1

PopGen – 1.1

DATA STRUCTURES

2

Data Structures

- When providing their own data, users need to follow certain data structure formats
- In this presentation, appropriate formats for specifying your own data in PopGen are presented

3

Data Structures

- All the input files have some common features
  - First row contains the variable names
  - Second row contains the variable types
    - Integers – bigint
    - Floating point values – double
    - Strings - text
  - First few columns in any table are compulsory followed by optional columns
  - In the sample file, categories for population attributes should:
  - Start with 1 and subsequent categories are in increments of 1
  - TAZ's should be treated as blockgroups when setting up data

4

Data Structure for Household Sample File

- The file provides the household attributes of the sample from which synthetic population is drawn
- Data structure
  - Compulsory fields – state, pumano, hhid, serialno
  - Optional fields – household attributes

state	pumano	hhid	serialno	<householdvariable1>	<householdvariable2>
<variable type>	<variable type>	<variable type>	<variable type>	<variable type>	<variable type>
value	value	value	value	value	value
value	value	value	value	value	value
value	value	value	value	value	value
value	value	value	value	value	value

5

Data Structure for Groupquarter Sample File

- The file provides the groupquarter attributes of the sample from which synthetic population is drawn
- Data structure
  - Compulsory fields – state, pumano, hhid, serialno
  - Optional fields – groupquarter attributes

state	pumano	hhid	serialno	<groupquartervariable1>	<groupquartervariable2>
<variable type>	<variable type>	<variable type>	<variable type>	<variable type>	<variable type>
value	value	value	value	value	value
value	value	value	value	value	value
value	value	value	value	value	value
value	value	value	value	value	value

6

Data Structure for Person Sample File

- The file provides the person attributes of the sample from which synthetic population is drawn
- Data structure
  - Compulsory fields – state, pumano, hhid, serialno, pnum
  - Optional fields – person attributes

state	pumano	hhid	serialno	pnum	<personvariable1>	<personvariable2>
<variable type>	<variable type>	<variable type>	<variable type>	<variable type>	<variable type>	<variable type>
value	value	value	value	value	value	value
value	value	value	value	value	value	value
value	value	value	value	value	value	value
value	value	value	value	value	value	value

7

Data Structure for Household Marginal File

- The file provides the household marginal distributions that the synthetic population should match
- Data structure
  - Compulsory fields – state, county, tract, bg
  - Optional fields – household marginal distributions

state	county	tract	bg	<householdvariable1category1>	<householdvariable1category2>
<variable type>	<variable type>	<variable type>	<variable type>	<variable type>	<variable type>
value	value	value	value	value	value
value	value	value	value	value	value
value	value	value	value	value	value
value	value	value	value	value	value

8

Data Structure for Groupquarter Marginal File

- The file provides the groupquarter marginals distributions that the synthetic population should match
- Data structure
  - Compulsory fields – state, county, tract, bg
  - Optional fields – groupquarter marginal distributions

state	county	tract	bg	<groupquartervariable1category1>	<groupquartervariable1category2>
<variable type>	<variable type>	<variable type>	<variable type>	<variable type>	<variable type>
value	value	value	value	value	value
value	value	value	value	value	value
value	value	value	value	value	value
value	value	value	value	value	value

9

Data Structure for Person Marginal File

- The file provides the person marginal distributions that the synthetic population should match
- Data structure
  - Compulsory fields – state, county, tract, bg
  - Optional fields – person marginal distributions

state	county	tract	bg	<personvariable1category1>	<personvariable1category2>
<variable type>	<variable type>	<variable type>	<variable type>	<variable type>	<variable type>
value	value	value	value	value	value
value	value	value	value	value	value
value	value	value	value	value	value
value	value	value	value	value	value

10

Data Structure for Geographic Correspondence File

- The file provides the correspondence between the geography and the PUMA to which the geography belongs
- Data structure
  - Compulsory fields –county, tract, bg, state, pumano, stateabb, countyname

county	tract	bg	state	pumano	stateabb	countyname
<variable type>	<variable type>	<variable type>	<variable type>	<variable type>	<variable type>	<variable type>
value	value	value	value	value	value	value
value	value	value	value	value	value	value
value	value	value	value	value	value	value
value	value	value	value	value	value	value

11

Data Structure for Synthetic Housing File

- The file lists frequencies and ID's of the sample *households/groupquarters* selected in the synthetic population
- Data structure
  - Compulsory fields – state, county, tract, bg, hhid, serialno, frequency, hhuniqueid

state	county	tract	bg	hhid	serialno	frequency	hhuniqueid
value	value	value	value	value	value	value	value
value	value	value	value	value	value	value	value
value	value	value	value	value	value	value	value
value	value	value	value	value	value	value	value

12

Data Structure for Synthetic Person File

- The file lists frequencies and ID's of the sample *persons* selected in the synthetic population
- Data structure
  - Compulsory fields – state, county, tract, bg, hhid, serialno, pnum, frequency, personuniqueid

state	county	tract	bg	hhid	serialno	pnum	frequency	personuniqueid
value	value	value	value	value	value	value	value	value
value	value	value	value	value	value	value	value	value
value	value	value	value	value	value	value	value	value
value	value	value	value	value	value	value	value	value

13

Resolution

- If you wish to synthesize population at the resolution of
  - County
    - Set tract and bg columns to 0 in the marginal tables
  - Tract
    - Set bg column to 0 in the marginal table
  - Blockgroup
    - No change
  - TAZ
    - Treat TAZ's as blockgroup

14

Data Inconsistencies

- Avoiding data inconsistencies when providing your own data can prevent crashes and improve fit of the synthetic population
- Data Inconsistency Types
  - Person Total Inconsistency: Person total derived from household size distribution is not consistent with the given person total
  - Marginal Distribution Totals: The population totals derived from household/groupquarter/person variables of interest are not consistent

15

Data Inconsistencies

- Data Inconsistency Types (continued)
  - Sample Serial ID's
    - Sample serial numbers are not consistent across input files
  - Geographic Correspondence
    - More than one PUMA entry for a geography

16