Recent Developments in Population Synthesis

Gaurav Vyas

Peter Vovsha (WSP)

What is Population Synthesis?

- Creation of synthetic population for entire region from Census sample:
  - Individual HHs with full set of characteristics

What Do We Know?

- Sample of HHs with a full set of HH and person characteristics
- Marginal controls (base year):
  - Total population and HHs by TAZ/MAZ
  - HH distribution by size, income group, #workers by TAZ/MAZ, etc..
  - Population distribution by age, gender, etc.
- Marginal controls (future years):
  - Total population and HHs by TAZ/MAZ
  - Average HH size, income, #workers by TAZ/MAZ etc.

Essence of Population Synthesis

- Expansion of sampled HHs:
  - Meet regional controls
  - Use HHs as uniformly as possible ("representative population")
- Allocation of expanded HHs to TAZs/MAZs:
  - Meet TAZ/MAZ controls

PopSyn III Recap

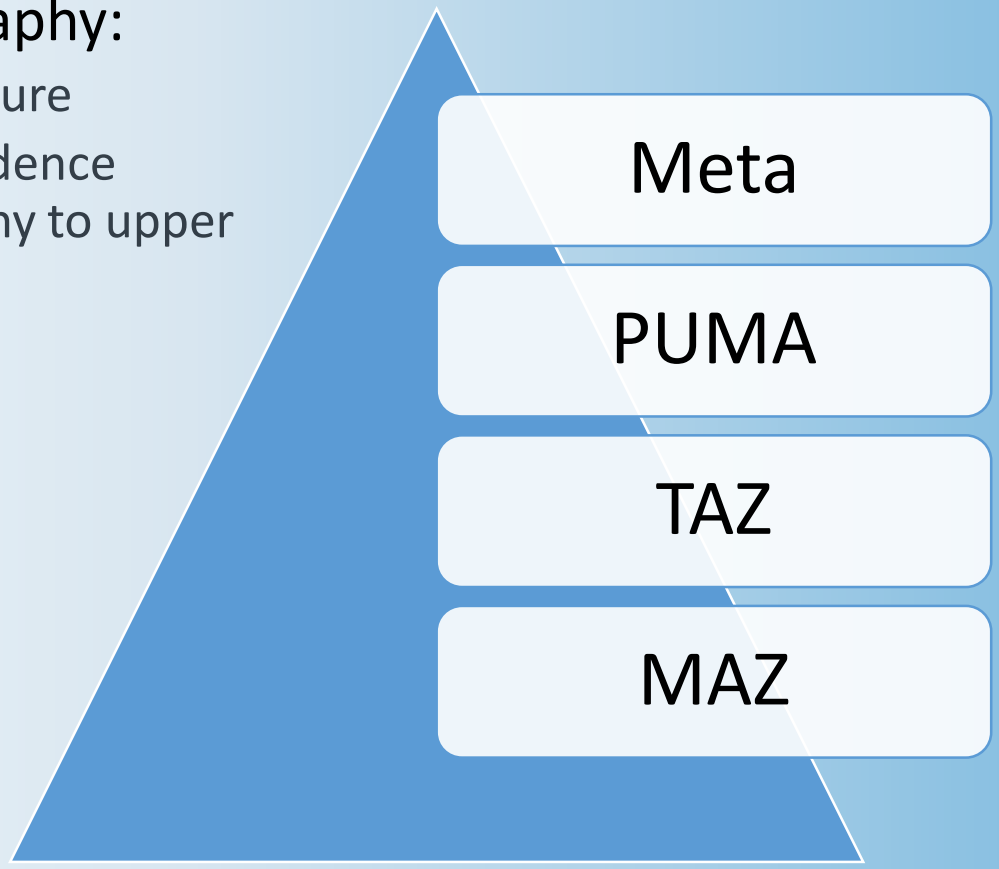
- Innovative theoretically consistent algorithm based on the maximum entropy principle, originally developed for Maricopa Association of Governments (MAG):
  - Core list balancing procedure:
    - Can handle any number of household-level and person-level controls at different levels of geography
    - User can specify differential importance weights reflecting relative importance and reliability of controls
    - Preserves uniformity of HH expansions as much as possible
  - LP discretizing method to convert fractional HH weights to discrete numbers:
    - Eliminates Monte-Carlo simulation error, all procedures are analytical and repeatable

PopSyn III Recap

- Core List balancing implemented w/relaxations:
  - Handles inconsistency among different controls
  - Produces a unique convergent solution w/controls satisfied to the extent possible
  - Degree of the necessary relaxation of each control is inversely related to the importance
- Multiple level of geography:
  - Important demographic & socio-economic trends:
    - Can only be translated into more aggregate controls than TAZ-level
    - Handled by upward meta-balancing
  - New generation of CT-RAMP ABMs operate with enhanced level of spatial resolution:
    - Location choices are modeled at the level of Micro-Analysis Zones (MAZs) nested within TAZs
    - Handled by downward allocation

PopSyn III Recap

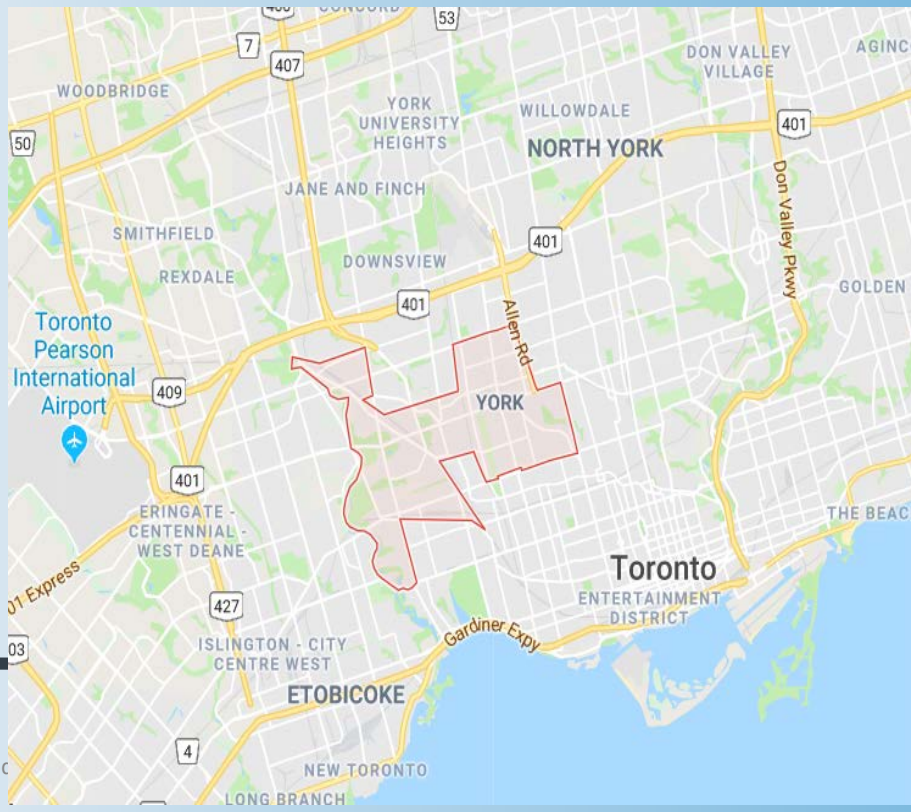
- Multiple level of geography:
  - Nested geography structure
  - Many-to-one correspondence between lower geography to upper geography



New Features of 4th Generation

1. Differential specification of controls by geography

- Detailed controls may not be possible to obtain for the entire modeling region, especially for future year scenarios
- Controls available at aggregate geography (smaller than Meta but greater than TAZ/MAZ):
  - Population by age at municipality level but not at TAZ level



1. Differential specification of controls by geography

- Still possible to take advantage of aggregate controls for certain part of the region:
  - For instance, income not available for certain PUMAs, but total households is available.
  - Total households for these PUMAs can be used as a control
  - For other PUMAs, detailed controls by household income used
- Such flexible specification allows for using controls which would have been dropped in previous versions of PopSyn
  - More detailed controls for a certain sub-area of interest and more aggregate controls outside of the sub-area of interest

1. Differential specification of controls by geography: Validation, Columbus OH

Income control for PUMA = "all"

Test 2

Control	Difference (PopSyn - Control)					
puma	NumHH	Hinc1	Hinc2	Hinc3	Hinc4	Hinc5
1	-	(24)	2	(2)	9	9
2	-	(12)	(1)	(9)	8	11
3	-	(9)	(1)	(10)	10	10
4	-	(16)	2	(8)	6	12
5	-	(9)	(5)	(7)	10	8
6	-	(2)	3	(7)	(7)	16
7	-	(13)	(5)	(9)	8	15
8	-	(2)	-	(8)	(1)	14
9	-	8	7	(16)	(4)	35
10	-	(8)	(2)	(8)	6	2
11	-	(982)	(1,537)	(4,052)	(4,438)	11,303
12	-	(7)	7	(10)	5	16
13	-	(19)	3	3	7	4
14	-	(36)	-	4	19	11
15	-	(15)	1	(2)	10	10
16	-	(172)	46	37	58	22
17	-	(52)	19	14	21	5
18	-	(59)	21	12	21	6
19	-	(80)	13	29	35	10

Income control for PUMA = 1,2,3,4,5

Test 1

Control	Difference (PopSyn - Control)					
puma	NumHH	Hinc1	Hinc2	Hinc3	Hinc4	Hinc5
1	-	(24)	2	(2)	9	8
2	-	(12)	-	(9)	8	10
3	-	(8)	(1)	(10)	10	9
4	-	(15)	2	(8)	6	12
5	-	(9)	(4)	(7)	9	8
6	-	-	-	(297)	(162)	444
7	-	(835)	-	(261)	71	1,021
8	-	(192)	-	(696)	-	1,314
9	-	692	296	(1,610)	(762)	2,908
10	-	(440)	-	(389)	605	112
11	-	(968)	(2,579)	(4,021)	(5,815)	13,677
12	-	(360)	335	(427)	-	940
13	-	(1,233)	-	124	797	334
14	-	(2,819)	-	-	1,866	1,073
15	-	(834)	-	-	497	425
16	-	(15,694)	-	684	5,491	9,090
17	-	(3,261)	-	360	2,151	757
18	-	(3,662)	398	260	2,242	764
19	-	(6,129)	-	1,298	3,902	879

For first test, better income distribution for PUMA list specified for the control

2. Multiple sources for household samples

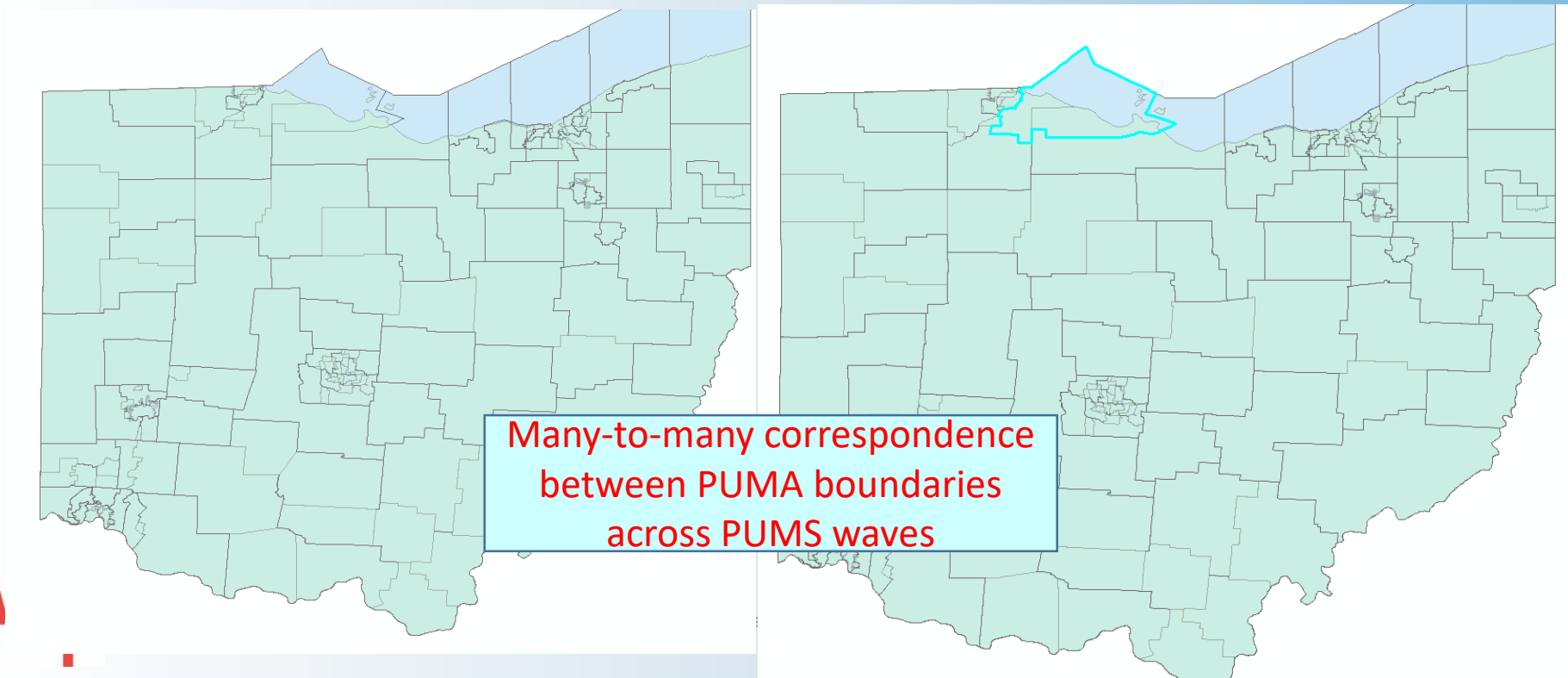
- Most population synthesizers cannot handle multiple sources of seed data:
  - US applications:
    - ACS 2007-2010 vs ACS 2011-2015
  - Outside-US applications:
    - Rich HTS vs other data sources
      - TTS (5% sampling rate) vs PUMF (Toronto)
      - HTS (2.5% sampling rate) vs Census (Jerusalem)
      - Conceivable although not common in US
- Two options with existing PopSyn:
  - Choose one of the waves
  - Try to combine seed data sources

2. Multiple sources for household samples

- Choose one of the waves of PUMS/data sources as seed:
  - Seed data would be 5% sample instead of richer accumulated 10-15%
  - Under-representation of certain population segments (university students, highest/lowest income group)

2. Multiple sources for household samples

- Combine different waves into one:
  - Significant effort in bringing different waves to common denominator: variable coding, geography layer overlap
  - Not transferable from one region to other



2. Multiple sources for household samples

- PopSyn IV incorporates multiple seed data waves:
  - Handle differences in data structures across samples:
    - Census based or household survey based from different years
  - Overlapping non-nested geographies across different samples:
    - PUMA 2000 layers and PUMA 2010 layer
  - No need to pre-process multiple seed data to bring them to common denominator

3. Simultaneous List Balancing

- PopSyn III geography structure is highly hierarchal:
  - Nesting geography
- Sequential allocation from larger to smaller geography:
  - PUMAs, TAZs, and MAZs are balanced sequentially
  - Error propagation to the last TAZ/MAZ (as shown by Paul et.al. 2017)

3. Simultaneous List Balancing

- PopSyn IV features simultaneous global list balancing:
  - Region, PUMA, TAZ, MAZ
- Alleviates error propagation due to ordering of PUMAs, TAZs, and MAZs
- Seamlessly incorporates any non-nested geographies for different controls

Conclusions

- Differential control specification by geography:
  - Full advantage of all available information as opposite to setting uniform controls for the entire population
  - Incorporate additional important controls for a subarea which would have been difficult to handle using a standard population synthesizer
- Enriched seed data by combining several sources:
  - Better representation of certain segments for future year scenarios
- Simultaneous global list balancing:
  - No error propagation
  - Controls at non-nested geography

Contact(s)

Gaurav Vyas

Systems Analysis Group
Gaurav.Vyas@wsp.com

Additional contact

Peter Vovsha, PhD

Systems Analysis Group
Peter.Vovsha@wsp.com