Synthetic Population Generation: A Heuristic Data-Fitting Approach and Validations

Sivaramakrishnan Srinivasan
Lu Ma
Dept. of Civil and Coastal Engineering
University of Florida

The past several years have seen tremendous developments in disaggregate travel-demand models. This interest is motivated by several factors such as (1) reduction of aggregation errors, (2) ensuring sensitivity to demographic shifts like the ageing of the population, (3) capturing differential response of travelers to policy actions, and (4) addressing special travel-needs of certain population groups.

The application of such models for predictions and policy evaluations requires as inputs detailed information on the socio-economic-mobility characteristics of the population. Synthesis methods are used to first generate the population for a base year (current year/census year) and this, in turn, is used as an input to generate the target year (forecast year) population. The state-of-the-practice approach to population synthesis involves the use of the Iterative Proportional Fitting (IPF) method. While there have been several applications of this approach, the following issues still remain. First, the number of controls used in the synthesis of the base-year population has been limited. In particular, most practical applications control only for household-level attributes (e.g., household size and dwelling-unit type) and not for person-level attributes such as age and gender. Thus, the synthesized base-year population may not truly match the observed person-level distributions. This would affect the accuracy of the target-year population as the synthesized base-year population is used as an input to generate the target-year population. Second, documentation of the validation of the synthesis procedure, especially in the context of a “target” year population, is limited. The broad focus of this research is to contribute towards synthetic population generation (SPG) literature by addressing these two issues.

The rest of this abstract is organized as follows. A heuristic data-fitting algorithm is first described that can be used to synthesize populations by simultaneously controlling for household-level and person-level characteristics. Next, the results of the application of the algorithm for both base-year and target-year population synthesis are presented.

As already indicated, state-of-the-practice methods for population synthesis fundamentally involve the development of a joint multi-way distribution using IPF. The development of this joint multi-way distribution requires that all controls are at the same “universe” (such as households). This condition is violated when both household- and person-level controls are present. Thus, a heuristic data-fitting algorithm was developed to systematically draw households from a “seed” dataset (such as the PUMS) such that several control tables (at household and person levels) are satisfied. In each iteration, a “fitness” value is calculated for each household in the seed dataset. This value is a measure of the extent to which the household contributes to satisfying the target values in all the different control tables simultaneously. The household with the highest fitness is drawn into the synthetic population of the census tract – thus a “greedy” heuristic is employed. When adding a household would violate several control tables, its fitness
would have a negative value. The synthesis procedure stops when all households in the seed data have negative fitness and hence none can be selected into the synthesized population. Empirical testing indicates that, with this stopping criterion, the number of synthesized households is approximately equal to the actual number of households in the census tract.

The procedure was implemented in GAUSS, a matrix programming language. The code was used to synthesize the year 2000 population for 13 census tracts of varying populations and areas in Florida. For each tract, two sets of populations were estimated – the first with only household-level controls (seven control tables from SF1 and SF3 files of the US Census) and the second with both household- and person-level controls (total of twelve control tables). Validation analysis indicates that the second synthesized population matches the true distributions much better. In fact, for the first population, the extent of mismatch with the (uncontrolled) person-level tables is significant.

As a second step, the populations of 1990 were synthesized for the same 13 census tracts. Once again, two sets of populations were synthesized. One used the year 2000 population synthesized with both household and person controls as the seed data whereas the second used the year 2000 population synthesized with only household controls as the seed data. Only three attributes (household size, age, and gender) were controlled in the target-year synthesis. The aggregate characteristics of the synthesized populations were compared with several other control tables from the 1990 SF1 and SF3 files. The results indicate that the first population (using seed data generated with both person and household controls in the base year) match the target-year true values more closely.

In summary, this research develops a heuristic data-fitting methodology that can be applied to synthesize populations by controlling for several attributes at both household and person levels simultaneously. The procedure was applied to synthesize both base-year and target-year populations for thirteen census tracts in Florida. The results indicate that the greedy-heuristic procedure results in synthetic populations that match rather closely with the true distributions. Further, the results also highlight the improvements that can be achieved by controlling for both household and person level attributes.