**SEMCOG Population Synthesis**

For 2045 Regional Development Forecast, SEMCOG adopted a new population synthesis procedure to develop its 2015 base year households and population. This procedure includes three major steps: first, use UrbanSim Synthpop to generate households and population by Census Block Group; second, assign synthesized households to housing units within Block Group; third, refine population age distributions to match REMI base year controls.

1. **UrbanSim Synthpop with adjusted household control totals**

SEMCOG reviewed and explored major population synthesizers at the time of work and decided to adopt the “Synthpop” population synthesizer (<https://github.com/UDST/synthpop>) from UrbanSim Inc. This synthesizer implements IPU (Iterative Proportional Updating) fitting method initially developed by Ye *et al.* at Arizona State University. A major advantage of this synthesizer, comparing with the previous synthesizer used by SEMCOG, is that the synthesis process matches distributions of both household and population attributes, resulting much improved synthesized population at small geographies. This is important as personal attributes are essential inputs for both SEMCOG Regional Development Forecast (RDF) and Regional Transportation Plan (RTP) Models. For details about IPU based population synthesis method, please refer to this paper. (<http://www.scag.ca.gov/Documents/PopulationSynthesizerPaper_TRB.pdf>)

SEMCOG used 2015 PUMS household and person records as synthesis “seed”. For synthesis controls, based on the needs of RDF and RTP models, SEMCOG selected total of 13 household and personal marginal variables:

*Household variables (10): age of head; race of head; Hispanic head or not; household size; income; vehicle; workers; presence of children; presence of seniors; housing tenures*

*Personal variables (3): age; gender; race*

Since 2015 is not a decennial Census year, above marginal summaries are only available at Block Group level in Census 5-year ACS. Two significant issues with 5-year ACS are that Block Group marginal patterns are “averaged” between 2011 and 2015 and the BG counts do not add up to the latest 2015 1-year county estimates. Those issues will cause household placement difficulties in next synthesis step and make synthesized population miss its true targets. Therefore, SEMCOG developed a MCD-to-Block Group crosswalk table and distribute its own MCD estimates, which is controlled to 2015 1-year ACS county estimates, to Block Groups. As a result, synthesized demographic data will meet both SEMCOG MCD and Census county estimates as close as possible. Original ACS 5-year Block Group marginal distributions by household and population attributes were then adjusted according to new Block Group targets and served as the new marginal controls in synthesizer.

Early synthesis tests showed that Synthpop had good matches to household targets but population results were poor. Depends on the specific county in SEMCOG region, the synthesized population could be 2% to 8% less than its control totals. Further analysis showed that this discrepancy was largely caused by ACS household size distribution, which is heavily skewed toward small households. SEMCOG corrected this issue by shifting overall household size distribution upwards while holding both BG household and population controls constant. At the end, synthesized household and person counts at county level are around 99.9% and 99% accuracy.

1. **Household Assignment to Residential Units**

At this stage, synthesized households were further placed into individual residential housing unit within Block Groups. Households from step 1 synthesis carry PUMS housing information such as property types and year built, allowing the possibility to match households to suitable residential units from 2015 SEMCOG building data. SEMCOG developed a special household assignment program and implemented the standard “Stable Matching Problem” algorithm (<https://en.wikipedia.org/wiki/Stable_marriage_problem>). This algorithm guarantees that there is no pair of household and building where the building is more suitable to the household than its current home and the household is more suitable for the building than its current occupant. The quality of the placement hinges on the definition of suitable used in the algorithm.

The function chosen considered a series of comparisons in order:

1.       Do they have the same County?

2.       Do they have the same MCD?

3.       Do they have the same Tract?

4.       Do they have the same Block Group?

5.       Do they have the same Tenure status?

6.       Are the buildings type and households building code compatible?

7.       Are the number of housing units is compatible with the household’s building info?

8.       How similar are the quantile of the property value of the building to quantile of the home value / rent of the household?

9.       How similar is the age of the building to the age of the households dwelling, recoded into Census categories?

10.   How similar are the quantile of the property value of the building to quantile of the income of the household?

1. **Refine synthesized population to match REMI’s 2015 age distribution**

REMI population forecast provided essential demographic controls for SEMCOG forecast model. Its base year age distribution is based on 2015 1-year ACS, which is different from the 5-year ACS synthesized population. This difference broadcasts into the forecast period and causes some volatility in demographic forecast simulation, especially in early forecast years. Therefore, SEMCOG applied a refinement process to better align synthesized household ages to REMI baseline age distribution. The refinement program randomly selected synthesized population that are not the heads of households from each 5 year age group and “aged” them into the next higher age group. Care was taken to maintain the single year age distributions within each age group.

The final synthesized households and persons structure:

Households:

1. age of head
2. race of head (White, Black, Hispanic, Others)
3. income
4. # of workers
5. # of cars
6. # children
7. # of persons
8. building ID

Persons:

1. age
2. race (White, Black, Hispanic, Others)
3. worker status
4. household ID