-occupied households. We create two new variables, SNAP_OWNER and SNAP_RENTER. These variables are then coded as follows:

SNAP_OWNER: The original variable detailing SNAP usage is FOODSTMP. FOODSTMP is coded as 1 if the household does not receive SNAP benefits and 2 if the household does receive SNAP benefits. We thus code SNAP_OWNER as 0 if FOODSTMP is 1, or if FOODSTMP is 2 and TENURE_RECODE is 0. SNAP_OWNER is coded as 1 if FOODSTMP is 1 and TENURE_RECODE is 1.

SNAP_RENTER: SNAP_RENTER is coded as 0 if FOODSTMP is 1, or if FOODSTMP is 2 and TENURE_RECODE is 1. SNAP_RENTER is coded as 1 if FOODSTMP is 1 and TENURE_RECODE is 0.

A screenshot of the PUMA dataset following recoding is shown below. The full dataset can be accessed as data file **pums scenario 2 recoded.xlsx** (Excel) or **pums scenario 2 recoded.dta** (Stata).

SERIAL	HHWT	UNITSSTR	OWNERSHP	AGE	HISPAN	HHINCOME	FOODSTMP	UNIT_RECODE	TENURE_RECODE	AGE_RECODE	HISPANIC_RECODE	INCOME_RECODE	SNAP_OWNER	SNAP_RENTER
1000070	13	3	1	52	0	276783	1	1	1	2	0	3	0	0
1000078	17	3	1	36	0	69532	1	1	1	2	0	2	0	0
1000192	47	8	2	23	0	9830	2	2	0	1	0	1	0	1
1000271	21	3	1	57	0	86398	1	1	1	2	0	3	0	0
1000319	11	7	2	30	0	7346	1	2	0	2	0	1	0	0
1000378	14	3	1	83	0	43044	1	1	1	3	0	2	0	0
1000434	23	3	2	29	0	61358	1	1	0	2	0	2	0	0
1000647	10	6	2	35	1	37456	1	2	0	2	1	2	0	0
1000705	22	3	1	57	0	58978	1	1	1	2	0	2	0	0
1000724	22	4	1	44	0	5898	1	1	1	2	0	1	0	0
1001841	22	8	2	38	0	2483	2	2	0	2	0	1	0	1
1000990	12	3	1	72	0	36008	1	1	1	3	0	2	0	0
1001028	26	10	2	84	0	50700	1	2	0	3	0	2	0	0
1001042	13	3	1	71	0	158309	1	1	1	3	0	3	0	0
1001044	15	3	1	59	0	39319	1	1	1	2	0	2	0	0
1001057	13	9	1	26	0	75744	1	2	1	2	0	3	0	0
1001060	15	3	2	44	0	25868	1	1	0	2	0	2	0	0
1001153	17	3	2	20	0	414	1	1	0	1	0	1	0	0
1001167	17	3	1	53	0	254464	1	1	1	2	0	3	0	0

Screenshot of Recoded PUMS Data File for Scenario 2