Digital Populations User Tutorial

What is Digital Populations?

Digital Populations (or "DP") is a software suite for synthesizing a plausible population census. DP starts with statistics and partial data from an actual census, and generates a complete census that resembles the input data. The generated census contains locations and statistics for every household and individual even though the input data contains no location data and only partial statistics.

The primary purpose of DP is to facilitate the detection of statistical clusters. Cluster detection requires a complete population data set, however most governments don't release that data for privacy reasons. DP can generate configurations of households that can be used for cluster detection.

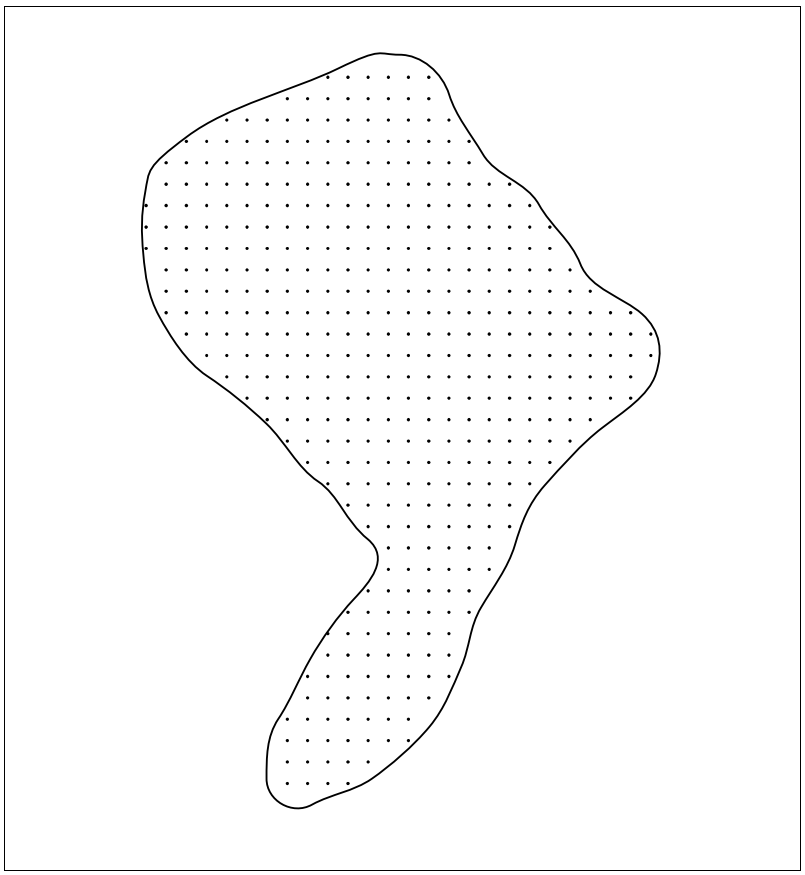
Required Input Data

Digital Populations requires input files that describe some portion of an actual census, along with control files that describe the files and specify which statistics are important to the analysis. The files include:

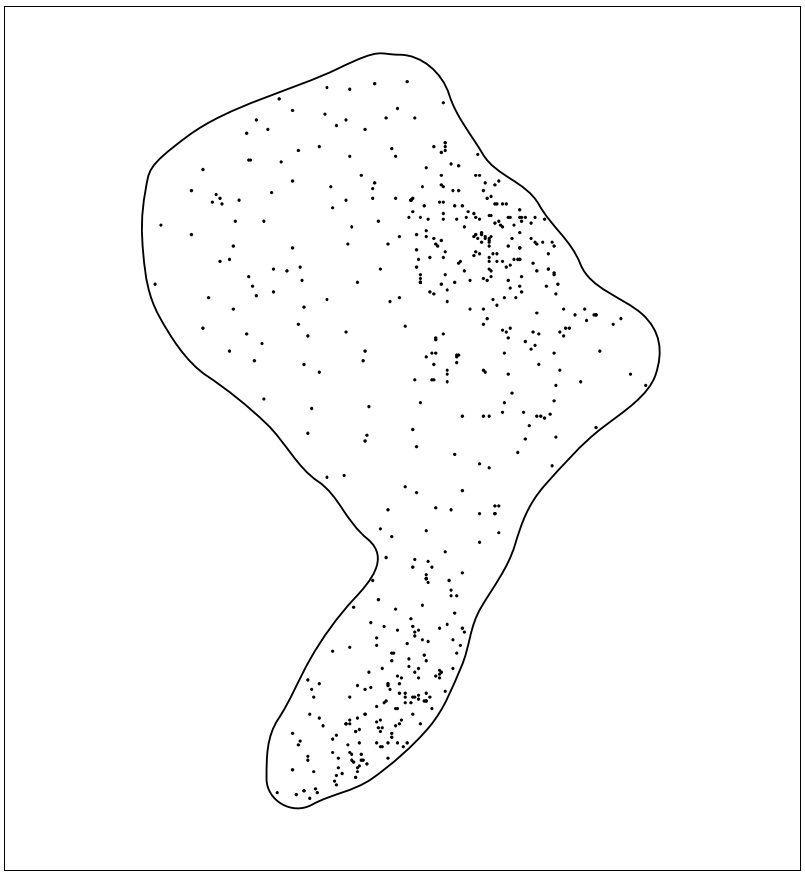
- Tract statistics. Censuses are generally collected by dividing a section of landscape into tracts or regions, then collecting statistics a region at a time. DP uses this data to guide the generation and placement of households and individuals.
- Partial data sample (aka "public use microdata sample," or PUMS). While governments generally refuse to release the full census data, they will usually release an anonymized subset of the data. DP clones the households and individuals in this data as needed to produce the full population for the area being analyzed.
- Relationship file. A control file is required to describe the input files and their structure. This file also contains a listing of "traits" which name an interesting characteristic of a households or individual, and specifies how to calculate the characteristic from the region data as well as from the PUMS data. By comparing the two calculations, DP can determine the quality of its generated census.

Sample Walk Through

The ultimate goal of Digital Populations is to distribute households across a map in way that resembles an actual census. Since DP doesn't have complete census data to work with, it starts with random locations, then nudges households into positions that agree with statistical data given as input. This is called a "plausible" arrangement – it's totally fake, but resembles an actual census.



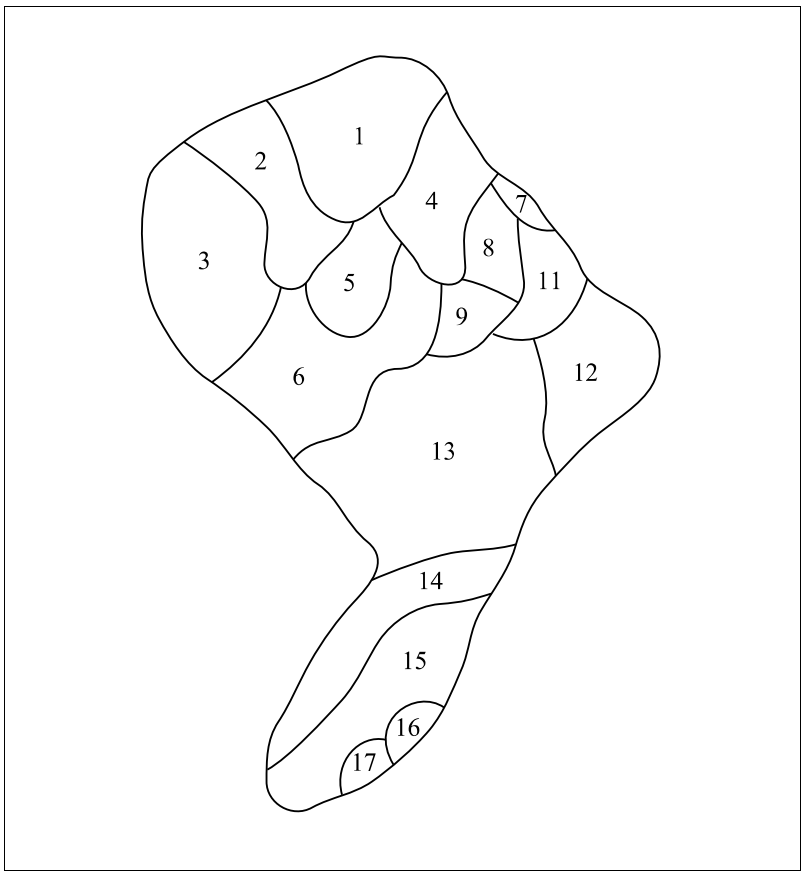
Map with poor placement of households.



Map with plausible placement of households.

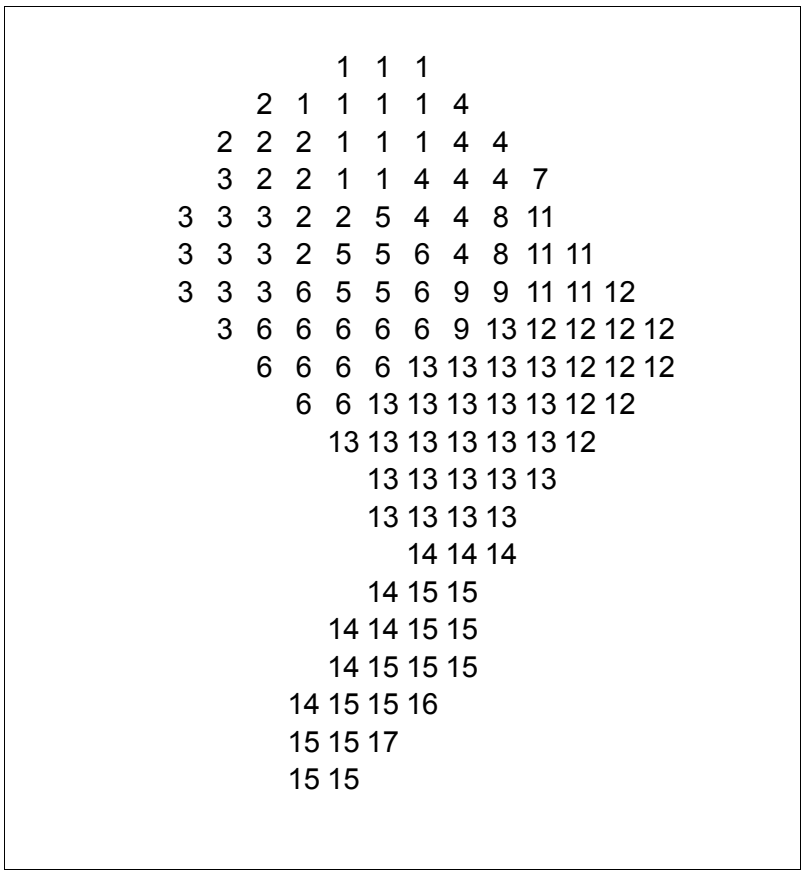
To generate such a result, DP needs pieces of data from an actual census. Each piece contains only a subset of the data in the census, so DP computes an arrangement of households that closely matches every input. In essence, the pieces of data collectively form a huge equation, and DP tries to find the best solution to it.

The first requirement is a set of region statistics. The geographic region to be analyzed must be broken down into small regions, then composite numbers must be provided for each region. DP works at a resolution of regions – households will be placed into regions so as to best match the statistics, but each household's location within a region will be totally random. Generally, the most convenient regions to use are the tracts defined by a recent government census.



Geographic map with regions.

DP doesn't work with vector maps; it requires a raster version of the region map. The raster map simply contains a region number in each cell.



Region map, rendered as a raster.

A table file provides composite statistics for the regions defined above. Each row in this table defines one region, and lists overall values for the region. DP will use these values as goals, and attempt to place households in a way that best matches them.

Region Number	Rabbits	Mouses	White Animals	Grey Animals
1	23	18	20	21
2	5	15	4	16
3	4	6	5	3
4	104	31	99	36
5	15	23	35	3
etc.				

Region statistics for above map.

The second piece of data is a set of archetype households. Generating fake households from scratch is far too complicated, so instead DP takes a set of actual household and population data as input, and uses them as inspiration for the households placed in the final map. The archetype list need not be

complete; DP will duplicate the ones provided as many times as necessary to fully populate the map. The selection and placement of households will be governed by the region statistics defined above.

Household Number	Dwelling Type	Building Size
1	Apartment	8-unit
2	House	2-story

Sampling of households from the map.

Household Number	Species	Color
1	Rabbit	Grey
1	Rabbit	Grey
2	Mouse	White
2	Mouse	Grey
2	Rabbit	White

Members of above households.

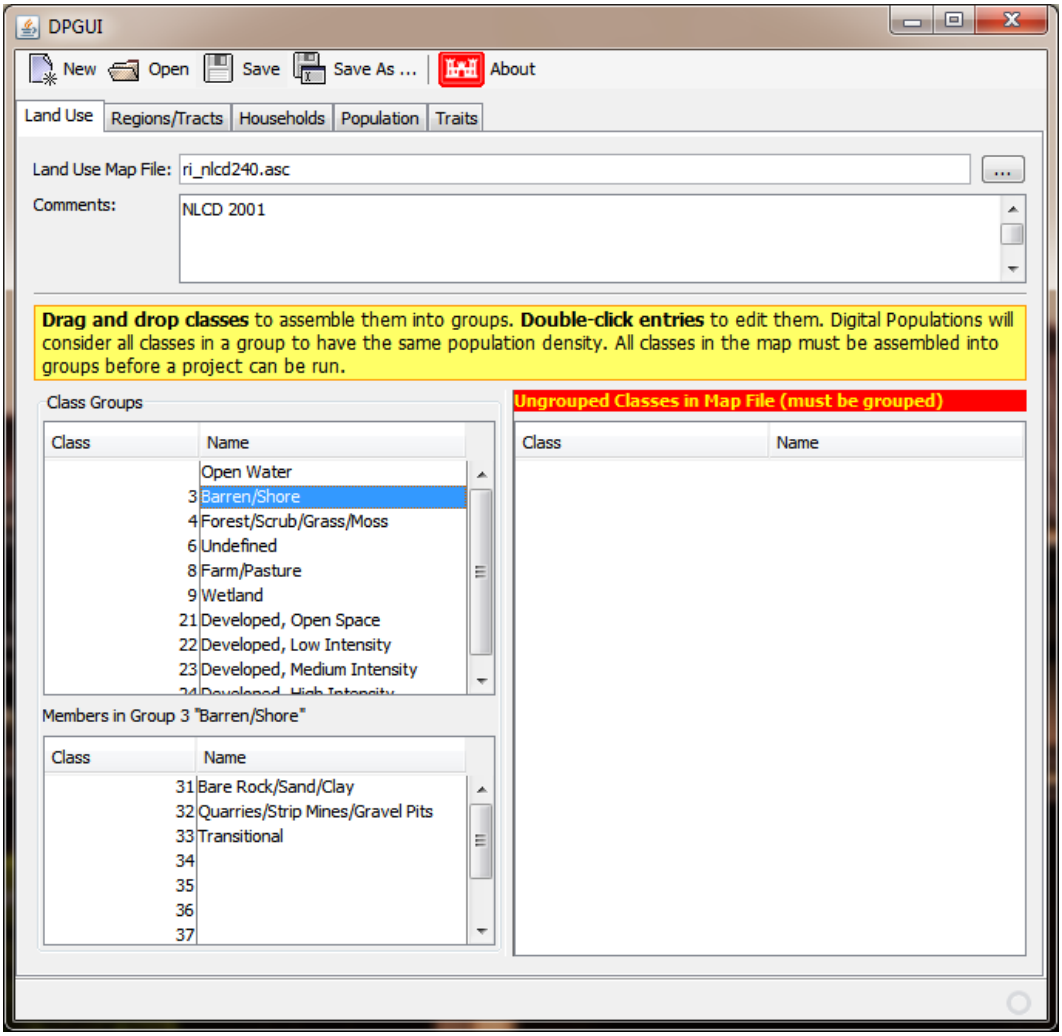
The final piece of data is the control structures that describe the above files, and also specify the exact problem to be solved (i.e. how to compare region goals to copies of the archetypes, and which values are important to a run.) The **relationship file** lists the files and links them together, and the **fitting criteria** file selects the important criteria from the relationship file. These files can be created with the GUI, documented below.

Relationship File GUI

To help create the control files, Digital Populations provides a GUI to help the user select and describe the input files. Operation is as simple as walking through the tabs one at a time, and filling out the fields.

Tab “Land Use”

The Land Use tab specifies a raster map as described above plus simplifying criteria.



Land Use Map File specifies the raster map file itself. DP doesn't care about the meaning of the groups; it only cares about the population density of each cell in the map. The lower part of this tab is used to create **Class Groups**, where the classes in each group are expected to have similar densities. DP will compute the actual density value, then try to mimic this value when placing households on the map.

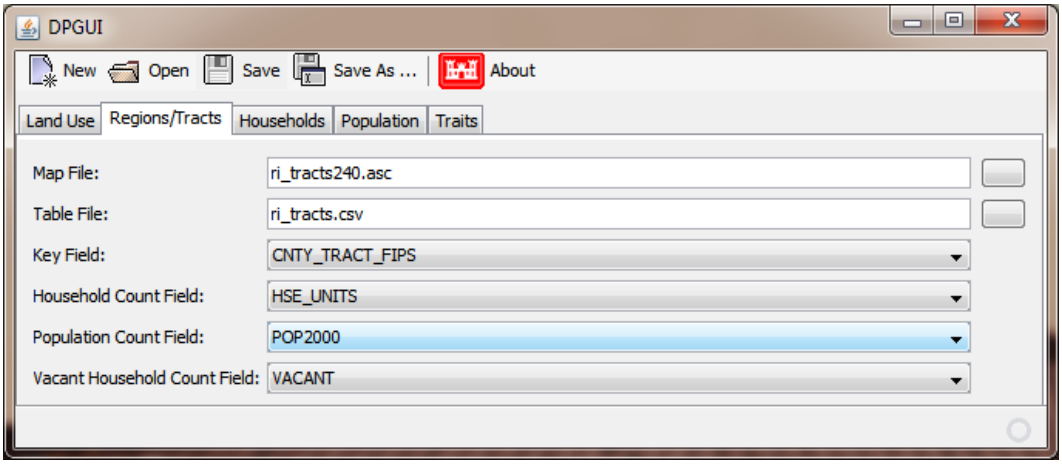
To create a group,simply select one or more classes from the list under **Ungrouped Classes**, then drag them into **Class Groups**. DPGUI will automatically create a group with a default class number and name. Both of these can be freely changed. Selecting a group will display its contents under **Members in Groups**, and classes can be dragged in and out of this list as well.

The first line of Class Groups is the “vacant cell” class, and always implies a total lack of households. Classes that have no residents should be dragged to this group. The name can be changed, but this group has no number.

DP requires all classes to be grouped before a run, so please ensure the Ungrouped Classes list is empty before using the relationship file.

Tab “Regions/Tracts”

To compute the best location for each household, DP needs a set of statistics that it can target. It also needs a map defining the regions for which goal statistics are provided. These regions also define the resolution at which DP will work: households will be placed into regions in a way that closely mimics the statistics, but the exact location of each household within the region will be random.



Map File specifies the raster map that provides the location of each region, and **Table File** provides the composite statistics for each region. Each cell in the map has a number indicating the region it belongs to, and each row in the table provides data for one region. **Key Field** specifies the column within the table file that links the two files, specifying the region for which the row provides data.

Household Count Field specifies the column that gives the total number of households in each region, while **Population Count Field** does the same for individuals. **Vacant Household Count Field** gives the number of households in each region that are unoccupied.

Tab “Households”

Rather than fabricating households from whole cloth, DP instead takes a table of “archetype households” which will be cloned as needed to populate the final map. This table should contain a thorough sampling of real households, as characteristics (combinations or traits of household members) absent from this table will be absent from the output map as well.

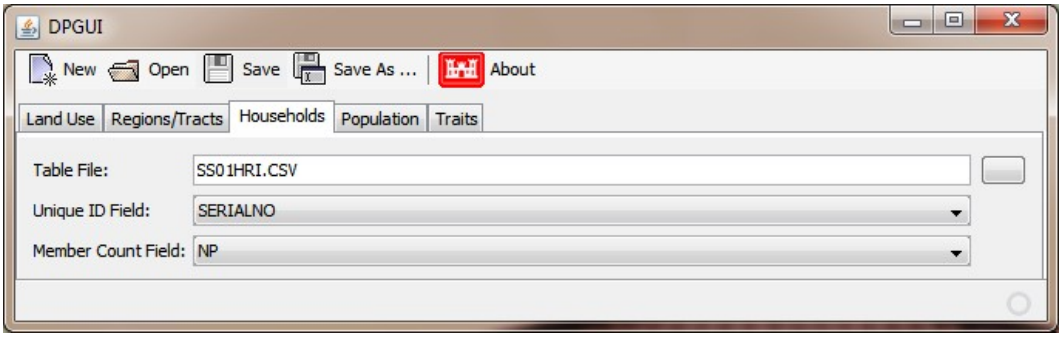


Table File specifies the household table where each row describes one household. **Unique ID Field** gives each household a unique number or code that the population table can refer to. **Member Count Field** specifies the number of people living in each household. DP will compute this count from the population table if provided.

Tab "Population"

DP also accepts an optional archetype population table that provides specific data for the members of each household.

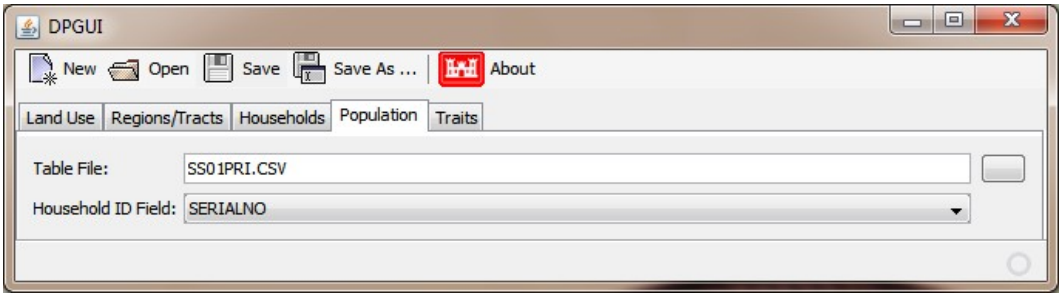


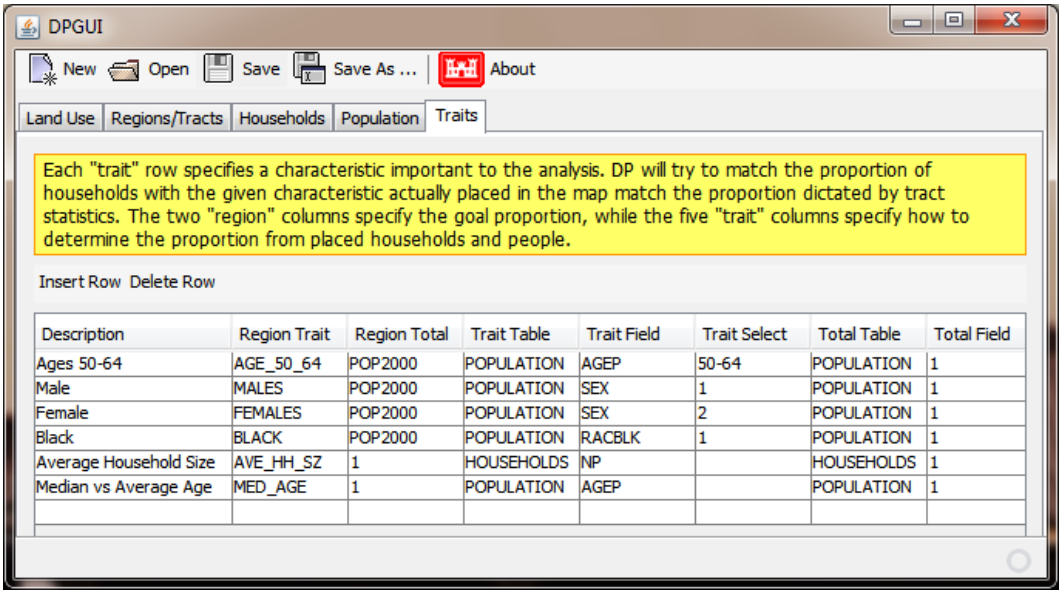
Table File specifies the population table where each row describes one individual. **Household ID Field** provides the value from the household table's Unique ID Field that identifies the specific house this person lives in. Records without household ID values will be ignored.

Tab "Traits"

The Traits tab is the hardest part of the system to understand, as it describes how the above tables are to be used to generate a new census. Each row here describes a characteristic along with how to calculate the characteristic from the above tables. DP uses these rows to evaluate the quality of arrangements of households.

DP starts by internally creating two traits from Household Count Field and Vacant Household Count Field above. These ensure the raw numbers of households are correct, and along with the population density data computed from the land use files, that the households are distributed where the people actually live.

The Traits tab allows the user to specify additional criteria so that the households place actually mirror important characteristics of the actual population being studied. DP does not necessarily use all the criteria listed here; it requires another "fitting criteria" file to select which of these rows will be used in a given run.



Description	Region Trait	Region Total	Trait Table	Trait Field	Trait Select	Total Table	Total Field
Ages 50-64	AGE_50_64	POP2000	POPULATION	AGEP	50-64	POPULATION	1
Male	MALES	POP2000	POPULATION	SEX	1	POPULATION	1
Female	FEMALES	POP2000	POPULATION	SEX	2	POPULATION	1
Black	BLACK	POP2000	POPULATION	RACBLK	1	POPULATION	1
Average Household Size	AVE_HH_SZ	1	HOUSEHOLDS	NP		HOUSEHOLDS	1
Median vs Average Age	MED_AGE	1	POPULATION	AGEP		POPULATION	1

Description is a free-text field that lets the user give each row a useful name.

Region Trait and **Region Total** specify columns from the Regions/Tracts table that together form a target proportion for each region. For example, the second row “Male” specifies the number of males in each region as the numerator, and the total number of people in each region as the denominator, resulting in the percentage of each region that is male. DP will place households in each region so that the proportion of individuals placed that are male is close to this target.

The remaining five columns describe how to compute a proportion from the households placed in a map. **Trait Table** specifies whether the numerator will come from the household table (on the Households tab) or the population table. **Trait Field** specifies which column from that table provides the required value. **Total Table** and **Total Field** work the same way, but provide the denominator.

Trait Select controls the way the values from Trait Field are used. If left blank, DP will simply add together all the values, and divide by the sum of the Total Field values, effectively providing and average over the households in question. If a list of values or ranges is given here, DP will instead *count* the number of objects (households or persons) that have a Trait Field value that is one of the values listed, then divide by the Total Field sum as above.

In the above example, the row labeled Male needs to compute the proportion of a region's population that is male. This is accomplished by counting the number of males, and dividing by the total number of people. The Population table doesn't have a “MALE Y/N” column, but instead provides a “SEX” column containing a code that indicates the gender of each person. So POPULATION/SEX/1 selects all individuals with a SEX value of “1”, indicating male. And POPULATION/1 simply counts the number of people in the region.

Trait Field and Total Field can contain a fixed number instead of a field name, and this value will be used instead of the data tables. For example, HOUSEHOLDS/1 will provide a count of the households that DP has placed into a region, and POPULATION/1 will provide a count of the people placed.