Population Synthesizers

John L. Bowman

May 2008

Microsimulation models that forecast the activities and travel of urban populations first create a synthetic population and then simulate the behavior of the households and persons in that synthetic population. Typically, they create a base year synthetic population from census data, and then use aggregate demographic and land use forecasts to create a synthetic population for each forecast year. The synthesis procedure involves two main steps. First a demographic distribution of households is estimated for each transportation analysis zone or small census area (zone), and then a matching sample of households is drawn from a set of household records for which nearly complete census information is available (microdata sample).

The demographic distribution is defined discretely by the cartesian product of several categorical control variables (dimensions), with each multidimensional category (or cell) defined as a unique value combination of the one-dimensional control variables. The number of households in each cell is estimated through an iterative proportional fitting procedure (IPF). The IPF procedure starts with an initial joint distribution available for aggregate census geographical units. It then cycles iteratively through a set of control totals, one total for each category of each control variable. For each control total it adjusts the joint distribution to satisfy the total by proportionally adjusting all cells with attributes matching the control total's category. If the control totals are mutually consistent, then IPF eventually converges so that all control totals are satisfied and the correlation structure of the initial joint distribution is preserved. Control totals are taken from census tables for the base year, and for the forecast years they come from demographic and land use forecasts, which may be less detailed. In estimating a base year distribution, nearly all population synthesizers control for household income, household size and number of workers. Additional household characteristics used as controls in some cases include age, gender or race of householder; presence of children; and family vs. non-family household. Most of the controls are one-dimensional, but some synthesizers make heavy use of available 2- and 3-dimensional controls.

Once the joint distribution is determined, with each cell indicating the number of households of a certain type residing in the zone, then for each cell its number of households is drawn randomly from the corresponding subset of household records available in the microdata sample. This process typically includes four steps. Since the estimated distribution specifies non-integer values for the number of households of each demographic category within a geographic unit, the first step is usually to adjust all those values to integers. Second, a Monte Carlo procedure is employed to draw the correct number of households of each type from the census sample. Third, the useful household and person variables are extracted from the drawn census household and retained for use by the model system. Finally, in some cases, an additional procedure is used to assign each household to a more precise location within its geographic unit. The end result is a synthetic population in which each synthetic household and its members have many clearly defined characteristics for use in the model system, and together they match the estimated demographic distribution within each zone.